

2011

# Multidimensional Epidemiological Transformations: Addressing Location-Privacy in Public Health Practice

AbdelMalik, Philip

<http://hdl.handle.net/10026.1/901>

---

<http://dx.doi.org/10.24382/3438>

University of Plymouth

---

*All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.*

***This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's prior consent.***



**MULTIDIMENSIONAL EPIDEMIOLOGICAL TRANSFORMATIONS:  
ADDRESSING LOCATION-PRIVACY IN PUBLIC HEALTH PRACTICE**

**VOLUME I**

**PHILIP SAMI ONSY ABDELMALIK**  
MHS (Community Health & Epidemiology)

**Thesis Submission  
in partial fulfillment for the Degree of**

**DOCTOR OF PHILOSOPHY (PHD)**

**Submitted to the University of Plymouth  
Faculty of Health and Education**

**November 2011**

**© Philip AbdelMalik. All rights reserved.**



# **MULTIDIMENSIONAL EPIDEMIOLOGICAL TRANSFORMATIONS: ADDRESSING LOCATION-PRIVACY IN PUBLIC HEALTH PRACTICE**

**Philip Sami Onsy AbdelMalik**

## **Abstract**

The ability to control one's own personally identifiable information is a worthwhile human right that is becoming increasingly vulnerable. However just as significant, if not more so, is the right to health. With increasing globalisation and threats of natural disasters and acts of terrorism, this right is also becoming increasingly vulnerable. Public health practice – which is charged with the protection, promotion and mitigation of the health of society and its individuals – has been at odds with the right to privacy. This is particularly significant when location privacy is under consideration. Spatial information is an important aspect of public health, yet the increasing availability of spatial imagery and location-sensitive applications and technologies has brought location-privacy to the forefront, threatening to negatively impact the practice of public health by inhibiting or severely limiting data-sharing. This study begins by reviewing the current relevant legislation as it pertains to public health and investigates the public health community's perceptions on location privacy barriers to the practice. Bureaucracy and legislation are identified by survey participants as the two greatest privacy-related barriers to public health. In response to this clash, a number of solutions and workarounds are proposed in the literature to compensate for location privacy. However, as their weaknesses are outlined, a novel approach - the multidimensional point transform - that works synergistically on multiple dimensions, including location, to anonymise data is developed and demonstrated. Finally, a framework for guiding decisions on data-sharing and identifying requirements is proposed and a sample implementation is demonstrated through a fictitious scenario. For each aspect of the study, a tool prototype and/or design for implementation is proposed and explained, and the need for further development of these is highlighted. In summary, this study provides a multi-disciplinary and multidimensional solution to the clash between privacy and data-sharing in public health practice.



# TABLE OF CONTENTS

## VOLUME I

<b>1. Impetus</b> .....	<b>2</b>
1.1. Introduction.....	2
1.2. Cross-national validity .....	4
1.3. Research Objectives .....	4
1.4. Originality and Contribution .....	5
1.5. Structure & Organisation .....	6
<b>PART I THE BACKDROP</b> .....	<b>8</b>
<b>2. Establishing Common Ground</b> .....	<b>10</b>
2.1. Introduction.....	10
2.2. Public Health "Practice" & "Functions" .....	10
2.3. Privacy & Personally Identifiable Data .....	13
2.4. Public Health and Privacy.....	15
2.5. Focus on Location .....	17
<b>3. Privacy Concepts and Principles</b> .....	<b>19</b>
3.1. Introduction.....	19
3.2. Psycho-Social Dimensions of Privacy.....	20
3.3. Privacy and Information Practice "Principles".....	22
<b>4. International Whirlwind Tour of Legislation</b> .....	<b>28</b>
4.1. Overview .....	28
4.2. Definitions .....	31
4.3. Application and Exceptions.....	36
4.4. Governance.....	37
4.5. Implications and Final Thoughts .....	38
<b>5. Public Health Practitioner Perceptions</b> .....	<b>40</b>
5.1. Introduction.....	40
5.2. The Perceived Impact of Location Privacy .....	45
5.2.1. Overview.....	45
5.2.2. Background .....	46
5.2.3. Methods.....	49
5.2.4. Results .....	53
5.2.5. Discussion .....	58
5.2.6. Conclusions .....	62
<b>PART II CONCESSIONS</b> .....	<b>65</b>
<b>6. Brief Overview of General Solutions</b> .....	<b>67</b>
6.1. Introduction.....	67
6.2. Access Control and Security.....	67
6.3. Suppression Of Information .....	68
6.4. Data Aggregation .....	69
6.5. Data Anonymization .....	71
6.6. Analytical Software Agents .....	72
6.7. Transformations .....	73
<b>7. Paving the Path To A Novel Transform</b> .....	<b>75</b>
7.1. Introduction.....	75
7.2. Uniqueness .....	75
7.3. Re-identification Risk.....	77
7.4. Managing re-identification risk .....	78
7.4.1. Creating the Canada Grid .....	80
7.5. Rethinking Spatial Aggregation.....	89
<b>8. The Multidimensional Point Transform</b> .....	<b>95</b>



8.1.	Background .....	95
8.2.	Objectives .....	97
8.3.	Methods .....	98
8.3.1.	Algorithm: Overview .....	98
8.3.2.	Data .....	100
8.3.3.	Algorithm: Preliminary Proof-of-Concept Implementation .....	102
8.4.	Results .....	104
8.5.	Discussion .....	109
8.5.1.	Re-Identification Risk .....	113
8.5.2.	Limitations .....	117
8.5.3.	Strengths .....	118
8.5.4.	Using Synthesised Populations .....	122
8.5.5.	Algorithm Refinement .....	123
8.6.	Conclusions .....	126

**PART III GUIDANCE ..... 127**

<b>9.</b>	<b>Towards A Conceptual Framework.....</b>	<b>129</b>
9.1.	Introduction.....	129
9.2.	Framework Overview.....	130
9.3.	The Recipient .....	136
9.3.1.	Trust .....	138
9.3.2.	Security.....	142
9.3.3.	Training.....	145
9.3.4.	Contract.....	146
9.4.	The Data .....	149
9.4.1.	Granularity .....	151
9.4.2.	Multiplicity .....	151
9.4.3.	Sensitivity .....	152
9.4.4.	Size .....	154
9.5.	The Purpose.....	156
9.5.1.	Contribution .....	157
9.5.2.	Necessity .....	157
9.5.3.	Effort.....	158
9.5.4.	Impact.....	160
9.6.	The Transform.....	166
9.6.1.	Uniqueness.....	167
9.6.2.	Identification Risk.....	168
9.6.3.	Erosion .....	168
9.6.4.	Analytical Effect .....	169
9.7.	The Output .....	170
<b>10.</b>	<b>From Conceptual to Concrete .....</b>	<b>173</b>
10.1.	Introduction.....	173
10.2.	Scenario Setup.....	174
10.2.1.	Recipient Domain .....	177
10.2.2.	Data Domain.....	179
10.2.3.	Purpose Domain .....	182
10.2.4.	Transform Domain .....	185
10.2.5.	Output Domain.....	187
10.3.	Scoring Bias .....	190
10.4.	Summary and Conclusions.....	191
<b>11.</b>	<b>Facilitating Practice Through Tools.....</b>	<b>195</b>
11.1.	Introduction.....	195
11.2.	Demistifying the Legislation .....	197
11.3.	Applying the Multidimensional Point Transform .....	201
11.4.	Guiding Data-Sharing .....	204
<b>12.</b>	<b>Future Directions.....</b>	<b>208</b>

12.1. Legislation .....	208
12.2. Transformations .....	209
12.2.1. Smart Aggregation.....	209
12.2.2. Synthesised Populations.....	210
12.2.3. The Multidimensional Point Transform.....	211
12.3. Governance Structures & Data-sharing .....	212
12.4. Tools .....	213
<b>13. Review and Concluding Remarks .....</b>	<b>214</b>
13.1. Objectives Revisited & Accomplishments Reviewed.....	214
13.1.1. Objective 1.....	214
13.1.2. Objective 2.....	214
13.1.3. Objective 3.....	215
13.1.4. Objective 4.....	215
13.2. Concluding Thoughts.....	216
<b>PART V REFERENCES .....</b>	<b>221</b>

## LIST OF TABLES

TABLE 1: INCLUSION OF HEALTH AND LOCATION INFORMATION IN THE DEFINITIONS OF "PERSONAL INFORMATION" IN CANADIAN LEGISLATION .....	33
TABLE 2: SECTIONS OF THE SURVEY .....	50
TABLE 3: NUMBER AND PERCENTAGE OF SURVEY PARTICIPANTS BY MAIN ROLE AND GEOGRAPHICAL SCOPE.....	54
TABLE 4: SUMMARY OF SURVEY FINDINGS .....	56
TABLE 5: FABRICATED TUBERCULOSIS CASES AND THE 2-ANONYMISED RESULT .....	71
TABLE 6: CANADIAN GRID DEVELOPMENT SECTIONS AND ATTRIBUTES .....	82
TABLE 7: SIMPLIFIED HYPOTHETICAL EXAMPLE OF THE WEIGHTED ASSOCIATION BETWEEN CENSUS TRACTS AND FORWARD SORTATION AREAS .....	86
TABLE 8: CENSUS TRACT POPULATION COMPARISON BETWEEN CREATED POPULATION GRID AND 2001 CENSUS PROFILE .....	87
TABLE 9: PROVINCIAL DIFFERENCES BETWEEN PROFILE AND GRID CENSUS TRACT POPULATIONS .....	88
TABLE 10: RESULTS OF THE MULTIDIMENSIONAL POINT TRANSFORM ALGORITHM WITH DIFFERENT PATIENT DATASET SIZES FOR NEW YORK COUNTY.....	106
TABLE 11: OUTLINE OF THE PROPOSED FRAMEWORK DOMAINS AND THEIR CORRESPONDING DIMENSIONS .....	134
TABLE 12: IMPACT MATRIX SHOWING EXAMPLES OF THE BENEFITS AND RISKS OF DISCLOSURE AND NON-DISCLOSURE OF DATA TO IMPLICATED GROUPS .....	161
TABLE 13: EXAMPLE OF A CUSTODIAN WEIGHTED-SCORE MATRIX .....	176

## LIST OF FIGURES

FIGURE 1: SECTIONS AND FLOW OF THE WEB-BASED SURVEY TO COLLECT PRACTITIONER PERCEPTIONS ON THE IMPACT OF PRIVACY ON PUBLIC HEALTH PRACTICE .....	42
FIGURE 2: PARTICIPANT RATING OF THE DEGREE TO WHICH PRIVACY RESTRICTIONS POSE AN OBSTACLE TO PUBLIC HEALTH PRACTICE .....	43
FIGURE 3: RELATIONSHIP BETWEEN SELF-RATED KNOWLEDGE OF PRIVACY LEGISLATION AND POLICIES AND THE DEGREE TO WHICH PRIVACY WAS RATED AS AN OBSTACLE BY SURVEY PARTICIPANTS.....	44
FIGURE 4: DISTRIBUTION OF 2001 CENSUS TRACTS ACROSS CANADA.....	81
FIGURE 5: ARCGIS MODEL FOR BUILDING GRID-DISSEMINATION AREA – FORWARD SORTATION AREA – CENSUS TRACT INTERSECT POLYGONS .....	85
FIGURE 6: EXAMPLE CENSUS TRACT - FORWARD SORTATION AREA SUB-AREA OVERLAY TO ILLUSTRATE THE HYPOTHETICAL EXAMPLE .....	86
FIGURE 7: DISTRIBUTION OF CENSUS TRACT POPULATION DIFFERENCE BETWEEN GRID-CALCULATED POPULATION AND 2001 CENSUS PROFILE.....	87
FIGURE 8: AN EXAMPLE OF A FORWARD SORTATION AREA “K1G” AND THOSE ADJACENT TO IT.....	90
FIGURE 9: AGGREGATION OPTIONS FOR FORWARD SORTATION AREA POLYGONS ADJACENT TO K1G: (A) “CLUSTERED”, MINIMIZING DISTANCE BETWEEN AGGREGATED FSAs, AND (B) “STRING”, WHERE AGGREGATION IS STRETCHED BASED ON OTHER PARAMETERS, IRRESPECTIVE OF GEOGRAPHY .....	92
FIGURE 10: ILLUSTRATION OF THE “HOMOGENEITY METRIC” BASED ON THE ADJACENCY OF FORWARD SORTATION AREAS .....	93
FIGURE 11: EXTENT OF THE K0A FORWARD SORTATION AREA AND THE OTTAWA CENSUS SUBDIVISION IN ONTARIO, CANADA (2006 CENSUS).....	94
FIGURE 12: MULTIDIMENSIONAL POINT TRANSFORM FLOW.....	101
FIGURE 13: SIMPLIFIED EXAMPLE OF AGE CATEGORISATION USING ONE-YEAR INTERVALS WITH 5 LEVELS AND K=5.....	104
FIGURE 14: MEAN CUMULATIVE PERTURBATION DISTANCE FOR SUCCESSIVE RUNS OF THE TESTED PATIENT DATASET SIZES .....	106
FIGURE 15: STRUCTURE OF THE PROPOSED PUBLIC HEALTH DATA-SHARING FRAMEWORK SHOWING DOMAINS AND THEIR CORRESPONDING DIMENSIONS .....	132
FIGURE 16: FLOW OF THE PROPOSED PUBLIC HEALTH DATA-SHARING FRAMEWORK .....	133
FIGURE 17: TWENTIETH CENTURY POSTER PROMOTING SMOKING THROUGH THE MEDICAL PROFESSION.....	153
FIGURE 18: YAHOO PIPE CREATED TO AUTOMATE RSS FEED COLLECTION FROM SEVERAL WEBSITE ON PRIVACY AND HEALTH HEADLINES .....	196

FIGURE 19: SCREENSHOT OF THE WEBSITE HOMEPAGE SHOWING THE RESULTS OF THE YAHOO PIPE FOR THE NEWS FEED MASHUP ..... 197

FIGURE 20: SPLASH SCREEN OF THE DEMO TOOL DESIGNED AS A PUBLIC HEALTH GUIDE TO PRIVACY LEGISLATION (AVAILABLE THROUGH THE STUDY WEBSITE)..... 199

FIGURE 21: WORLD MAP IN THE PUBLIC HEALTH GUIDE TO PRIVACY LEGISLATION; SELECTABLE COUNTRIES - CURRENTLY CANADA ONLY - APPEAR DARKER..... 199

FIGURE 22: SUMMARY OF RELEVANT PRIVACY LEGISLATION FOR CANADA AS SHOWN IN THE PUBLIC HEALTH GUIDE FOR PRIVACY LEGISLATION TOOL ..... 200

FIGURE 23: EXAMPLE OF A USER INTERFACE FOR A TOOL TO IMPLEMENT THE MULTIDIMENSIONAL POINT TRANSFORM ..... 202

FIGURE 24: MULTIDIMENSIONAL POINT TRANSFORM TOOL INTERFACE DESIGN AS ORIGINALLY PUBLISHED SHOWING EXAMPLE SETTINGS..... 204

FIGURE 25: PROTOTYPE TOOL AVAILABLE THROUGH THE STUDY WEBSITE SHOWING AN EXAMPLE OF HOW A PRACTICAL IMPLEMENTATION OF THE FRAMEWORK MIGHT LOOK..... 205

FIGURE 26: ASSESSMENT WORKSHEET OF THE MICROSOFT EXCEL WORKBOOK DEVELOPED AS AN EXAMPLE IMPLEMENTATION OF THE DATA-SHARING FRAMEWORK. .... 206

FIGURE 27: THE ADMINISTRATION WORKSHEET OF THE MICROSOFT EXCEL WORKBOOK DEVELOPED AS AN EXAMPLE IMPLEMENTATION OF THE DATA-SHARING FRAMEWORK ALLOWING USERS TO SET DOMAIN AND DIMENSION WEIGHTS AS WELL AS ASSESSMENT SCORING THRESHOLDS..... 207

## **DECLARATION**

The author declares that at no time during registration for this research degree of Doctor of Philosophy was he registered for any other University award.

This study was partially financed and supported by the Public Health Agency of Canada.

During the course of this research, the author was invited to participate in collaborative work with other groups and authors. Research arising from such collaborations was informed by the author's current degree work and is clearly identified as such throughout the thesis. Similarly, original work and publications arising directly from this work are identified and were conceived and completed in their entirety by the author.

Relevant publications, presentations, reports, awards, grants and activities are identified below. Please note that workshops, seminars and meetings attended by the author in which he was a participant and not a presenter are not listed.

In addition, a Website was created for the study and can be found at <http://www.personplacetime.org>. Prototype tools developed through the course of this work can be found through this site and are referenced and described herein where relevant.

The University of Plymouth is hereby granted permission to allow the thesis to be copied and or distributed, in whole or in part, for academic purposes, subject to the acknowledgement of the author.

## Document & Achievement Summary

<b>Document Summary</b>	
<b>Word count:</b>	
Abstract:	298
Main body of thesis:	53,676
Appendices <sup>1</sup> (Volume II):	52,220
<b>Total Word Count<sup>2</sup>:</b>	<b>110,120</b>
<b>Number of tables:</b>	13
<b>Number of figures<sup>3</sup>:</b>	27
<b>Number of references:</b>	204

<b>Relevant Achievement Summary</b>	
<b>First-author publications</b>	2
<b>Original published contributions</b>	2
<b>Conference Abstracts</b>	1
<b>Dedicated Presentations</b>	6
<b>Related Presentations</b>	>12
<b>Co-authored publications</b>	3
<b>Awards</b>	1
<b>Grants (co-investigator)</b>	1
<b>Conferences co-chaired</b>	2
<b>Research Reports</b>	2
<b>Advisory Roles</b>	2

**Study Website:** <http://www.personplacetime.org>

---

<sup>1</sup> Does not include word count for Appendix H (published manuscripts)

<sup>2</sup> Does not include references or Appendix H (published manuscripts)

<sup>3</sup> Does not include four Appendix A figures

## **AUTHORED WORK ARISING DIRECTLY FROM THIS RESEARCH**

### **Peer-reviewed publications**

**AbdelMalik P**, Boulos MNK: Multidimensional point transform for public health practice. *Methods of Information in Medicine*. (In press; ePub ahead of print available online)

<http://dx.doi.org/10.3414/ME11-01-0001>

**AbdelMalik P**, Boulos MNK, Jones R: The Perceived Impact of Location Privacy: A web-based survey of public health perspectives and requirements in the UK and Canada. *BMC Public Health*, 8:156 (2008)

<http://www.biomedcentral.com/1471-2458/8/156>

### **Conference Papers/Abstracts**

AbdelMalik P. The Impact of Privacy on Public Health Practice. 19th IUHPE World Conference on Health Promotion & Health Education. IUHPE, Vancouver, BC, Canada (2007)

### **Presentations**

1. Is Privacy Good for the Public's Health?  
Presenter, Public Health Agency of Canada Lunch and Learn Session, Ottawa, ON, Canada (May 7, 2008)
2. Is Privacy good for the Public's Health?  
Panellist, *Privacy, Security and Technology – Affirming Our Rights*. Riley Information Services Seminar. Ottawa, ON, Canada (March 31, 2008)
3. The perceived impact of location privacy: a web-based survey of public health perspectives and requirements in the UK and Canada  
Peninsula Postgraduate Health Institute / Peninsula College of Medicine and Dentistry Annual Research Event. (March 14, 2008)
4. Multidimensional Epidemiological Transformations  
Spatial Analysis and Space Statistics Day, University of Montréal, St. Hyacinthe. (March 5, 2008)
5. AbdelMalik P, Boulos MNK, Jones R: The Perceived Impact of Privacy: Public Health Perspectives and Requirements in the UK and Canada. Access to Information and Privacy Division, Health Canada (2007)
6. AbdelMalik P, Assessing public health perceptions on the impact of privacy. Access to Information and Privacy Division, Health Canada (2006)

Over a dozen additional presentations on spatial epidemiology have been delivered by the author in which material informed directly by this research was used due to its relevance and importance to public health practice.

### **Awards**

Public Health Agency of Canada – 2007 Most Promising Researcher Merit Award  
Awarded in recognition of research with significant impact in the related fields of front-line public health practice and future trends for privacy protection within Canada. Awarded June 2008.



## CO-AUTHORED WORK INFORMED BY OR ENGAGED IN AS A DIRECT RESULT OF THIS RESEARCH

### Peer-reviewed publications

Khaled El Emam, Ann Brown, **Philip AbdelMalik**, Angelica Neisa, Mark Walker, Jim Bottomley, Tyson Roffey: A method for managing re-identification risk from small geographic areas in Canada. *BMC Medical Informatics and Decision Making*. 10:18 (2010)

<http://www.biomedcentral.com/1472-6947/10/18>

Maged N. Kamel Boulos, Andrew J. Curtis, **Philip AbdelMalik**: Musings on privacy issues in health research involving disaggregate geographic data about individuals. *International Journal of Health Geographics*. 8:46 (2009)

<http://www.ij-healthgeographics.com/content/pdf/1476-072X-8-46.pdf>

Khaled El Emam, Ann Brown, **Philip AbdelMalik**: Evaluating predictors of geographic area population size cut-offs to manage re-identification risk. *Journal of the American Medical Informatics Association*, 16:256-266 (2009)

### Research Reports

1. **A Method for aggregating small geographic areas to protect privacy**

Khaled El Emam, Fida Dankar, **Philip AbdelMalik**

Research report produced for the Knowledge, Information and Data Systems Division, Office of Public Health Practice, Public Health Agency of Canada. March 2009.

2. **Uniqueness in the Canadian population**

Khaled El Emam, Ann Brown, **Philip AbdelMalik**, Angelica Neisa

Research report produced for the Knowledge, Information and Data Systems Division, Office of Public Health Practice, Public Health Agency of Canada. January 2009.

### Conference Involvement

**2009 Electronic Health Information and Privacy Conference**

November 19, 2009 – Ottawa, Ontario, Canada

**Session Chair:** Privacy Considerations in Disease Surveillance

**2008 Electronic Health Information and Privacy Conference**

November 3, 2008 – Ottawa, Ontario, Canada

**Conference co-chair**

### Advisory Roles

1. GeoConnections Federal Privacy Advisory Committee, Natural Resources Canada. 2007-2009

Public Health Representative on behalf of the Public Health Agency of Canada

2. GeoConnections Cross-Jurisdictional Privacy Advisory Committee, Natural Resources Canada. 2007-2009

Public Health Representative on behalf of the Public Health Agency of Canada

## Grants

2009—2013 Canadian Institutes of Health Research; **Facilitating access to health data for research and planning in light of laws and ethical norms**; \$160,529 CDN.  
Principal Investigators: D Willison, K El Emam, E Gibson.  
Co-Investigators: **P. AbdelMalik**, S. Anand, M. Burgess, A. Dawson, C. Emerson, V. Goel, M. Hadskis, A. Holbrook, M. Loeb, J. McLaughlin, P. Raina, L. Schwartz, V. Steeves

The author hereby asserts that all declarations and statements made above are true and that this thesis and the work described herein are his original work and compilation.

**Thesis Title:** Multidimensional Epidemiological Transformations: Addressing Location-Privacy in Public Health Practice

Author:

Date:

\_\_\_\_\_  
Philip S. O. AbdelMalik

\_\_\_\_\_



## A Personal Word & Acknowledgements

*You have zero privacy anyway. Get over it.*  
Scott McNealy, CEO Sun Microsystems

Pursuing a PhD 5,000 kilometres from one's academic institution while also pursuing a career, attempting to maintain some semblance of a life and simultaneously courting a wonderful woman living over sixteen thousand kilometres away in the last year and half of study has been for me a most...“extraordinary” experience. It began with a zeal bent on solving the world’s problems, and ended with the overwhelming desire to accomplish what was necessary to complete the degree requirements. In between, my appreciation for Charles Dickens' words increased as his introduction to *A Tale of Two Cities* took on renewed personal meaning:

*It was the best of times, it was the worst of times...*

It is amazing to look back and reflect on the history that has been written in the five years it took me to complete this work; history that I can think of no better way to describe than some of the best and worst of times: the 2009-2010 H1N1 pandemic, the 2010 winter Olympics in Vancouver, the G8/G20 in Muskoka and Toronto in 2010, all of which I was involved with in my capacity as an epidemiologist with the Canadian federal government; the meltdown of the economy in 2009, and its slow, irregular and continuing “recovery”; dramatic work changes and career decisions; the pain of a breakup and the joy of an engagement and imminent marriage to a wonderful and supportive woman...I watched with concern as floods and cyclones of historical significance battered her country of birth, and with interest and apprehension as democratic revolution prevailed in mine; Australia saw its first ever female prime minister and the United States its first black president. I could go on, but suffice it to say that much history has been written in these past few years...but perhaps none that has brought as much relief, satisfaction and growth to me personally as the completion of this degree!

The topic for this research was actually forged through real personal experiences. Working in health research, I went from a clinical setting where I had unfettered access to patient data to a government setting in which I had unreasonable access – at least from my perspective. I found the right to “privacy” being used as a catch-all to justify non-release of data, though ironically there often remained acknowledgement of the importance and necessity of the research requiring them. As I looked into the literature and chatted with colleagues, it quickly became obvious that this was a common problem in public health across the globe. What made it even more complex was the increasing use of geography in public health analyses. While most commonly-collected direct identifiers – such as name or health card/insurance numbers – are not relevant to public health research questions, and can therefore be stripped, anonymised or pseudonymised, this is not the case with granular location information. Interactions with the space around us – including environmental factors and socio-demographic attributes – are obviously dependent on the location of an individual and the ability to detect these interactions by the scale of data release. Alter either in any major way and you can cast serious doubts on the validity of the findings. And so began my search to survey and document the issue and existing proposed solutions in order to come up with recommendations and, if endowed with the creative zeal, a novel solution.

The journey proved to be extremely multidisciplinary, and I found myself going beyond my existing expertise and comfort zone of health and epidemiology into what to me at the time seemed like ethereal realms of law, mathematics, physics, computers, geography, ethics, spatial statistics, and of course, politics. At times, the task seemed overwhelmingly daunting, but taking the cue from Charles Darwin, I agreed that *doing what little one can to increase the general stock of knowledge is as respectable an object of life as one can, in any likelihood, pursue*. Now that I am able to breathe again, turn around and look back and reflect on the past five years, I am humbled by the enriching and rewarding experience it has been. I could not have done it without the support and encouragement of a lot of individuals, though at times I wonder if perhaps

it was the lack of energy to quit that kept me going! Either way, the result is the tome you now hold in your hands and with which I am rather pleased, in spite of the fact that there remains much work to yet be done in this area of public health.

It is my sincere hope that the recommendations and ideas presented in this thesis will help facilitate "public health practice" (a loaded and controversial phrase). By allowing access to critical data integrated across multiple domains, we dramatically increase the power and efficacy of the practice, providing quality evidence to correctly and holistically inform policies and decisions aimed at improving public health. In doing so, we improve the quality of the most primal of all rights, the rights to life and health.

I would be remiss to reflect on all that has occurred in the past five years without also reflecting on the wonderful people who made the best of times the best, and the worst of times bearable. To list them all by name would require yet a third volume and a memory far better than mine; even to list those who directly helped me complete this work, be it through as intense an involvement as editing what to them was likely a boring and irrelevant piece of work, or the simple encouragement of a positive smile and compliment. Nonetheless, I would like to acknowledge at least some of those whose patience and encouragement brought me to this point:

Dr. Maged N. Kamel Boulos, my director of studies, who believed in my ability to complete this work, and whose tireless encouragement, patience, faith and prayers kept me hanging on through thick and thin;

Professor Ray Jones, my secondary supervisor, whose encouragement, support and insight brought an invaluable perspective to the work;

Professors Gerard Rushton and Markku Löytönen for providing insightful comments on the survey during the piloting phase and encouragement as I began the journey;

Dr. Gregory Taylor, who generously made it easier for me to juggle both work and the PhD simultaneously;

Dr. Anne Bassett, Dr. Eva Chow, Dr. Janice Husted and Dr. Ian Johnson, who started me on my path to and through epidemiology and initiated me into the world of medical and health research;

Dr. David Mowat, Dr. Michael Goddard and David Lewis for encouraging me to pursue the PhD to begin with, and for providing professional, career and financial support without which I could not have come this far;

Philip Ng and Jennifer Siushansian, whose calm resolve kept me going when career issues were tumultuous, and whose support was absolutely critical in the final stages of this work - I pray that my support to you both has been as comforting as yours has been to me;

Kara Hayne, Jeff Wingeat, Dolon Chakravartty and Alex Hewitt, my work team, who tirelessly put up with me and listened to my PhD woes, but also made work a fun place to be;

Dr. Khaled El Emam, whose kindness and academic insight helped keep me focused and moving forward through the challenges of being over 5,000 kilometres from my university;

Daniel Hynes, whose optimism and sense of humour brightened many days;

Dr. Pascal Michel, whose caring and cheerful attitude kept me motivated as he challenged me intellectually – and with whom I have had the immense pleasure and privilege of working;

Janet Honig, whose insightful discussions and joyful encouragement made me see that there IS light at the end of my tunnel;

Erin L. Schock, who gave tirelessly of her time to encourage me to always move forward;

My parents, who held me close in prayer and offered much encouragement and advice when progress looked bleak and quitting seemed appealing;

And last, but certainly not least, my wonderful fiancée (now my wife), Miriam Rawson, whose love, patience, commitment and comments lit my journey's end; no doubt God saw that my PhD was coming to a close and decided to bless me with the most wonderful congratulatory gift I could have asked for.

As I am humbled looking back over all that has transpired, and all the people who directly or indirectly played a role, no matter how minute, I am deeply grateful for God's blessing and provision – a thread I can follow throughout my life. Ultimately, I have not a shred of doubt that it was His hand that sustained me, at times through some of these people, and I thank Him for his grace that allowed me to persevere – to Him be the glory, for great things He has done...and great things He continues to do.

*Soli Deo Gloria*





# 1. Impetus

## 1.1. INTRODUCTION

The state of public health, it seems, is one of confused conflict.

Public health has been defined as one of society's organised efforts to protect, promote and restore people's health [1]. Unfortunately it is not that organised and in spite of this generally accepted definition, much of what is at its core lacks clear and globally accepted definitions, often times creating confusion. Nonetheless, the contributions of public health to individuals, societies and nations are tangible, plentiful and significant, ranging from outbreak response and investigation to vaccination programs, prenatal care to smoking cessation, substance abuse to injury prevention, disease surveillance to risk factor analysis and emergency preparedness and response to name only a few. At the core of these activities is a chain of health data, starting with the building blocks at the micro, or individual, clinical level, and ending at the macro, or population level. Without the starting point – without the micro health data – public health would be impossible to do. Yet the acquisition of this data – so fundamental to the "practice" – is arguably in conflict with another fundamental human right.

For decades, and in spite of their interdependence on one another [2], the debate has raged between the fields of privacy – an acknowledged human right and evolving "principle as old as the common law" [3] – and public health [3]. So much so, in fact, that one sometimes cannot help but wonder if privacy is, indeed, the enemy of public health [4] and whether the two could ever peacefully co-exist [5]. With e-health already a reality in countries like Canada, the United Kingdom and the United States, and as information giants continue to pursue their stake in health information [6,7], privacy continues to become an increasingly critical concern. Catalytic to this concern is the increasing use of Geographic Information Systems (GIS) – and therefore the incorporation of place – in public health.

A key discipline at the heart of public health is epidemiology – "the study of the distribution and determinants of health-related states or events in specified populations and the application of this study to the control of health problems" [1]. The three fundamental pillars of descriptive epidemiology are *person*, *place* and *time*. Traditionally, the privacy debate has revolved around the first of these pillars – person. This has resulted in a variety of anonymisation techniques in public health. Numeric codes are assigned, names and other identifiers are stripped or abbreviated, k-anonymity [8] and other techniques are applied, and the privacy issue is addressed. The problem with place is that, at its most granular – and arguably useful – scale, it identifies us. The greater the level of geographic detail one has regarding an individual, the more readily that individual can be identified. Consequently, the acquisition of geographic data tends to be either limited, or at a sub-optimal or unusable scale [9,10]. Not only do privacy issues impact data acquisition and use for analysis, but also visualisation and dissemination of the results. Researchers have been able to "reverse engineer" maps, for example, to successfully re-identify individuals [11-13].

While some argue that this debate is the product of a lack of understanding of the legislation and regulations by the public health community [2,14], there has been no formal collection and synthesis of the corresponding views and perspectives of those directly involved in public health activities. Although some solutions have been developed [8,10,15-18], they not only generally result in a concerning loss of data quality, but none to date are comprehensively adaptive, adjusting for important public health dimensions such as age and sex in concert with location. Lastly, there is currently no framework to guide public health professionals, custodians and research ethics boards in the appropriate assessment of the privacy implications - particularly including location privacy - for data-sharing. This research addresses all three of these issues: a public health survey was conducted and the findings published; novel solutions were explored with several being co-authored and a novel dynamic, adaptive

algorithm published; and a conceptual framework with a prototype application has been developed and is presented.

## **1.2. CROSS-NATIONAL VALIDITY**

Cross-national research provides a rich source of data and information, allowing countries to learn from the methods, successes, and failures of one another. It is not uncommon to find cross-national comparisons of health care systems around the world – particularly between member countries of the Organisation for Economic Co-operation and Development (OECD). Within World Health Organisation (WHO) and Commonwealth Fund reports on health care systems, four countries repeatedly appear in comparisons: Australia, Canada, the United Kingdom, and the United States [19-22]. When the Naylor Report was completed in response to the 2003 SARS outbreak in Toronto, Canada, a comparison of international systems focused solely on these countries, because it was felt their “organisation and governance of public health to be particularly informative” [23]. Comparisons between various aspects of public health and privacy have therefore been included, where applicable, for these countries, as well as, where appropriate, the European Union, the WHO and the OECD.

## **1.3. RESEARCH OBJECTIVES**

The objectives of this research are to contribute to the resolution of the public health-privacy debate by:

1. Reviewing privacy legislation as it pertains to place and public health in Canada, the UK, and various other countries around the world;
2. Formally collecting and synthesising the perspectives and requirements of public health professionals in Canada and the UK on the current issue, with a focus on the role of place;
3. Exploring the development of novel techniques to allow spatial public health analysis at a granular level without compromising privacy;

4. Developing a conceptual framework to guide public health practice in the appropriate evaluation of the privacy implications of data-sharing with a particular emphasis on location-privacy.

In pursuing this work, this research heeds the warning of Curtis *et al.*: "...health and spatial scientists should be proactive and suggest a series of point level spatial confidentiality guidelines before governmental decisions are made which may be reactionary toward the threat of revealing confidential information, thereby imposing draconian limits on research using a GIS." [11]

#### **1.4. ORIGINALITY AND CONTRIBUTION**

The originality and contribution of this research can be found in the last three of its objectives as stated above:

1. The formal collection and synthesis of public health perspectives and requirements on the issue of privacy – including location privacy – and public health. This is the first such survey of its kind and was published in *BMC Public Health* in 2008. This survey contributes British and Canadian perspectives to the body of research and knowledge.
2. The exploration of the development of novel techniques to allow spatial public health analysis at a granular level without compromising privacy. To address this, a novel multidimensional algorithm was developed and has been accepted for publication in *Methods of Information in Medicine*. The algorithm has been named the "MPT" - Multidimensional Point Transform. It is a dynamic, adaptive algorithm that addresses many of the recognised deficiencies in already existing techniques as discussed in the manuscript. In addition to this novel technique, direct contributions were made to three related original approaches, two of which have been published in *BMC Medical Informatics and Decision Making* and the *Journal of the American Medical Informatics Association* and the third of

which was submitted as a report to the Public Health Agency of Canada and is currently being prepared for peer-review submission.

3. The development of a conceptual framework to guide public health practice in the appropriate evaluation of the privacy implications of data-sharing with a particular emphasis on location-privacy. Currently, no such framework has been found.

In addition to the above, a summary of legislative findings was authored and published in the *International Journal of Health Geographics*. A prototype tool or tool idea has also been developed to facilitate each of the legislative, transform and framework aspects of this work as they apply to public health and privacy, with particular emphasis on location privacy, offering some contribution to the future development of such implementations.

The concepts and overall findings need not be limited to any particular country or health event, and have the potential to promote further research into more complex and comprehensive functional analyses involving the complete epidemiological triad. This research is unique and innovative in that it takes a holistic epidemiological approach whilst building on existing and novel technologies and concepts.

## **1.5. STRUCTURE & ORGANISATION**

The thesis is organised into two volumes. Volume I contains the body of the research, composed of five parts reflecting the three original aspects of the study as stated above, future developments and the references:

**Part I** sets the stage by identifying the issues underlying this study. It establishes common ground for definitions relating to public health practice and privacy, reviews privacy concepts and legislation as related to location and public health, and describes the results of a novel public health practitioner survey and its findings.

**Part II** focuses on concessions that attempt to reconcile public health practice with location privacy concerns by reviewing existing solutions, describing contributions to novel applications through collaborative efforts, and ending with a detailed description of an original novel algorithm and its preliminary implementation.

**Part III** is dedicated to the development of a proposed generalisable novel framework for guidance on data-sharing, particularly in light of location privacy issues, and gives a sample implementation using a fictitious scenario.

**Part IV** concludes the thesis body, describing suggested prototypes for implementing some of the research, proposing future research ideas informed by the current work, reviewing the objectives and accomplishments, and concluding the study.

**Part V** contains all references.

Volume II forms **Part VI** of this work and contains the appendices. These include a description of the survey logistics and business specifications, original copies of the surveys, the full results of the survey findings, the Multidimensional Point Transform algorithm code (written in SAS), the preliminary code developed for creation of a synthesised Canadian population using census data (also in SAS), and copies of all publications and reports arising from this work, either directly or in collaboration with others.

It is my hope that you find this study informative, inspiring, enjoyable to read and, quite possibly, even useful to public health.

**PART I**  
**THE BACKDROP**





## **2. Establishing Common Ground**

### **2.1. INTRODUCTION**

Before delving into the sort of complex multidisciplinary discussion required by this study, it is necessary to first establish some common terminology and definitions on which to stand as well as set the backdrop against which this thesis unfolds. Chapter 1 began by stating that public health is seemingly in a state of confused conflict. No doubt this begged the questions: what is the confusion and where is the conflict? While these were briefly described, a more thorough explanation is required for concepts fuelling the confusion and conflict form the foundation on which this study is built.

The confusion referred to is at the very heart of public health and has to do with the very definition of the "practice" of public health and the identification of its core or essential functions. The conflict is equally as concerning and critical since it revolves around the perceived clash of two fundamental human rights: the right to privacy and the right to health. The right to privacy has obvious restrictive implications for data-sharing. Data-sharing, however, is a fundamental requirement for public health. And so it is that the two continue to lock horns with increasing concerns on both sides and no clear resolution in sight.

### **2.2. PUBLIC HEALTH "PRACTICE" & "FUNCTIONS"**

Since the entire content of this study is intended for use in public health practice, it behoves us to define the phrase before we start using it everywhere. The phrase "public health practice", much like the focal "privacy" issue at hand, lacks a globally accepted or consistent definition. While its overall goals are more or less universally acknowledged - the improvement and protection of the health of individuals and communities - the methods and functions implicated by the word "practice" are not, particularly where research is involved [1]. Some define public health practice in terms of its vehicles, as both a science and art that includes education, promotion, research, intervention and prevention [24]. Others argue that a distinction must be made between

the direct practice itself as a legally authorised and ethical duty and the research which serves to inform it [25,26]. Others still have defined it in terms of functions, further differentiating between those essential for and unique to public health and its essence, and those that establish, maintain and enable public health to operate [27]. In many cases, frameworks and guidelines have been developed to attempt to clarify some of these distinctions [25,27,28], but in the absence of common definitions, such guidelines are region-specific at best and no global or definitive approach exists.

Contributing to the ambiguity of the “practice” is the similar lack of agreement on the core or essential functions of public health (sometimes referred to as EPHF - Essential Public Health Functions) and even how to define the term. Some define it in terms of the “conditions” required to facilitate and improve public health practice [29], while others in terms of “activities” undertaken to achieve its objectives [30]. To illustrate the inconsistent approaches, consider the following: in 1997, the World Health Organisation (WHO) in consultation with over 130 public health experts from around the world identified 37 functions grouped into nine EPHF categories [29,31]. A few years later, the Pan American regional office of the WHO (PAHO) added two more to make them eleven [29] while the Western Pacific regional office (also of the WHO) made some slight modifications but kept them at nine [32]. The original nine did not include research as an EPHF; the two revisions mentioned above did. The United States, operating under a different definition, identifies only three overarching public health functions (assessment, policy development and assurance) attained through ten essential “services” which do include research [33,34]. Canada identifies six core functions [30] based on the recommendations of the post-SARS Naylor report [23], while Australia subscribes to nine [35] and the UK to ten (the tenth being identified as quality assuring the other nine functions) [36]. In the case of both Canada and Australia, research is not considered to be among the core functions, whereas in the United Kingdom it is. It is no wonder that public health is struggling with the concept of privacy and personal control over health information – we cannot even universally

agree on a concise set of the core functions that define it! A comprehensive comparison of the functions across groups is beyond the scope of this research as it is complicated by conflicting terminologies and inconsistent sub-categorisation of services and activities. However, a useful overview is provided by the Ministry of Health Services for the Province of British Columbia, Canada, in their work to develop their own framework for core functions in public health [27].

The current study adopts a broad definition of the phrase "public health practice" to reflect the range of activities and tasks necessary to achieve its goal (i.e. improving and protecting the health of individuals and communities). This therefore captures core public health functions *as well as the activities required to perform them*. The latter include research, which is therefore considered to be a part of public health *practice* though not a public health *function*. Another example which is perhaps more obvious is immunization; while it contributes to the function of prevention and control of disease, it is not in and of itself a core function of public health. Yet it is a vital component of public health practice. By adopting this wide definition, this study allows relevance and application to whatever definitions and distinctions may be made with regards to functions, services, activities and conditions that together make up the overall practice. It is important to note that within this one all-encompassing word, "practice", the ability to differentiate between types of activities is important, as they may have legal, ethical, political and procedural implications. For example, classifying an activity as research ordinarily requires ethics approval by a review board or committee; surveillance of some diseases may be mandated by law; and preparedness and response for emergency situations and public health threats are necessary governmental and societal responsibilities.

Since many uses of the work of this study will likely revolve around research, it merits a brief commentary. While a discussion on research methodology is beyond the current scope, it should be noted that there are different kinds of health research and study

designs, such as population-based research, targeted clinical trials, retrospective studies, longitudinal studies and so on. As will be clarified and emphasised throughout, this current study does not apply to all research scenarios, but rather research involving secondary use of personally identifiable data in the absence of consent. This often leads to debates on privacy, which invariably leads to the next set of terms requiring definition.

### 2.3. PRIVACY & PERSONALLY IDENTIFIABLE DATA

*The problem with the word privacy is the term itself. Everybody has their own idea of what it means, which allows the courts to stretch from contraception to abortion to whatever.*

Taken from <http://davidboyd.org/posts/1131586492.shtml>

The concept of something being “private” is embodied in ownership, engrained in creation, and evolving through history. We speak of various species being territorial, and as humans we set up fences and mark our geographical boundaries, reflecting the concept of private property. We have clear definitions of what is “mine” – indeed, one need only spend a few minutes or hours with a child to observe this in action. However as suggested above, and just like public health practice and functions, privacy is a complex, multifaceted construct that is heavily context-dependent and inconsistently defined. It can refer to personal opinions, communication, actions, information, space or physical property. Therefore, when discussing privacy, we must ask “privacy of what?”

Today, privacy is generally differentiated from the concepts of confidentiality and security. Privacy and security are often discussed hand in hand, but by mitigating either one does not necessarily mitigate the other. In order to establish common ground, a distinction is made between various terms as used throughout this study:

**Personally Identifiable Data:** This generally refers to any information that can be used to identify an individual and it is within this context that privacy is used within this study. A review of personally identifiable information as a component of privacy legislation is provided in Chapter 4.

**Privacy:** This reflects the ability of an individual to control the collection, use, retention, disclosure and destruction of his or her personal information. It has been defined as a right in various legislative pieces.

**Confidentiality:** Distinguished from privacy, this is not a right but rather an obligation of a second party to respect the privacy of an individual concerning his or her personal information. This may be at the wish of the individual to whom the data pertain, or guardians, representatives, etc. responsible for the individual (e.g. in the case of children or patients who are incapable of making decisions for themselves).

**Security:** This refers to the protocols and procedures put in place to safeguard the confidentiality of personal information and respect the privacy of the individual.

Now that the main terms have been defined, let us clarify the nature of the conflict before probing more deeply into some of the background and legislation surrounding privacy.

## 2.4. PUBLIC HEALTH AND PRIVACY

*There is nothing more public than privacy*

**Berlant and Warner**  
**Sex in Public, Critical Inquiry (1998)**

Issues of patient privacy and confidentiality in public health and medicine are not new [8]; they have been, and continue to be, perceived as a major obstacle for public health research, complicating data mining, sharing, acquisition, and analysis, ultimately impeding accurate, evidence-based decision-making [15,16,23,37-45]. The genesis of much of the health data used for public health purposes begins with the clinician. In many parts of the world, however, this fiduciary relationship between patient and doctor is so revered that it has been compared to that between a parishioner and priest [46]. Yet at what point does it become acceptable to compromise this confidentiality, and by extension privacy, and inevitably trust? Is such a compromise justified when to do otherwise could potentially jeopardize the health of society or a third party? Consider the following example: when presented with the same scenario involving an injured drunk driver presenting to the emergency department in the absence of a police escort, six different physicians gave varying and opposing responses as to whether or not they would breach confidentiality and inform the authorities [46]. The ethical question for the clinician therefore becomes whether the preservation of this confidentiality is subservient to a greater good [46]. In this particular case, the subject has violated a socially defined and accepted norm. While many members of the public may therefore, in this instance, advocate a breach of confidentiality, it seems that when it comes to matters of public health – which, by definition, is concerned with the greater good – most individuals within a society would be less inclined to support the breach. A breach of confidentiality, however, would also constitute a breach of privacy since as per the

definitions established above the patient would have no control over the release of his or her information.

The inability to share information due to privacy and confidentiality legislation was cited as a major obstacle in Toronto's response to SARS in Canada in 2003 [23]. Even critical events relating to national security, such as bio-terrorism surveillance, are negatively impacted by this problem [42]. The issue at hand is not so much the importance of and need for privacy laws and confidentiality agreements – protecting the identity of patients is an acknowledged and recognized necessity [37,40,46]. Rather, it is the constraining effect of such legislation on a practice that, by definition, can only appropriately improve the health of populations by using data on the very individuals it comprises. The perceived good of the individual seems to outweigh the effective betterment of the whole - which, ironically, is intrinsically dependent back on the good of the individual; an interesting contrast to the drunk driving scenario previously presented.

As electronic patient health records become a reality, there is a growing concern that patients may not seek medical attention, or may withhold information from their care-providers, for fear of privacy breaches. While this assumes that the average individual values privacy more than his or her health and the health of the society in which they live, all too often the debate focuses on highlighting the losses people might incur by revealing personal information, as opposed to the gains of sharing this information with the appropriate individuals and/or organisations. It is also important to differentiate between personal information recipients and their underlying agendas or motives (e.g. journalist vs. practitioner). The media, which thrives on sensationalism and society's curiosity of the personal affairs of others, is more likely to attempt re-identification of individuals given any amount of personal information. However, a public health practitioner is less likely to jeopardise his or her career, reputation and ethical duty in this way, and typically has very different motives and objectives.



Further complicating the conflict is the rapid rise of technology and the ubiquitous availability of information. With increased availability comes increased utility – as well as the potential for increased intrusion. This is particularly the case with the explosion of location-based technology. Satellite and aerial imagery dominate Web-based mapping tools, geomatics tools are no longer for the privileged few and the power of today's normal personal computer makes the integration and linkage across volumes of information a growing concern. All one need do is simply open up a popular mapping Website, such as Google Earth, and let curiosity follow its natural course.

## **2.5. FOCUS ON LOCATION**

We are, by nature, curious, nosy and resourceful creatures. Given access to the appropriate tools and/or skill set this can become a recipe for innovation and success, or privacy breaches and lawsuits.

Clara Poole, a young, fictitious and ambitious entrepreneur, decides to open up her own pool business. In order to identify her potential customers, she uses Google Earth to carefully dissect various neighbourhoods. Using the aerial imagery, Clara quickly identifies neighbourhoods and houses with in-ground pools. Armed with this knowledge, she aggressively targets pool-less homes in pool-rich neighbourhoods for new pool projects, and develops neighbourhood-targeted pool-maintenance packages.

Meanwhile, Rob Beri (also fictitious) has just finished using Google Earth to draw up his own plan. Though he has been looking for exactly the same thing as Clara, his intent is completely different. Pool-rich neighbourhoods, he reasons, generally have higher incomes...and homes with pools in those neighbourhoods are particularly more appealing for his line of work – breaking and entering.

As mentioned in Chapter 1, in spite of “place” being one of the three fundamental pillars of descriptive epidemiology - which is at the heart of public health - it is, of the three pillars, the most weakly utilised. This is, however, changing – albeit relatively

slowly. As location-specific information becomes more readily available and analysable, its significance to public health practice is becoming more recognised as it contributes the lacking and important dimension of spatial intelligence and relates health states to the world around us. However, location-specific information can immediately identify individuals: given a patient's age, one has little to work with. Given an address, however, the whole picture changes, and passions in the privacy debate ignite. Even releasing (or not releasing) a large and seemingly harmless geographic area, such as province in Canada, can be court-of-law material [47] depending on what additional information is released or known.

Location information, therefore, while immensely useful and in many cases vital to good public health practice, is also of increasing concern in the public health-privacy world. In their 2004 review of data-sharing and the development of a Web-enabled geographic information statistics service, the South East Public Health Observatory (SEPHO) in the UK cited privacy and confidentiality issues as a major challenge to data-sharing, and the greatest barrier uncovered by a preliminary feasibility study [43]. As will become evident throughout the body of this work, the frustrating experience of the UK's SEPHO is not unique, and the call for the identification and implementation of appropriate solutions must be taken seriously. To do so, one must first gain a better understanding of the legislative privacy landscape that currently frames the debate, and the perceptions and requirements of the public health community as they pertain to privacy, and more specifically, location privacy.

### 3. Privacy Concepts and Principles

*Good people do not need laws to tell them to act responsibly, while bad people will find a way around the laws.*

**Attributed to Plato**

#### 3.1. INTRODUCTION

Viruses, bacteria, parasites...disease agents in general, along with their vectors, come in all shapes and sizes. However, they could not care less about political and administrative boundaries. Indeed, they know no borders, save those that lead to unfavourable environments or conditions...and even then, they do not necessarily *know* this, or opt out of travelling to such places; they simply do not survive if they do. They, and by extension the concept of health and well-being, transcend jurisdictional boundaries. Legislation, on the other hand, which also comes in all shapes and sizes, has traditionally been very much the opposite.

For over a century, political leaders have presented public health as society's most important responsibility. Benjamin Disraeli reportedly stated that "the health of the people is really the foundation upon which all their happiness and all their powers as a state depend". Health, as a state of wellbeing, has been dubbed our "most valuable asset", worthy of vigorous defence [48], and a universal and fundamental human right that is a critically significant world-wide social goal [49]. As the active and fundamental social construct of this right, it is imperative that public health impact and influence the legal system as much as it is impacted and influenced by it. Yet simply introducing legislation to cover any and all scenarios is an impossible task and damaging to society. After all, legislation may be implicated in the moral decay of society.

Laws attempt to infuse society with structure and rules that, to a very large extent, define how we relate to one another. It is no wonder, therefore, that they suffuse every facet and discipline of the world in which we live. The world's response to atrocities and periods of difficulty has typically been legislative. When problems strike fear in the

population, governments respond by instituting laws and enforcing them. Since ancient times, the resolution of issues or disputes has typically involved escalation through some sort of hierarchical structure. In most cases, this is the preferred and most frequently used approach. Eventually, however, if all else fails and as a “last-resort”, legal action is taken. Modernised and enforceable legislation that accurately reflects current practices and future directions is therefore an important societal requirement. However, the proliferation of legislation also has a negative side effect: it begins to create a prescriptive method for living, thereby slowly eroding our ability to think and act in an ethical and moral manner. When a question or issue arises, our response is not “what is the right thing to do” but rather, “what is the legal thing to do”; more specifically, what does the legislation say?

The role of the legal system in the field of public health has been the driving force behind this study. More and more, particularly with increasing emphasis on individual rights and concerns, public health practitioners around the world are recognising the importance of having some understanding of the legal system, and a working relationship with the legal profession [2]. Unfortunately, the relationship typically tends to be unidirectional. Just as privacy is a multifaceted and complex concept, so too is the required collaboration resulting from the interdependency of public health and legislation. And yet, in this particular area, the legal profession has not fully recognised the interdependence of the two fields [2]. After all, a sense of control over one’s own information brings with it a certain peace of mind that is an important contributor to psychological well-being.

### **3.2. PSYCHO-SOCIAL DIMENSIONS OF PRIVACY**

*We must know you to serve you*

The psycho-social dimension of privacy is influenced by the prevailing societal and individual philosophical perspectives. As defined within this study, privacy has become such an important personal concept because of the sense of control it affords the

individual and the view that it is a precondition for equality and social justice [50]. While the issues of consent are outside the scope of this study, it is common practice for any research involving human subjects to first seek the consent of the subjects for the collection, use and disclosure of their personal information. And indeed, based on the literature, most individuals would not object to the use of their information for health-specific purposes [14,51]. In these cases, however, individuals have been given the power to choose what to disclose and to whom. But they are very largely powerless when it comes to the health event itself; pathogens do not seek consent prior to infection. So in a sense, while people do have some degree of control over the determinants of their health (such as exercise, smoking, etc.), health in and of itself is dictated by a hotchpotch of controllable and non-controllable factors.

In his review of the history and importance of privacy legislation in civil society [50], Martin Lengwiler of the University of Zurich suggests that there has been a move from defining privacy in negative terms in the nineteenth century, to a more positive one in the twentieth century. He relates this to a change in paradigms, contrasting libertarianism and communitarianism (or what he refers to as a sociological paradigm). Libertarian views prioritise the rights of the individual as superseding those of a group, society, nation, religion, ideology, etc. This is in direct contrast with communitarian views which focus on the responsibility of the individual to communities and societies when addressing ethical questions. Similarly, privacy has been framed within a political context, with distinction being made between historical “paternalistic privacy” and more recent “democratic privacy”. While the latter facilitates the libertarian paradigm, it does not exclude communitarian ideology; giving individuals the fundamental democratic right to own and control their own information simply places the power to choose between individual and societal good in the individual’s hands. Legislation must therefore counter-balance this power by considering the cost of privacy protection; failing to do so and focusing solely on legitimising the power of the individual can have dire consequences for individuals and society [52].

Another changing social consideration is the relative terminology used in the legislation. In the US, for example, there is a *reasonable* expectation of privacy. Reasonable is very much individually, socially and culturally defined; what is reasonable to one individual may not be so to another, and differs between countries. Even within a nation, “reasonable” can take on different meanings within different contexts. Wire-tapping a suspected terrorist, for example, may be justified since the suspect does not have a reasonable expectation of privacy; or perhaps more accurately, the breach of the person’s privacy is not unreasonable. There is unfortunately no clear definition, however, of what would be considered a reasonable context for “breaching” an individual’s privacy.

It is obvious that differing psychological, societal and cultural norms and expectations play a major role in the concept of privacy, making it necessarily context-dependent. “Privacy laws are most burdensome and least effective when they apply broadly, without proper concern for the settings in which they operate, the types of information that they cover, the obligations that they impose and the purposes they were designed to serve” [52]. Unfortunately, in spite of an increasingly globalised world, it is difficult to transcend the jurisdictional boundaries and create universally acceptable context-specific and public health sensitive privacy legislation. Before embarking on a whirlwind tour of the legal landscape, however, let us pause, reflect on and consider the disharmony that currently exists in the concept’s underlying principles.

### **3.3. PRIVACY AND INFORMATION PRACTICE “PRINCIPLES”**

In the late nineteenth century, privacy invasions were perceived to primarily be the result of proliferating media attention as suggested by Samuel D. Warren. In his 1890 seminal article in the Harvard Law Review, Warren suggests the natural development of legislation as “man” becomes more aware of the importance of various faculties [3], culminating in his focus on the necessity of privacy legislation – privacy then being defined as the “right to be left alone”. Warren’s passionate argument is made in light of

increasing invasions of privacy by the media extending beyond idle gossip to “satisfy a prurient taste” and “occupy the indolent”. Since then, definitions of privacy have evolved to capture more than just the right to be left alone, but rather a prime social value and human right encapsulating “the claim of an individual to determine what information about himself or herself should be known to others” [53]. It was this emphasis on control by the individual that formed the basis for the development of fundamental privacy “principles”.

In 1973, the United States Advisory Committee on Automated Personal Data Systems in the Department of Health, Education and Welfare proposed five principles which later underpinned the US Privacy Act: no personal data record-keeping systems whose existence is secret (transparency), availability to an individual to find out what information is in a record about him or her, and how it is used (access), ability of an individual to prevent personal information that was obtained for one purpose from being used or made available for other purposes, without the person’s consent (use limitation), ability of the individual to correct or amend a record (correction, data quality), and the assurance of the reliability of the data and prevention of its misuse (data quality, security) [54]. Of these five, three very clearly emphasise the concept of control by the individual to whom the data pertain.

Since then, privacy frameworks, guidelines and legislation have pivoted around the continued creation and expansion of such principles – also referred to as Fair Information Practice Principles (FIPPs) – intended to offer a transparent foundation and set of governing concepts. As described in the recommendations issued by the Organisation for Economic Cooperation and Development (OECD), the implementation of these principles serves two functions: to preserve the right of individuals to privacy as a fundamental human right and to provide harmonised guidelines to facilitate national data flows [55]. However, despite the underlying right to privacy being recognised as a fundamental one, accepted and adopted as a standard in many

countries, the principles have remained vague and variable, creating multiple recipes with varying ingredients. For example, eight general and overarching principles are espoused by the OECD [55] compared to five by the Federal Trade Commission [56] and nine by the Asia-Pacific Economic Cooperation group [57] – though all three have overlapping content. National legislation has been enacted around such principles and different business areas have also made their contributions, but they remain inconsistent. For example, the American Institute of Certified Public Accountants (AICPA), together with the Canadian Institute of Chartered Accountants (CICA), have adopted ten Generally Accepted Privacy Principles (GAPP, not to be confused with GAPS, Geographic Area Population Size, which is a concept used in re-identification risk minimisation [58] and discussed in more detail in Part II of this study) to govern business practices [59]. It is worth noting in this context that these principles do not all map directly to those of the relevant Canadian federal legislation, the Personal Information Protection and Electronic Documents Act [60], although there is some overlap. Comparisons of and comments on some of these and others have been published [54,61], demonstrating a lack of standardisation, differing interpretation and resultant difficult translation into implementable requirements.

There are two fundamental issues that should be brought to light with regards to “Fair Information Practice Principles”. The first is simply a question around “fairness” and the second lies in the expensive and illusory “devolution” of their application.

The question of fairness is an important one for public health. The development of these principles is anchored in the libertarian rights of the individual, particularly to control his or her own personal information, hence principles around consent, accessibility of the information to the individual, the ability of the individual to correct the information and to limit or restrict its use. But in much of public health practice, such principles are often impractical and un-implementable since much of public health is concerned with secondary use of clinically collected data; use that is not detailed at the



time of data collection. As such, much has been published on the impracticability of consent and the potential biases and implications to research when individuals are given such control [62-66]. Given a requirement for such data for public health practice, does this therefore make the use of such data in public health – and therefore the very application of public health – “unfair”? The issue lies in the universal application of this control as being a “fair practice”. Furthermore, this application focuses exclusively on the individual’s right to privacy in absolute isolation; it is not a comprehensive approach that takes into consideration other fundamental rights and societal implications. But privacy is not an “absolute right” [67]. Public health is the societal effort to protect, promote, restore and maintain health [1], both for society and for the individual [68], which itself also happens to be a fundamental human right. In the absence of good public health practice, including measures and controls, we significantly compromise our individual rights to health, life, security, freedom of movement, peaceful assembly and leisure, to name a few. Imagine the implications if we had no societal means to detect and control infectious diseases, monitor and warn of the health effects of radiation, respond to and control outbreaks or simply advocate for and promote healthier lifestyles. The identification of these “fair information practice principles” as necessary for the right to privacy has ignored the interplay between these rights; it has promoted them within a vacuum and reduced them to mechanically independent silos as opposed to meaningfully interacting ethical values for the holistic good of individuals and the societies they compose.

The related issue of the devolving application of these types of principles – that is, increasing individual control of and access to personal information – is that we have become reliant on legislation within an increasingly restrictive paradigm of control-based data protection. As local and global information flows increase, so, too, do privacy risks. We attempt to compensate by implementing and revising regulative control mechanisms through globally-inconsistent directives and legislation. In doing so, we increase costs and bureaucracy whilst giving the illusion of individual control

through notices and lengthy consent forms [54,62]. Meanwhile, we are crippled by overemphasised potential costs, the potential benefits are not realised, and as the underlying issues remain unaddressed, we find ourselves having to re-visit and update the regulations. To borrow from and rephrase Fred H. Cate, we have become so enamoured with a control-based approach that we are blinded to the need for developing better alternatives [54]. But the need to alter our approach is critical, and the current methods, whilst emphasising only one aspect of the individual's rights, fail to address real and context-specific issues. They also provide an easy scapegoat for prioritising one right at the expense of others. In today's socially-networked world of instant access to an overwhelming amount of linked information, these "fair information practice principles" are, for public health purposes, out of date, inappropriate, unrealistic, and detrimental. Either that or we are forced to label public health as an "unfair" practice.

As stated earlier, we have moved from a right to be left alone by the media to a right of control over one's information in all circumstances. However, using this definition leaves little room for privacy in the world of public health. Public health is not interested in "proclaiming from the housetops what is whispered in the closet", or the "evil of the invasion of privacy by the newspapers" [3], but neither can it operate under the mercy of individual control. More specifically, public health is typically more concerned with context than with identity; it does not care that John Doe specifically has a particular condition, unless that condition poses some grave risk to John Doe or to others. Rather, public health is more concerned with the fact that a 32 year old male in a particular neighbourhood with specific demographics has a given condition at a given time, and the surrounding context in which that condition occurs (including others with and without the same or other conditions). It is these characteristics that form the pillars of the descriptive epidemiological triad and not the actual identity of the individual. An effective and efficient public health system is neither privacy-centric libertarian nor socialist communitarian. Rather it must be understood and framed as a means for

achieving a harmonised balance between societal protection (“the whole”) and the general protection of the individuals of which that society is composed (“the parts”); the two are interrelated and inseparable. Within such a construct there is no room for individual consent or control over personal data, but there is a role for its ethical use. Public health practice should not be seen as a foe of privacy protection but rather the latter as a necessary component of the ethical pursuit of the former. In many scenarios, this will mean that either privacy will have to be implemented at the potential expense of public health, or public health at the potential expense of privacy. Which potential outweighs the other depends on a variety of factors, and it is these factors that are at the heart of this study.

It is important to emphasise that public health is not being used to justify or suggest a dilution of the right to privacy. Privacy proponents suggest that public safety and national security have been used as excuses to do just this [67] just as the argument is being made that privacy has also been used as an excuse to impede these activities. However, as stated, the clashing of rights is inevitable and some must sometimes be prioritised at the expense of others. To deny this would be to deceive ourselves, and the development and implementation of appropriate solutions requires that we recognise and accept this. Threats to privacy have certainly grown over the years, but so too have threats to health, life, security and safety.

This study therefore, and in particular the framework proposed in Part III, attempts to burrow down to the underlying concepts that instil fear over misuse of data and individual identification rather than implement incongruous control-centric privacy principles. The approach is not an attempt to reconcile the free flow of information with individual privacy and data protection concerns through regulatory control, but rather a drive towards governance of the necessary sharing of information based on a balanced and ethical approach to individual and societal good.

## 4. International Whirlwind Tour of Legislation

*Original contribution as published in the International Journal of Health Geographics (2009) [69]*

A discussion on location privacy solutions for health research would be incomplete without reflecting on some of the underlying reasons that necessitate their development. The very notion of privacy is itself a complex fabric of interwoven philosophical and psychosocial threads. Perhaps this is why the associated bureaucratic and legal landscape is as complex as it is – and often blamed for the issue. A large majority of public health professionals consider privacy to be an obstacle to public health; when asked for the underlying reasons, survey respondents in Canada and the UK most commonly identified bureaucracy and legislation [70].

There is no universal legislation to guide and govern the activities of public health professionals, particularly where issues of privacy are concerned. Instead, nations have their own constraining or enabling privacy and data protection laws, with some being such a maze of cross-referenced “legalese” that familiarising oneself with them – let alone gaining a thorough understanding of them – becomes a daunting task. What ensues is a brief compilation and comparison of relevant personal information and privacy legislation in Canada and the United Kingdom (UK), with particular focus on location and public health as seen and understood by an epidemiologist.

### 4.1. OVERVIEW

The Canadian privacy-legislation landscape is additionally muddled by its political system: ten provinces and three territories, each with its own legislation and jurisdiction over its own health system. Overarching is the federal government, providing guidelines, support, oversight and funding. Although the words “privacy” and “personal information” do not occur anywhere in Canada’s Constitution (Charter of Rights and

Freedoms) [71], Section 7, granting the right to life, liberty and security, and Section 8, guaranteeing protection from unreasonable search and seizure, have been determined by the courts to capture the right to privacy [72,73]. These cases have expanded on the Charter sections to include privacy as related to protection from government or other intrusion, autonomy, and dignity.

Federally, Canada has two privacy laws. The *Privacy Act* [74] governs roughly 160 federal public bodies, whereas the *Personal Information and Protection of Electronic Documents Act* (PIPEDA) [60] governs private sector organisations regulated federally and provincially. Provinces with privacy legislation similar to *PIPEDA* are exempt from its provincial aspect. At the time of writing, British Columbia, Alberta and Québec have such legislation, and Ontario has health-specific legislation that exempts it from the corresponding section.

All provinces and territories have legislation similar to the *Privacy Act*, whereas only three provinces have private-sector legislation similar to *PIPEDA*. In addition, four provinces have specific health *information* legislation: Alberta, Manitoba, Ontario and Saskatchewan.

The UK has three legal jurisdictions: England and Wales, Scotland and Northern Ireland. However, it itself is also part of a larger community - the European Union (EU). European Union legislation is generally intended to "direct" that of its member states, and takes precedence in cases where there is no concurrence; the UK is obligated to align itself with EU law (referred to as Community law) [75] or else give way in a court of law to the latter [76]. Let us therefore begin with the EU.

The concepts of privacy and personal information are captured in core EU legislative documents as fundamental rights. The *European Convention for the Protection of Human Rights and Fundamental Freedoms* (ECHR), building on the 1948 *Universal*

*Declaration of Human Rights* [77], includes a “Right to respect for private and family life” in Article 8 [78]. The *Charter of Fundamental Rights of the European Union*, proclaimed in 2000, builds on the ECHR [79]. Updated in 2007, the Charter includes two particularly relevant articles. Article 7 reiterates the *ECHR*’s position on the respect for private and family life, whereas Article 8 explicitly limits the processing of personal data to specified purposes, requiring either individual consent or legislated “permission”.

Recognising the importance of data-sharing and the threats and benefits of developing technologies, the EU introduced a number of legislative pieces to harmonise, regulate and facilitate the flow of personal information. In 1995, *Directive 95/46/EC* was adopted for the protection of personal data [80] - the core directive at the heart of data protection in EU member states. It does not, however, apply, to personal information used solely for personal reasons, household activities, public security, national defence or criminal law enforcement, and falls short when dealing with issues around communication. Two years later, the EU adopted *Directive 97/66/EC* for protecting privacy and confidentiality in telecommunications [81]. As technology and the Web became increasingly ubiquitous, this directive quickly became limited in scope. It was therefore replaced in 2002 by *Directive 2002/58/EC* [82] covering electronic communications more broadly, and updated again in 2006 by *Directive 2006/24/EC* [83]. In addition, *Data Protection Regulation (EC) 45/2001* [84] ensures the protection of personal information in EU institutions and bodies, such as the European Parliament, for example, and accountability to a governing body, the European Data Protection Supervisor.

In the UK, the *Data Protection Act* was first enacted on July 12, 1984, thereby preceding the *Directive on Data Protection* adopted by the European Union (EU) by more than a decade. Upon adoption of the EU directive, however, the Act was amended in 1998. Though simpler than Canadian legislation in the sense that it applies

to both public and private entities, it is none-the-less a complex document. In 2003, Lord Phillips of the Supreme Court of Judicature, Court of Appeal (Civil Division) in the UK referred to it as "...a cumbersome and inelegant piece of legislation" [85]. Other UK health-related Acts have been amended to reference the *Data Protection Act 1998*, including the *Access to Health Records Act 1990*, the *Access to Medical Reports Act 1988* and the *Access to Personal Files and Medical Reports (Northern Ireland)*. The UK also has a *Health and Social Care Act 2008* [86], which replaced its 2001 predecessor and legislated the creation of a Care Quality Commission for the protection and promotion of the health, safety and welfare of the public. The Act makes it an offence to recklessly disclose confidential personal information obtained by the Commission that "relates to and identifies an individual." (S. 76)

Scotland has a *Freedom of Information Act 2002*, but a search on the UK Office of Public Sector Information Website [87] yielded no specific data protection legislation for either Scotland or Northern Ireland. Scotland also has a *Public Health Act* enacted in 2008 [88], which obligates Scottish Ministers, health boards and local authorities to protect public health. It allows for the disclosure of information to facilitate its directives despite any other legal prohibition or restriction, except, interestingly, the *Data Protection Act 1998* (S. 117(6)). Northern Ireland's *Health and Social Care (Reform) Act 2009* [89] has a similar clause (S. 13(8)).

Both Canada and the UK have a tapestry of legislative documents in place to protect the privacy of personal information "...as something worth protecting as an aspect of human autonomy and dignity." [90] But what, exactly, constitutes personal information?

## **4.2. DEFINITIONS**

There is no consistent definition for "personally identifiable data" or "personal information" in Canadian legislation. Where a definition is included, it ranges from "information about an identifiable individual" in Alberta's *Personal Information*

*Protection Act* [91] to very well-defined and explicit components in Manitoba's *Freedom of Information and Protection of Privacy Act* [92]. Of the 30 acts and regulations reviewed, four include health information in their definition of personal information, three include location information, 14 include both and nine include neither (Table 1).

This definition of personal information as pertaining to an "identifiable individual" appears quite often in legislation, including in *Directive 95/46/EC*. However, the *Directive* goes one step further to clarify: "...an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity" [80]. Health information is defined as a "special" category of personal information (S. III, Article 8 (1)), but there is no specific mention of location information in the *Directive*.

In the UK, the *Data Protection Act 1998* defines "personal data" vaguely as any information that, in isolation or in concert with other data available to the data controller, can identify a living individual. The *Act* also includes health in the definition of "sensitive personal data", but does not capture location information specifically. As mentioned previously, the *Health and Social Care Act 2008* also identifies confidential personal information as that which "relates to and identifies and individual", but does not specifically identify location as part of that definition.

As recent as April 2009, the Supreme Court of Canada stated that "Privacy analysis is laden with value judgements that are made from the independent perspective of the reasonable and informed person who is concerned about the long-term consequences of government action for the protection of privacy" [93]. As described, the definition of "personal information" in most cases casts a wide net, capturing anything and everything that can subjectively be argued as identifying. This has obvious implications



**Table 1: Inclusion of health and location information in the definitions of "personal information" in Canadian legislation**

Jurisdiction	Act	Reference	In Definition	
			Health	Location
Canada	The Privacy Act [190]	R.S.C. 1985, c. P-21	✓	✓
Canada	Personal Information Protection and Electronic Documents Act [191]	S.C. 2000, c. 5 P-8.6	✓	
B.C.	Freedom of Information and Protection of Privacy Act [395]	R.S.B.C. 1996, c. 165		
B.C.	Personal Information Protection Act [396]	S.B.C. 2003, c. 63		
B.C.	Freedom of Information and Protection of Privacy Regulation [397]	B.C. Reg 323/93		
B.C.	Personal Information Protection Act Regulations [398]	B.C. Reg. 473/2003		✓
B.C.	British Columbia Cancer Agency Research Information Regulation [399]	B.C. Reg. 286/91	✓	✓
B.C.	Privacy Act [400]	R.S.B.C. 1996, c. 373		
AB	Health Information Act [401]	R.S.A. 2000, c. H-5	✓	✓
AB	Freedom of Information and Protection of Privacy Act [402]	R.S.A. 2000, c. F-25	✓	✓
AB	Personal Information Protection Act [393]	S.A. 2003 c. P-6.5		

**Table 1: Inclusion of health and location information in the definitions of "personal information" in Canadian legislation (continued)**

Jurisdiction	Act	Reference	In Definition	
			Health	Location
AB	Personal Information Protection Act Regulation [403]	AR 366/2003		✓
SK	The Health Information Protection Act [404]	S.S. 1999, c. H-0.021	✓	
SK	The Freedom of Information and Protection of Privacy Act [405]	SS. 1990-91, c. F-22.01		✓
SK	The Local Authority Freedom of Information and Protection of Privacy Act [406]	SS. 1990-91, c. L-27.1	✓	✓
MB	The Personal Health Information Act [407]	C.C.S.M., c. P-33.5	✓	
MB	The Freedom of Information and Protection of Privacy Act [394]	C.C.S.M., c. F-175	✓	✓
ON	Personal Health Information Protection Act [408]	S.O. 2004, c. 3	✓	
ON	Freedom of Information and Protection of Privacy Act [409]	R.S.O. 1990, c. F-31	✓	✓
ON	Municipal Freedom of Information and Protection of Privacy Act [410]	R.S.O. 1990, c. M.56	✓	✓

**Table 1: Inclusion of health and location information in the definitions of "personal information" in Canadian legislation (continued)**

Jurisdiction	Act	Reference	In Definition	
			Health	Location
QC	An Act respecting Access to documents held by public bodies and the protection of personal information [411]	R.S.Q., c. A-2.1		
QC	An Act respecting the Protection of personal information in the private sector [412]	R.S.Q., c. P-39.1		
N.B.	Protection of Personal Information Act [413]	S.N.B. 1998, c. P-19.1		
N.S.	Freedom of Information and Protection of Privacy Act [414]	S.N.S. 1993, c. 5, s. 1	✓	✓
N.S.	Health Protection Act [415]	S.N.S. 2004, c. 4, s. 1		
P.E.I.	Freedom of Information and Protection of Privacy Act [416]	R.S.P.E.I. 1988, c. F-15.01	✓	✓
NL	Access to Information and Protection of Privacy Act [417]	S.N.L. 2002, c. A-1.1	✓	✓
YK	Access to Information and Protection of Privacy Act [418]	R.S.Y. 2002, c. 1	✓	✓
N.T.	Access to Information and Protection of Privacy Act [419]	S.N.W.T. 1994, c. 20	✓	✓
NU	Access to Information and Protection of Privacy Act [420]	S.N.W.T. 1994, c.20	✓	✓

on the use of disaggregate geographic data in health research. Or does it? The answer depends on the applications and exceptions made in the legislation.

### **4.3. APPLICATION AND EXCEPTIONS**

Legislation in Canada, the EU and the UK specifically limits the processing of personal information. What constitutes “processing”, however, is not consistently defined across legislation. The broadest definition to capture what this means is found in EU *Directive 95/46/EC*: “any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organisation, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction”. Generally, any such processing of personal information is prohibited *in the absence of the individual's informed consent*, unless it is first stripped of all identifying information (thereby ceasing to be personal information according to the legal definition).

In public health research, however, it is often impossible or impractical to pursue informed consent. Despite being incredibly information and data-rich, health researchers in both Canada and the UK have often expressed frustration over their inability to use existing data due to privacy concerns [70]. Is the prohibition based on the legislation?

Generally, in the absence of an individual's consent, the legislation does explicitly allow for some exceptions, particularly in the interests of national security. However, there is a lack of clarity and consistency, specifically around processing for public health purposes. Article 35 of the *Charter of Fundamental Rights of the European Union* emphasises the right to health care, and states “A high level of human health protection shall be ensured in the definition and implementation of all Union policies and activities” [79]. In almost all cases, exceptions are also made for research, as long as the

individuals whose data are processed are not identified in the results. Generally, the individual whose information has been disclosed should be informed; however, provisions are also made for cases where doing so is impossible or unreasonable.

The decision around whether or not the processing of the information is permitted under these exceptions is somewhat vague and inconsistent. In Canada, for example, the four provinces with health information legislation delegate the decision making authority to research ethics boards; otherwise, it is generally delegated to the head of the data-holding organisation. In the case of EU institutions, processing is only permissible after consultation with the European Data Protection Supervisor [84], whereas the UK *Data Protection Act 1998* exception for research (S. 4(33)) is unclear as to the decision-making authority. This leads to issues around governance.

#### **4.4. GOVERNANCE**

In Canada, the Office of the Privacy Commissioner (OPC) is responsible for protecting and promoting the privacy rights of Canadians by overseeing compliance with Canadian federal privacy legislation [94]. Each province and territory also has its own privacy commissioners who oversee their respective jurisdictions. As previously noted, health information legislation in Alberta, Saskatchewan, Manitoba and Ontario also delegates decision-making authority on these matters to research ethics boards.

The EU, as previously mentioned, has established the office of the European Data Protection Supervisor [95] for oversight of EU institution activities. The UK's equivalent of Canada's Office of the Privacy Commissioner is the Information Commissioner's Office (ICO) [96]. The legislation does not specifically mention research ethics boards or committees, and is unclear as to decision-making authority – in most cases, it seems to lie with the data controllers.

#### **4.5. IMPLICATIONS AND FINAL THOUGHTS**

The privacy of personal information is a recognised and important human right, protected through multiple intertwined acts and regulations in Canada, the EU and the UK. In the absence of informed consent, the legislation generally allows for the processing of an individual's personal information – which is any information that can identify the individual, and therefore includes health and disaggregate location information – for research purposes, subject to approval by the appropriate authority. However, guidelines are lacking, and authorities tend to err on the conservative side, resulting in much expressed frustration by health researchers. In the absence of frameworks to inform the processing of personal information, the only other alternative (besides seeking informed consent from every individual) for health researchers is the use of de-identification techniques, such as might be applied through privacy-preserving solutions involving disaggregate geographic data.

It has been suggested that privacy in the United States, Canada and the European Union have their bases in slightly different philosophical constructs: in the United States, privacy is anchored in protection from the government; in Canada, in principles of autonomy and control; and in the European Union, the focus is more on dignity and public image [97]. The argument is made that the Canadian model offers the appropriate “middle-ground” – after all, if individuals truly do have control over their own personal information, then they can choose to protect it from the government and others, and their dignity as far as public image is concerned is in their own hands. If we accept this definition of privacy – that is, having control over one's own personal information – then one might ask whether de-identification really solves the issue. Perhaps what is really needed is public health specific clarification in the legislation, public and practitioner education, and clear and concise frameworks and guidelines.

The importance of having some understanding of the legal system and a working

relationship with the legal profession is becoming increasingly recognised in public health [2]. However, while the privacy debate in public health may be fuelled in part by misperceptions of public health practitioners, it is very much coupled with a lack of understanding of the requirements of public health by legal practitioners. “Privacy laws are most burdensome and least effective when they apply broadly, without proper concern for the settings in which they operate, the types of information that they cover, the obligations that they impose and the purposes they were designed to serve” [52]. The issue can only be truly addressed through interdisciplinary collaboration. Until that happens, and until we recognise the importance and value of public health research and its implications on the health of individuals, we will continue to grapple with alternate de-identification solutions and sub-optimal data.

## 5. Public Health Practitioner Perceptions

### 5.1. INTRODUCTION

Having established some of the historical, social and legal background to the concept of privacy, it is only fitting to proceed by documenting and quantifying its effects on public health as experienced and perceived by public health practitioners. A review of the literature revealed a gap in this area, leading to the development and implementation by the author of a Web-based survey conducted in Canada and the UK between November 2006 and January 2007.

Both Canada and the UK have many similarities in their health care models, and consequently share some of the same pros and cons. Canada's model was based on the UK's National Health Service (NHS) [19,98], while the Canadian health care system continues to be studied for lessons that may apply to the UK [99]. From a health research perspective, both Canada and the UK place strong emphasis on evidence-based public health policies and services [23], yet in both countries, this continues to be hampered by privacy issues.

The survey was developed in paper form by the author and converted to a Web-based format by the ALPHA Project [100] team at the Public Health Agency of Canada (PHAC). The process involved a complete business submission to the ALPHA team, and the necessary functional and logistical details are included in Appendix A (Volume II). Two country-specific versions of the survey were developed: one for Canada, and another for the UK, each with country-relevant response options. In addition, because Canada is officially bilingual, the Canadian content was also officially translated to produce English and French versions. The three survey versions are provided in Appendix B (Volume II), and a high-level overview of the survey's structure and flow is illustrated in Figure 1.



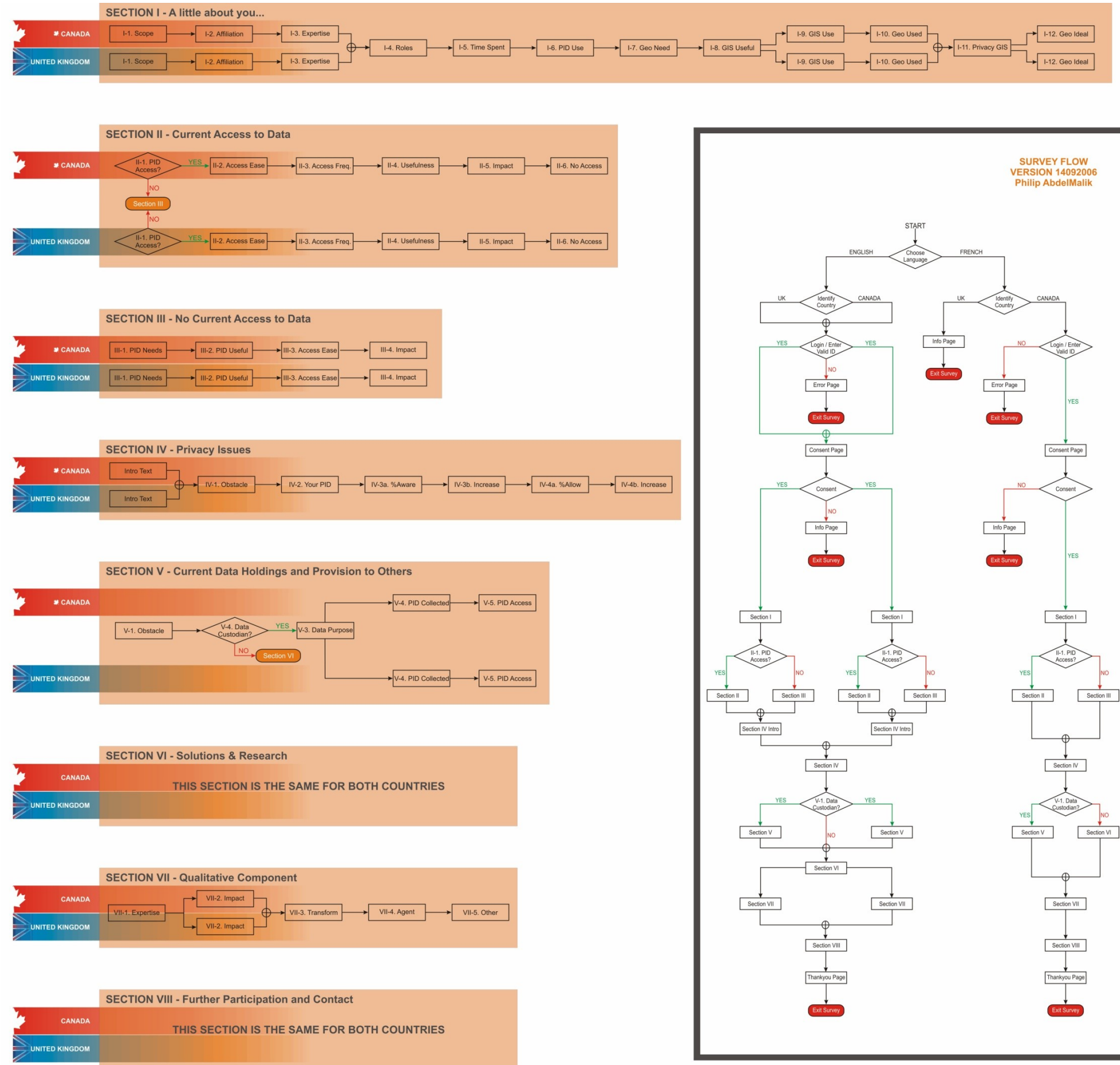
The survey contained both quantitative and qualitative components, and confirmed the following:

1. There is a definite requirement by public health professionals for personally identifiable data, including spatial data. The requirement for this spatial data is at its most granular level – latitude and longitude, or exact street address – which necessarily compromises patient privacy.
2. Participants generally rated privacy as an obstacle to public health practice (Figure 2); interestingly, the more highly the self-rated knowledge of privacy legislation, the more of an obstacle privacy was generally rated (Figure 3).
3. The most critical obstacles implicated in this perception of privacy as a critical issue in both countries are bureaucracy and legislation

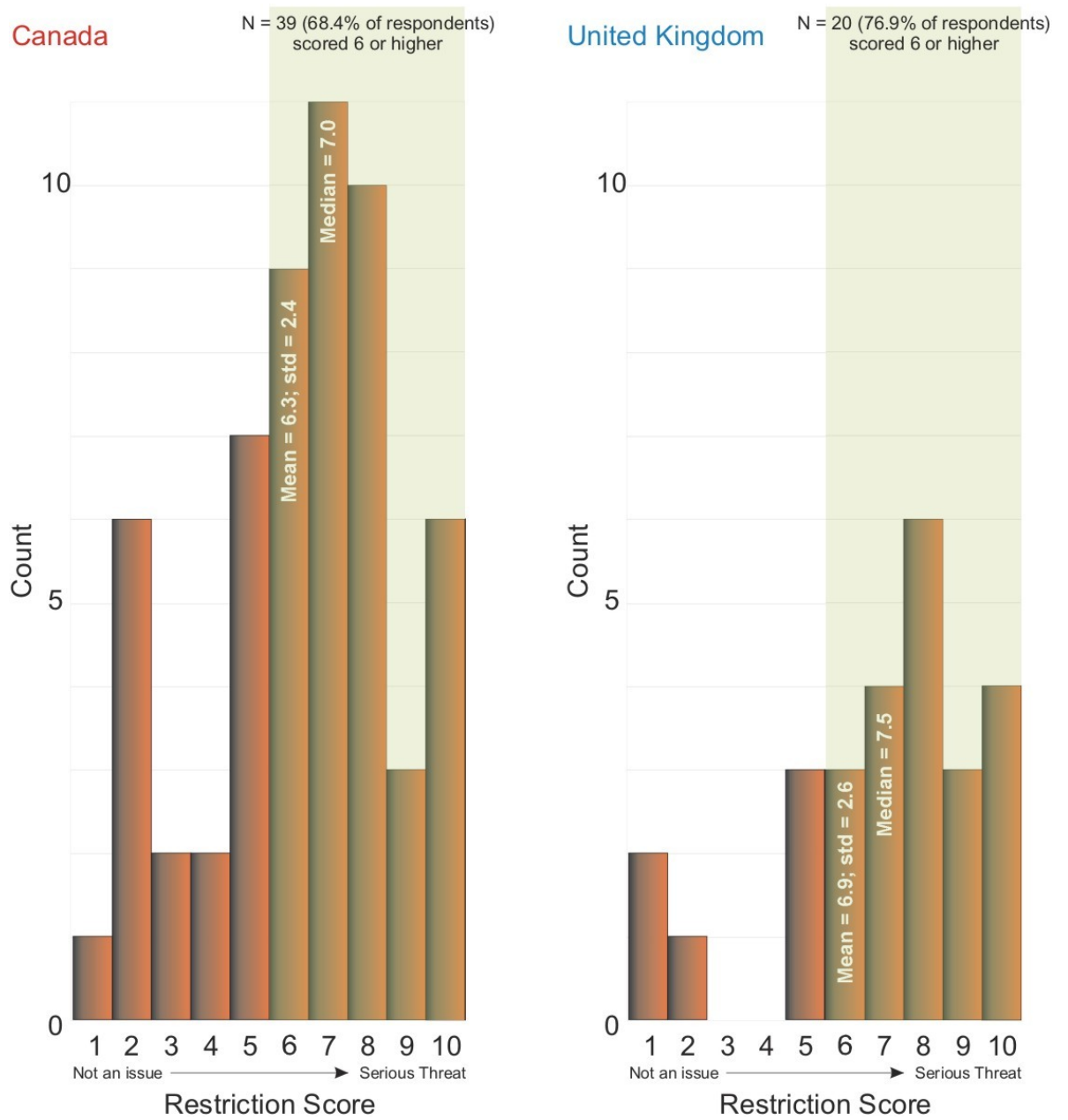
While many individuals recognised the importance of privacy legislation, participants generally indicated a concern and, in some cases, first-hand frustration that legislation unduly restricts public health activities, compromising surveillance and research.

All of the compiled results from the survey are provided in Appendix C (Volume II), and the most salient details were published as an original research article in *BMC Public Health* [70].

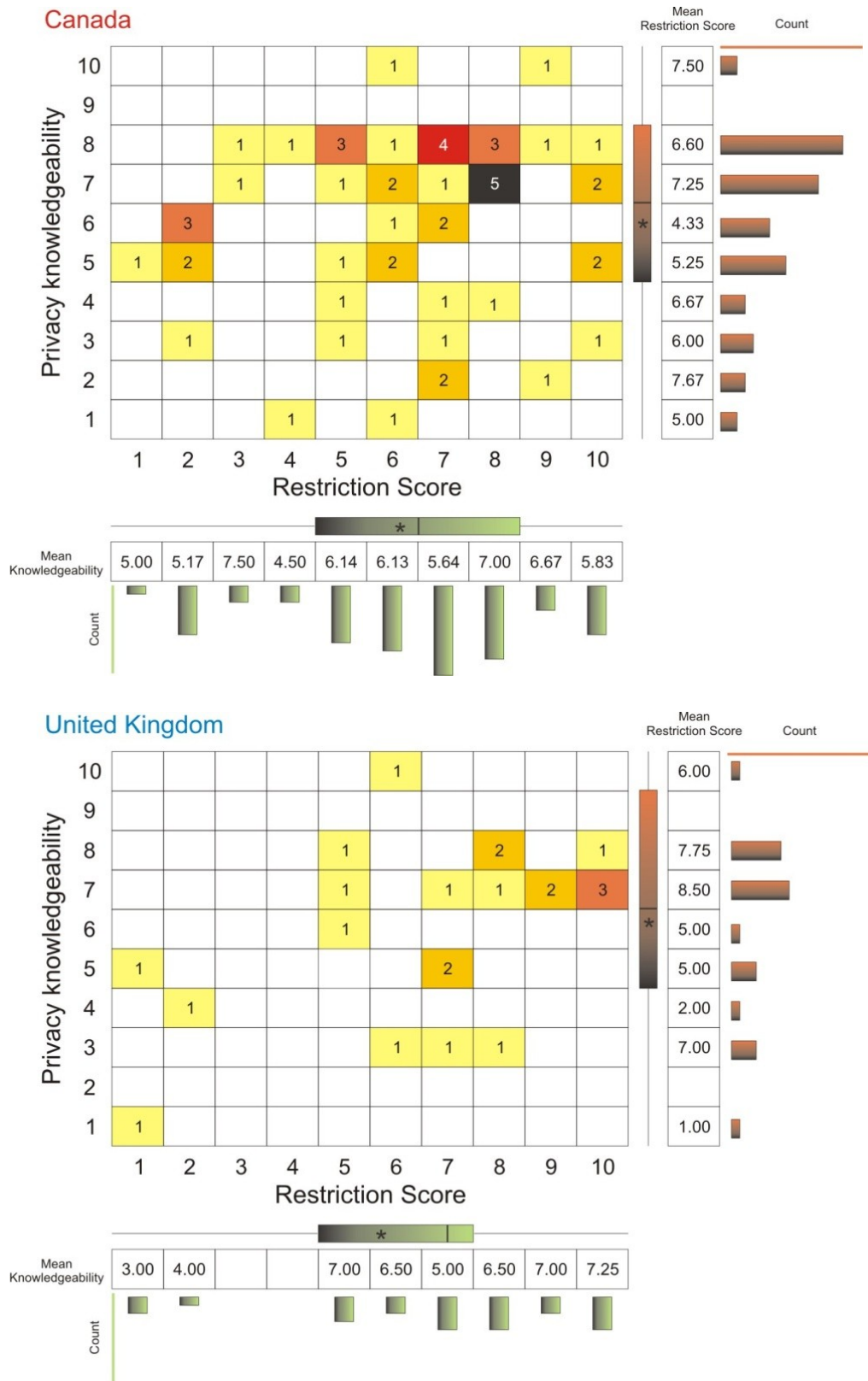
**Figure 1: Sections and flow of the Web-based survey to collect practitioner perceptions on the impact of privacy on public health practice**



**Figure 2: Participant rating of the degree to which privacy restrictions pose an obstacle to public health practice**



**Figure 3: Relationship between self-rated knowledge of privacy legislation and policies and the degree to which privacy was rated as an obstacle by survey participants.**



## 5.2. THE PERCEIVED IMPACT OF LOCATION PRIVACY

*Original research article as published in BMC Public Health (2008) [70]*

### 5.2.1. OVERVIEW

#### **Background**

The “place-consciousness” of public health professionals is on the rise as spatial analyses and Geographic Information Systems (GIS) are rapidly becoming key components of their toolbox. However, “place” is most useful at its most precise, granular scale – which increases identification risks, thereby clashing with privacy issues. This paper describes the views and requirements of public health professionals in Canada and the UK on privacy issues and spatial data, as collected through a Web-based survey.

#### **Methods**

Perceptions on the impact of privacy were collected through a Web-based survey administered between November 2006 and January 2007. The survey targeted government, non-government and academic GIS labs and research groups involved in public health, as well as public health units (Canada), ministries, and observatories (UK). Potential participants were invited to participate through personally addressed, standardised emails.

#### **Results**

Of 112 invitees in Canada and 75 in the UK, 66 and 28 participated in the survey, respectively. The completion proportion for Canada was 91%, and 86% for the UK. No response differences were observed between the two countries. Ninety three percent of participants indicated a requirement for personally identifiable data (PID) in their public health activities, including geographic information. Privacy was identified as an obstacle to public health practice by 71% of respondents. The overall self-rated median score for knowledge of privacy legislation and policies was 7 out of 10. Those who

rated their knowledge of privacy as high (at the median or above) also rated it significantly more severe as an obstacle to research ( $P < 0.001$ ). The most critical cause cited by participants in both countries was bureaucracy.

## Conclusions

The clash between PID requirements – including granular geography - and limitations imposed by privacy and its associated bureaucracy require immediate attention and solutions, particularly given the increasing utilisation of GIS in public health. Solutions include harmonization of privacy legislation with public health requirements, bureaucratic simplification, increased multidisciplinary discourse, education, and development of toolsets, algorithms and guidelines for using and reporting on disaggregate data.

### 5.2.2. BACKGROUND

Although “place” has been coined one of the three pillars of epidemiological data, only relatively recently has it garnered significant attention in the public health field, as Geographic Information Systems (GIS) have increasingly become more affordable, accessible, and intuitive. Indeed, the public health community’s “place-consciousness” is on the rise as spatial analyses and GIS, now defined as part of the medical and health literature [1,101,102], are rapidly becoming key components of the public health professional’s toolbox [44].

Privacy, an evolving “principle as old as the common law” [3], has been cited as an issue in a variety of public health events, reports, and media releases [23,43,103-106]. So much so, in fact, that one sometimes cannot help but wonder if privacy is, indeed, the enemy of public health [4], and whether they could ever peacefully co-exist [5]. A distinction should here be made between the related concepts of *privacy*, *confidentiality*, and *security* within the context of the current discussion (these were defined in Chapter 2 but are included here as published in the manuscript). *Privacy* is

attributable to the individual about whom identifiable information pertains, and refers to that individual's right to control such information, thereby freeing the individual from uninvited intrusion and identification. *Confidentiality* obligates others who have been entrusted with such information to respect the individual's privacy, and is therefore attributable to third parties; a breach of confidentiality violates the privacy of the individual because the individual has had no control over the release of the data. Finally, *security* refers to tools and methods used to safeguard confidentiality and privacy [2,107]. This research deals specifically with privacy issues as regulated and defined by legislation and ethical guidelines surrounding consent. From within this context, an individual's privacy is not deemed to have been violated if data shared in the absence of consent cannot be used to identify the individual. Exception clauses generally exist in legislation, allowing authorities to release personally identifiable data under various circumstances – such as where it is deemed to be in the best interest of society or where it is impractical to obtain consent. Examples include Section 60 of the UK's *Health and Social Care Act 2001* [108], and Sections 8 and 7 of Canada's *Privacy Act* [74] and *Personal Information Protection and Electronic Documents Act* [60], respectively. While an analysis of privacy legislation as it pertains to health data and the concept of "place" is beyond the scope of this paper, suffice it to say that such clauses are often ambiguous and subjective, particularly when combined with vague definitions of "sensitive personal information" and the scale at which geographic data become "identifiable". The concept of *place*, for example, is not explicitly specified as "sensitive personal data" in the UK's *Data Protection Act 1988* [109], nor in the generic *EU Data Protection Directive* of 1995 [80] (though it is explicitly mentioned in various telecommunications directives), but postcodes are specifically mentioned in a 2005 NHS data protection and medical research POSTnote [110]. In Canada's *Privacy Act* [74], "address" is specifically listed as "personal information", while in the *Personal Information Protection and Electronic Documents Act* [60], it is not (though implied). Such ambiguities deter the sharing of data, causing organisations and authorities to err

on the side of caution and not release identifying information [111], including spatial data.

It is no surprise, therefore, that the increasing popularity of “place” in public health has further exacerbated the public health research-privacy debate. Traditional health-data anonymisation techniques, such as pseudonymisation and aggregation, cannot be applied to spatial data without significantly altering or destroying the spatial relationships under investigation [9,10,15,112], and hence the very reason for which they are to be used in the first place. The problem with “place” is that it is most useful at its most precise, granular scale [15,107]. Yet with increasing spatial precision and accuracy comes a corresponding increase in the risk of identification, and therefore a breach of privacy [107]. This becomes particularly troublesome when the spatial data are linked to health, social or demographic data. The development of methods by which to mitigate these risks continues to be an active area of research, but thus far, proposed solutions have limitations, risks and tradeoffs, and lack guidelines on their appropriate use. Consequently, the acquisition of geographic data tends to be either limited, or at a sub-optimal or unusable scale. Not only do privacy issues impact data acquisition and use for analysis, but also visualisation and dissemination of the results. Researchers have been able to “reverse engineer” maps, for example, to successfully re-identify individuals [11-13].

While the debate between the fields of privacy and public health has raged on for decades [3] despite their interdependence on one another [2], tension continues to rise in concert with the rampant growth of information technology and e-Health. From a health research perspective, both Canada and the UK place strong emphasis on evidence-based public health policies and services [23], yet in both countries, this seems to be hampered by privacy issues. While some argue that this debate is the product of a lack of understanding of the legislation and regulations by the public health community [2,14,113], there is little in the way of formal collection and synthesis of the



corresponding views and perspectives of those directly involved in public health activities. This paper describes the views and requirements of public health professionals in Canada and the UK on privacy issues and spatial data, as collected through a Web-based survey. Given that Canada's health care and public health systems were both largely modelled after those of the UK [19,23,98], that each continues to be studied by the other for improvements and lessons learned [23,99], and that privacy issues for public health have been cited in both, it is expected that survey responses in the two countries will also be similar.

### **5.2.3. METHODS**

#### **Development & Content**

The survey was first developed on paper in the summer of 2006, and piloted with select public health individuals in Canada and the UK. It was then submitted for privacy assessment by the Access to Information and Privacy Branch of Health Canada, and for ethics review and approval from the Health Canada Research Ethics Board and the Southwest Multicentre Research Ethics Committee in the UK. Throughout the process it was clear that the survey would be developed as a closed Web-based survey, running between November 2006 and January 2007. The final paper version of the survey can be found on the research Website and in Appendix B (Volume II) of this work [114].

The paper survey was converted to a Web-version by the ALPHA Project [100] team at the Public Health Agency of Canada (PHAC), and piloted by the author and several colleagues within the PHAC. The survey launch was delayed by two weeks, with only some of the concerns identified during the pilot being implemented due to limitations of the ALPHA architecture. Issues and limitations with the design of the Web-based survey are addressed in a later section.

Three versions of the survey were developed and launched: Canada-English, Canada-French and UK-English. A summary of the survey's structure and contents is given in Table 2.

### Target

The survey targeted government, non-government and academic GIS labs and research groups involved in public health, as well as public health units (Canada), ministries, and observatories (UK). Potential participants were identified through Web searches of public health sites, mailing databases, personal contact, referrals / word of mouth, and postings on the research Website [114], a PHAC Public Health Portal Website [115], and the NHS Public Health Informatics Community Website [116].

**Table 2: Sections of the survey**

Section	Title	Description
I	A little about you...	Participant scope, roles, use of GIS, etc.
II	Current access to data	Asks participants with current access to PID to score 15 kinds of PID* on various dimensions, such as ease and frequency of access, usefulness and importance, etc.
III	No current access to data	Asks participants without current access to PID to score same as above
IV	Privacy issues	Collects participant opinions on the overall impact of restricted access to PID on public health practice (research, surveillance, health service delivery, etc.)
V	Current data holdings and provision to others...	Collects information on the sharing of PID within and between participant organisations

**Table 2: Sections of the survey (continued)**

<b>Section</b>	<b>Title</b>	<b>Description</b>
VI	Solutions and research	Presents two distinct solutions to overcome barriers posed by privacy to public health research, and gather participant views on usefulness, usability and preference for each
VII	Qualitative component	Allows participants to provide views and opinions on knowledge of privacy and confidentiality issues / legislation, impact of privacy, proposed research and solutions, and additional thoughts or comments
VIII	Further participation and contact	Allows participants to provide contact information if they choose, for follow-up, updates, or piloting of potential solution(s)

\* For all participants: first name; last name; initials; sex; date of birth; date of death; registered GP or family physician; street address; postal code; community name; city / town / village; region / geographic area; latitude / longitude.  
For Canadian participants: provincial health insurance plan number; hospital ID.  
For UK participants: old NHS number; new NHS number

### **Participation**

Potential participants were invited to participate through a standardised but personally addressed email outlining the reason for the invite, the mechanisms by which their contact information was retrieved, a brief summary of the research and survey, a description of the data handling methods, an estimate of the time it would take to complete the survey (approximately 20 minutes), a unique user ID and password, the URL to the survey site, the URL to the research Website, and the principle investigator's contact information. A copy of the email content is provided in Appendix A (Volume II).

The survey Website had no other content. In order to participate, invitees were required to (1) successfully log in, and (2) consent to participation. Only the most recent

responses for any given user ID were collected, ensuring only one survey was completed per participant. The consent screen outlined the voluntary and anonymous nature of the survey, indicated the approximate time it would take to complete the survey, the risks and benefits to the participants, the intellectual property and ownership of all data collected, and the protection of any personal data provided under Canadian and UK law. Failure to successfully complete either of these two requirements resulted in termination of the survey. After consenting, participants were given the option to select their country and language of choice, and the relevant survey then commenced.

All questions included a “Skip” option. Progress through the survey required the selection of a response for each question, and participants could terminate the survey at any time or complete it over multiple sessions, at their convenience. Questions were not randomized or alternated, but adaptive questioning was utilized. Question types varied, and included single-choice, multiple-choice, scale, and free-form response questions, thereby collecting both quantitative and qualitative responses. There was typically only one question per screen with multiple potential responses, the maximum number of which was 17. Depending on the responses of the participants, the survey was distributed over approximately 40 screens.

Key questions addressed by the survey included the following:

Is there a requirement for personally identifiable data, including spatial data?

What spatial resolution is ideal for public health research?

Is privacy perceived to be a significant obstacle to public health practice?

How knowledgeable do public health professionals consider themselves on  
privacy?

What is the most critical obstacle to the access and use of personally identifiable  
data?

What are the views of the public health community on public awareness and perceptions?

Which is preferred: raw, case-level data, or aggregated, anonymised data?

Collected responses were analysed using basic descriptive statistics and non-parametric methods in SAS 9.2. The Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [117] was used as a guideline in the reporting of the Web-based survey methodology.

#### **5.2.4. RESULTS**

Of 112 invitees in Canada and 75 in the UK, 66 (59%) and 28 (37%) participated in the survey, respectively. Of the Canadian participants, three responded to the French version. The completion proportion for Canada was 91%, and 86% for the UK.

There were no differences in the distribution of roles reported by participants in both countries, with most participants (49% in Canada; 64% in the UK) identifying their main role as falling within the research and analysis domain (Table 3). Participant expertise varied, and included aboriginal health (Canada only), chronic diseases, paediatric public health, infectious diseases, dental public health, emergency preparedness and response, environmental public health, ethics and public health law, food and nutrition, health services, injuries and disabilities, mental health and substance misuse, social determinants of health, surveillance, and education.

No response differences were observed between the two countries on each of the key questions, and the overall, combined results are therefore reported. A summary of the findings is given in Table 4.

#### **Is there a requirement for personally identifiable data, including spatial data?**

Almost all participants identified a need for personally identifiable data (PID) in their roles; only one Canadian participant indicated no need for PID. Five Canadian participants and one UK participant chose not to answer the question. In total 93% of

participants indicated a requirement for PID in their public health activities.

**Table 3: Number and percentage of survey participants by main role and geographical scope**

Scope	Main Role					
	Strategic decision / policy maker	Manager / Coordinator	Consultant	Research & Analysis	Front-Line Responder / Patient Care / Clinical	Other
<b>Canadian Participants</b>						
North American or National	3 (4.5%)	6 (9%)	-	9 (13.6%)	1 (1.5%)	2 (3.0%)
Provincial / Territorial	1 (1.5%)	3 (4.5%)	4 (6.1%)	6 (9.1%)	-	2 (3.0%)
Local / Regional	2 (3.0%)	7 (10.6%)	1 (1.5%)	17 (25.8%)	1 (1.5%)	1 (1.5%)
<b>Totals</b>	<b>6 (9.1%)</b>	<b>16 (24.2%)</b>	<b>5 (7.6%)</b>	<b>32 (48.5%)</b>	<b>2 (3.0%)</b>	<b>5 (7.6%)</b>

**Table 3: Number and percentage of survey participants by main role and geographical scope (continued)**

Scope	Main Role					
	Strategic decision / policy maker	Manager / Coordinator	Consultant	Research & Analysis	Front-Line Responder / Patient Care / Clinical	Other
<b>UK Participants</b>						
European or National	1 (3.6%)	1 (3.6%)	-	1 (3.6%)	-	-
Regional	2 (7.1%)	1 (3.6%)	2 (7.1%)	12 (42.9%)	-	-
Local	2 (7.1%)	-	-	4 (14.3%)	-	1 (3.6%)
<b>Totals</b>	<b>5 (17.9%)</b>	<b>2 (7.1%)</b>	<b>2 (7.1%)</b>	<b>17 (60.7%)</b>	<b>0 (0.0%)</b>	<b>1 (3.6%)</b>

\*One UK participant who identified a main role in research and analysis declined a response to the question on scope.

### **What spatial resolution is ideal for public health research?**

All participants identified geographic location of health data as a requirement for their roles or organisation. When asked "...what level of geography would you ideally like to visualise your data and/or conduct spatial analyses", 69% of respondents identified "latitude and longitude, exact street address, or exact household".

**Table 4: Summary of survey findings**

Question	Response Summary <sup>†</sup>
1. Is there a requirement for personally identifiable data?	Yes (93%)
2. What spatial resolution is ideal for public health research?	Lat/Long or address (69%)
3. Is privacy perceived to be a significant obstacle to public health practice?	Yes (71%)
4. How knowledgeable do public health professionals consider themselves on privacy?	High Knowledge* (53%)
5. What is the most critical obstacle to the access and use of personally identifiable data?	Bureaucracy (33%) Legislation (25%)
6. What are the views of the public health community on public awareness and perceptions?	Less than 30% of the public is aware (84%)
7. Which is preferred: raw, case level data, or aggregated, anonymised data?	Raw, case-level data (66%)

<sup>†</sup>Numbers in parentheses are the percent of participants who responded as described

\*Participants rating their knowledge as high were also more likely to rate privacy as a more severe obstacle ( $P < 0.001$ )

**Is privacy perceived to be a significant obstacle to public health practice? AND How knowledgeable do public health professionals consider themselves on privacy?**

When asked “Are you or have you been restricted in your use of GIS for any public health activity because of privacy concerns (i.e. map or data might identify an individual or community)?” 79% of respondents marked “YES”.

Of 83 participants who responded to the question “In your opinion, do current restrictions to PID pose an obstacle to any aspects of public health practice?” 59 (71%) agreed, rating the obstacle severity at 6 or higher. Of these 59, 36 (61%) rated their



knowledge of privacy and confidentiality issues/legislation at 6 out of 10 or higher, with a mean score of 7.5 (std = 1.0) and a median score of 7.

Using the median, respondents with a self-rated knowledge score lower than 7 were classified as “low” on knowledge (47%), while those at or above the median score were classified as “high” (53%). Those classified as high were more likely to rate privacy as an obstacle (one-sided Wilcoxon exact  $P < 0.001$ ). A trend was evident for the overall correlation between restriction score and self-rated privacy knowledge score (Spearman  $r = 0.22$ ,  $P = 0.057$ ).

### **What is the most critical obstacle to the access and use of personally identifiable data?**

The most common obstacles were reported as bureaucracy and legislation by 33% and 25% of the participants, respectively. Other responses included public disapproval/paranoia (15%), practitioner paranoia (7%), lack of knowledge (6%), combination of these factors (4%), other (2%), and none (skipped question, 7%).

### **What are the views of the public health community on public awareness and perceptions?**

Fifty seven percent of participants felt that under 10% of the public population is aware of the impact of restricted access to PID on public health practice; 74% felt it to be under 20%, and 84% felt the proportion to be less than 30% (cumulative frequencies). Most identified education and awareness (through media, reports, case studies, scenarios, etc.) as the best methods to increase this proportion. When then asked what proportion of the public they felt would allow the use of their PID if they were educated on the usefulness of such data to public health practice, 67% said 50% or higher.

### **Which is preferred: raw, case-level data, or aggregated, anonymised data?**

More respondents identified a preference for having access to granular-level rather than aggregate data (53 vs. 27; 66% of those responding to this question).

#### **5.2.5. DISCUSSION**

This survey and user-needs assessment on privacy and public health shows a definite requirement by public health professionals – in various fields and positions in both Canada and the UK – for personally identifiable data, including spatial data. The requirement for this spatial data is at their most granular level – latitude and longitude, or exact street address – which necessarily compromises patient privacy. It is not surprising, therefore, that public health professionals perceive privacy to be a significant obstacle to public health practice.

There are those who would argue that this perception is the product of a lack of understanding of the legislation and regulations by the public health community. The results of this research, however, indicate the contrary. Not only did public health professionals in both countries generally rate themselves high on knowledge of privacy legislation and related issues, but those with the highest self-rated scores also tended to rate privacy as more of an obstacle. That these self-ratings of knowledge are not representative of actual knowledge remains possible.

Participants perceived the most critical obstacles to sharing or acquisition of health data with PID to be bureaucracy, followed by legislation.

Bureaucracy surrounding health research in both Canada and the UK generally revolves around data ownership, academic competitiveness, ethics review boards or committees, and in particular, requirements for informed consent, even if they compromise public health, or are not in the best interests of the patients involved [118-120]. Since seeking subject consent with every new hypothesis to be tested or model

to be developed is an impossible task, some have suggested that thought be given to “blanket” consent. At the Canadian Institutes for Health Research (CIHR) 2003 workshop on the legal and ethical issues facing the Canadian Lifelong Health Initiative [121], participants spent some time discussing such issues, only to emphasise the importance of the establishment of ethical governance and structure; essentially, more necessary bureaucracy. Interestingly, while the debate continues, a relatively recent survey found that most of the British public did not consider the use of their National Cancer Registry PID for public health research and surveillance to be an invasion of their privacy [14]. While the ethics of blanket consent are not discussed in this study, it is nonetheless offered as a potential solution in light of the requirements of the public health community. This does not, however, address other issues of data ownership and control that contribute to the bureaucratic debate.

While many individuals recognised the importance of privacy legislation, participants generally indicated a concern and, in some cases, first-hand frustration that legislation unduly restricts public health activities, compromising surveillance and research. Many phrases were used by respondents to describe the implications of privacy legislation on public health, including, among others: “increasingly restrictive”; “serious”; “incomplete”; “fuzzy”; “does more harm than good”; “two-edged sword”; “causes challenges”; “delays and restricts access [to data]”; “[is a] hindrance to the improvement and efficiency of public health”; “disappointing”; “frustrating”; “difficult to interpret”; “very worrisome”; “disadvantages the public interest”; “not properly understood”; “over-protective”; “limiting”; “hinders knowledge”; and “used as an excuse not to share data”. A large proportion of the public health community represented in this sample clearly expressed major concerns with the impact of privacy legislation on their work – both in Canada, and in the UK – in spite of having a good understanding and acceptance of its purpose and necessity. It is also important for legislation to be written in an unambiguous manner that is clearly understood by both public health professionals and the general public [44].

Public health professionals are largely of the opinion that the general public's level of awareness of the impact of restricted access to PID on public health practice is extremely low. Surveys by the Office of the Privacy Commissioner in Canada [122] repeatedly show that the majority of Canadians surveyed (up to 80%) place an extremely high level of importance on strong laws to protect personal information, particularly health information, and that they feel that the level of protection of their personal information has declined over the past ten years. Yet interestingly, only 20% are clearly aware of existing laws, and even fewer (12%) are aware of their rights around the collection, use and disclosure of this information. The "need to raise Canadians' awareness about the current laws in place and what their rights are" [122] must therefore be coupled with the corresponding need to address this from within the context of public health requirements.

Educating the public, therefore, as well as practitioners, data users, policy makers and politicians, was not surprisingly identified by participants as a potential solution. Participants put emphasis on the utilisation of the media to educate and increase awareness, as well as demonstrating the impact of a lack of data, and the benefits of its use when available. Demonstration of the benefits to the individual (e.g. streamlining of the system, not being asked for personal information with every visit to a new clinician, improved dissemination of public health information and intelligence directly to the public) was also offered as a solution, and summed up by one participant in the phrase "seeing is believing". It is worth noting, however, that a number of participants displayed a certain level of pessimism that until a crisis or extreme event occurs, no amount of education or awareness-increasing activities would make a difference.

Public health professionals generally prefer disaggregate, case-level data, but access to this data is an issue. The limitations imposed by privacy on public health have resulted in the development of a variety of techniques for data anonymisation [15,17,107]. However, all unavoidably have their issues, risks and limitations, and there

is currently no framework to guide public health professionals in their appropriate use and interpretation.

### **Generalisability**

Although the findings of this paper may be generalisable to public health professionals in Canada and the UK, issues of privacy and public health are not unique to these countries. Privacy is defined as a fundamental human right in the legislation of many countries, and the concept is enshrined in Article 12 of the United Nations' *Universal Declaration of Human Rights* [77] and Article 8 of the *European Convention on Human Rights* [123]. Similarly, public health is an international discipline; both diseases and information are ubiquitous, and neither is constrained by political boundaries and oceans. The increasing requirement for spatial data and the inherent clash with privacy legislation therefore extend beyond the UK and Canadian contexts, and the results, requirements and conclusions drawn from this research can be generalised to wherever such a clash exists. The implementation of solutions by national governments may be further exacerbated by issues of social political trust. General public distrust in government initiatives and motives, such as in most countries of the European Union, Canada, and the United States [124,125], complicates changes that may be perceived by the public to be intrusions of privacy. Such issues may currently be less of a concern in countries such as Finland, Sweden, Denmark, and the Netherlands, where social political trust, though declining, has traditionally tended to be much higher [126-129]. However, even in such nations where privacy and health have traditionally not clashed, increased international data-sharing requirements and spatial data implications may pose unanticipated and challenging obstacles.

### **Limitations**

No comprehensive lists of public health and health GIS professionals were found in either country, so it was not possible to invite a random sample. In addition, the response rate in the UK was relatively low, and it is therefore uncertain that the sample

is representative of all public health professionals in the two countries. However, responses between the two countries were consistent, with no significant differences.

Since knowledge of privacy legislation and policies was based on self-rated scores, a thorough review and assessment of privacy legislation as it pertains to public health practice is required in both Canada and the UK to validate the findings of this survey.

A number of limitations and issues pertaining to the Web-survey were identified. Most notable of these was the presence of a scroll bar in sections II and III which most participants missed, thereby eliminating the ability to capture items in reference to “place”, such as usefulness. However, these items were also captured more broadly in other sections of the survey. Other issues involved the inability of the architecture to support various designs and types of questions that would have facilitated the completion of the survey, and shortened the length of time required. Participants also noted frustration with the navigation and structure of the survey pages. A document outlining these issues and others was submitted to the ALPHA team after the initial pilot for future enhancements to the architecture.

#### **5.2.6. CONCLUSIONS**

It is clear that privacy is perceived to be a major obstacle and issue for public health – the literature illustrates it, and the current study provides both quantitative and qualitative evidence. Together, these provide a more holistic portrayal of public health community viewpoints, and can be used to educate the public, and as evidence for decision makers to implement changes in policies and legislation. The clash between a requirement for personally identifiable data – including exact, individual location - by public health professionals, and the limitations imposed by privacy and its associated bureaucracy, must be addressed and appropriate solutions developed, particularly given the increasing utilisation of geographic information systems in public health and the imminent completion of comprehensive electronic health systems. Privacy

legislation is critical for the protection of this fundamental human right, and to prevent the abuse of personal information, particularly in the health field. However, the legislation must be harmonised with the requirements of public health practice if the health of societies and populations is to be maintained and improved. Since health is not limited by political boundaries, this must be pursued at an international level, and solutions must address these perceptions in the public health community, simplify the bureaucratic process, promote multidisciplinary discussions between legislators, bureaucrats and the public health community, educate communities, and develop and provide public health professionals with toolsets, algorithms and guidelines for using and reporting on disaggregate data. While the results of this study should inform and justify the development of techniques that better anonymise health data with minimal impact on their integrity and frameworks for implementing them, it seems fitting to once again echo the warning of Curtis *et al.*: "...health and spatial scientists should be proactive and suggest a series of point level spatial confidentiality guidelines before governmental decisions are made which may be reactionary toward the threat of revealing confidential information, thereby imposing draconian limits on research using a GIS." [11]





**PART II  
CONCESSIONS**



## 6. Brief Overview of General Solutions

### 6.1. INTRODUCTION

In light of the complications imposed by privacy issues on public health research, several strategies have been utilized and/or suggested for working around the problem. Much of the methodology devised to protect the privacy of individuals in health data has also been applied to location information. These methods include access control and the implementation of secure networks; suppression of information; data aggregation; data anonymisation; implementation of automated analytical software agents; and mathematical transformations. Unfortunately, none of these solutions is without its problems.

### 6.2. ACCESS CONTROL AND SECURITY

*“Computer security is not privacy protection” [8]*

One potential solution to the issue of privacy is the development of secure networks and technology for the sharing of information [8,15,44,130], provided the relevant legislation and policies are complied with. Unfortunately, this solution is not optimal for some jurisdictions. In Canada, for example, health is a provincial matter, and data ownership can adversely affect collaborative efforts. Furthermore, the establishment of data-sharing agreements, and the development of the required infrastructures and networks is costly in terms of financial, human and temporal resources. The area of health informatics has also been slow to develop; although Canada has consistently ranked among the most technologically advanced and healthy countries in the world by the United Nations [131], progress in marrying these two fields has been quite slow, as evidenced by the 2004 outbreak of Severe Acute Respiratory Syndrome (SARS) in Toronto. In fact, in response to Toronto’s inefficient response to SARS, the city’s chief medical officer of health at the time stated that part of the problem was due to the use of “nineteenth century tools to fight a twenty-first century disease” [23]. Access control

and security measures are data-indiscriminate, meaning they can be applied to any type of information including location information.

Even if one were to implement streamlined and efficient access control and secure networks, this would not equate to real privacy protection [8]; rather it equates to restricted information access, as well as increased politics and bureaucracy as researchers try to go through the necessary steps and channels to gain access to data that still do not protect privacy. So while security is a fundamental aspect of a governance structure, it does not, in and of itself, constitute privacy protection. This significance to governance is reflected in the suggested framework in Part III.

### **6.3. SUPPRESSION OF INFORMATION**

*“Suppression (of sensitive information) can drastically reduce the quality of the data, and in the case of statistical use, overall statistics can be altered, rendering the data practically useless.” [8]*

Suppression of sensitive information is also data-indiscriminate and typically involves the complete removal of identifying and potentially identifying (or “quasi identifying”) data – for example, gender, age, location, etc. – before sharing or releasing them. While this can protect the identity of the individual, it can also severely compromise the integrity and analytical capacity of the event data [8]. For example, two of the most fundamental and commonly used attributes in public health analyses are gender and age. When combined with location, however, even if the latter is only at a seemingly general scale such as zip code or postal code, the risk for identification is quite dramatic. In the United States, roughly 87% of the population can be identified by just the 5-digit ZIP code, gender and date of birth [8], while in Canada postal codes can be even more sensitive: almost 98% of the population in the city of Montréal, Québec, Canada is identifiable given these three fields [132].

Suppression need not only apply to the health data, however – it may also apply to contextual data. For example, it may mean refraining from using street-level geography

in a spatial analysis, or omitting other geographical layers that might help identify individuals [130] such as schools or other features. In this case, while the integrity of the health data may be maintained, the potential environmental and contextual relationships are compromised. Either way, suppression of information makes for poor data, resulting in poor analyses and inappropriate decisions.

#### **6.4. DATA AGGREGATION**

*“The areal units used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating.” [133]\**

Another solution to the privacy issue that is frequently used in epidemiological analyses is that of aggregation. Data may be spatially aggregated – or grouped – in several ways, depending on a variety of factors, such as environment, demography, event distribution, etc. Unfortunately, aggregation is fraught with problems. For example, information is invariably lost in aggregation, some of which may be related to the quantities of interest [134] demonstrating the sensitivity of analyses – particularly cluster detection – to the areal boundaries aggregated to. Several papers have been published on the types and effects of aggregation, particularly its impact on disease cluster detection [9,10,16,135]. While there potentially exist an infinite number of different ways one could aggregate the data [133], this is typically done based on existing political and administrative boundaries which may have no bearing on the investigation at hand [44]. Aggregation to existing political and administrative boundaries may also be particularly problematic if it results in a study area larger than that of interest. For example, the distribution of incident cases at a very local level, such as in a neighbourhood in a small rural community, or even urban core, would most likely not be accurately reflected through aggregation. This is referred to as a “zoning” issue.

---

\* Quoted in the cited article from a 1984 book by S. Openshaw entitled *The modifiable areal unit problem* in the series *Concepts and Techniques in Modern Geography*, No. 38, by GeoBooks – the author of this study did not have access to this original source.

In the context of the current discussion, the issue of how large a geographical area needs to be in order to maintain anonymity is entirely a function of its population; smaller urban areas with a much larger population are more likely to achieve anonymity requirements and thresholds than larger rural areas with sparser populations. This concept, based on "Geographic Area Population Size" - GAPS - is elaborated on in the next chapter, since the author has contributed to research and publication in this area.

From a statistical and technical perspective, the smallest possible spatial unit of analysis is the most preferred [136]. Aggregated datasets are subject to a problem called the Modifiable Areal Unit Problem (MAUP), akin to the ecological fallacy in epidemiology. The MAUP consists of two related issues – scale and zone – both of which can change the results of analyses [45,136-140]. The issue of scale deals with the magnitude of aggregation; for example, a series of street addresses can be aggregated up to a neighbourhood level, to a census subdivision level, to a public health observatory level, and so on. The choice may be arbitrary, or it may be defined by law to prevent identification, or may depend on the analysis required and event under investigation. The larger the areal unit of aggregation, the more likely it is to conceal variations that would otherwise be visible at smaller scales [139-141]. Not only is the degree of spatial aggregation modifiable, but so is temporal aggregation, if a temporal component is present. The issue of zone has to do with the geographies to which the data are aggregated, and was mentioned briefly above.

Since aggregation is modifiable, the best recourse to considering such areal allocation is to visualize the point data – just as one would in the case of tabular data with a chart. Indeed it may, at times, be essential for the public health practitioner to be able to visualize the individual level data in order to help determine the best analytical methodology to pursue; this is, after all, the first step taught to budding statisticians – to plot and visualize the raw data. This is also argued to be a necessary first step in all spatial data analysis [140], particularly in real-time, such as in outbreak, emergency

and bio-terrorist planning and response scenarios. For example, during the SARS outbreak in 2003, Hong Kong cases were mapped at the individual building level in near real-time, thereby “providing a unique and rare GIS opportunity that resulted in some very comprehensive public Internet mapping services.” [44] Individual level data also provide maximum flexibility for healthcare delivery planning and application based on analysis of actual trends and patient locations. This is a rapidly growing use of GIS and spatial analysis in public health. In summary, data at the smallest level of detail are the best data for accurate and efficient analysis and interpretation [136], since information is unavoidably lost in aggregation [45]. From a location perspective, the smallest practical level of detail is the street address.

### 6.5. DATA ANONYMIZATION

Lying somewhat between data suppression and data aggregation is data anonymisation. This method of identity masking involves the slight alteration of data – by suppression if necessary or generalization if possible [42] – such that every record (or *tuple*) becomes indistinguishable from a certain minimum number of other records. The term *k-anonymisation* refers to this technique where every record becomes indistinguishable from at least  $k-1$  other records [8,42]. To clarify, consider the Table 5 below showing (fabricated) incident cases of tuberculosis in a Native Canadian community, and the corresponding 2-anonymized table (“First” and “Last” refer to first and last name; “Ident.” refers to the aboriginal identity, such as Metis, Inuit, or First Nations):

**Table 5: Fabricated tuberculosis cases and the 2-anonymised result**

First	Last	Gender	Age	Ident.
Rob	Jones	M	19	Metis
Mark	Smith	M	29	Inuit
Mary	Swan	F	27	Inuit
Rob	Tobias	M	17	First Nation

→

First	Last	Gender	Age	Ident.
Rob	*	M	15-20	*
M	S	*	20-30	Inuit
M	S	*	20-30	Inuit
Rob	*	M	15-20	*

Based on the above 2-anonymized example (i.e.  $k = 2$ ), each record is indistinguishable from at least one ( $k-1$ ) other record; the first and last records are indistinguishable from each other, as are the second and third record.

This method of anonymisation is somewhat complex and time-consuming [42], but has been adopted in a variety of applications. The terminology is also often used to refer to anonymisation levels, and will be used and discussed further throughout the remainder of this study.

## 6.6. ANALYTICAL SOFTWARE AGENTS

A relatively novel and innovative approach to the problem privacy poses to public health research is the introduction of software agents. Essentially, these intelligent applications would perform the spatial analysis on the disaggregate data whilst still in the possession of the data owner, and only return the results of the analyses to the researcher [15,16,142]. The agents can also be designed to explore issues of scale by performing multi-scale analyses [15], or be sent to simply return a visual (map) that masks the identity of individuals [16]. The issue with having agents automatically perform the analyses is that often-times one may not know exactly what type of analysis is optimal until one has visualized the raw data. As was previously mentioned, this is the first step taught to statisticians, and is cited as a necessary first step in all spatial data analysis:

*“D. Unwin (1996) specifically focuses on visualization as a necessary first step in all spatial data analysis, simply because the position of particular attribute values on a map induces associative processes in the analyst, drawing upon analogies, possible prior information, or memory (for instance of possible sources of data error).” [140]<sup>#</sup>*

However, given an appropriate “transforming” task as described below, an agent would be the perfect tool to return a visual result and data that have been transformed to mask identity.

---

<sup>#</sup> Taken in the cited article from Unwin DJ, GIS, spatial analysis and spatial statistics. *Progress in Human Geography*, 20:540-551 (1996) – the author of this paper did not have access to this original source.



## 6.7. TRANSFORMATIONS

Another approach to preserving the privacy of the individuals when dealing with health data is to apply a mathematical transformation while preserving the relative locations of the events (to each other). For example, one could translate all the cases by shifting them over  $x$  number of degrees in longitude, and  $y$  number of degrees in latitude; or one could rotate the points around a fixed centre, or introduce random perturbations within a fixed radius [16]. Comparisons can then be made with the original data – if available – to assess the efficacy, accuracy and validity of the applied transformations.

Much of the recent focus on location transformations for public health has been on random perturbation algorithms [112,143-145], as they tend to “outperform” others in terms of the privacy-loss of information trade-off [16]. These algorithms, however, tend to either ignore important attributes of the population and its distribution (such as age and sex, for example), or treat the anonymisation of the location information separately from these features. In the case of the latter, approaches have adjusted for the attributes on an aggregate basis by “weighting” the degree to which the location is perturbed. Instead, what is needed is an algorithm that adjusts for location as part of the overall anonymisation design, and not merely as a weighted afterthought.

Another more fundamental geographic problem is that transformations can also perturb the relationships of the events themselves to critical or influential data in the surrounding geography. For example, if an outbreak of influenza were to occur in a little town in a group of children, public health practitioners would certainly be interested in the distribution of the cases relative to one another – this relationship can be somewhat preserved in the currently proposed transformations, and the effects can be quantified. However, in considering time and place (the other two components of the epidemiological triad), the practitioner is also interested in the time of year, as well as proximity of the cases to schools, nurseries, nursing homes, etc. By shifting or transforming only the health event data, relationships to these other factors are

potentially lost. Transformations can therefore also have a profound effect on common spatial statistics used in public health, and an appropriate transform must therefore allow for the perturbation to be influenced by such factors if required.

As the name of this study implies, the focus is on this category of solutions, and a novel solution that builds location into the overall anonymisation design *and* allows for the integration of additional features was developed and published as presented in Chapter 8. Additional critical review of similar existing methods is also provided in that chapter as part of the manuscript. Paving the way for the exploration of this novel approach, however, were various other projects focusing on concepts fundamental to a proper understanding of the required goal.

## 7. Paving the Path To A Novel Transform

### 7.1. INTRODUCTION

Before the “a-ha” moment of creative inspiration, the journey toward it requires a deeper understanding of some of the core concepts at the heart of anonymisation algorithms. Two related concepts emerge: *uniqueness* and *re-identification risk*. The understanding and implementation of these concepts is also related to previously mentioned methods, in particular *aggregation* and *k-anonymity*. By relating these together we begin to form a more comprehensive and accurate view of anonymisation and how together they can inform a spatial transform.

### 7.2. UNIQUENESS

*Uniqueness* is the degree to which an individual is unique in a given sample or population and can be influenced by aggregation. Aggregation in this sense is not merely limited to spatial aggregation, but also non-spatial, in which it can also be referred to as generalisation or categorisation. The higher the aggregation, the lower the uniqueness; this in turn translates to higher anonymity. For example, a 97 year old male in a certain neighbourhood is likely to be unique; he is probably the only one. Aggregating several neighbourhoods may capture additional 97 year-old males, thereby reducing his uniqueness or making it more difficult to distinguish him from other 97 year-old males if other attributes are not disclosed. Since it is likely that one would have to aggregate over a very large area to capture another one or more 97 year-old males, an easier method would be to generalise age into categories. Revealing that this male is over 50 years old, for example, will more than likely capture a large group of males within this age group in the given neighbourhood. The spatial aggregation and age categorisation can also be combined to limit overall information loss, depending on what is appropriate for the purpose. For example, while generalising age in the neighbourhood to 85 years and over will likely not change the individual’s uniqueness level, doing this in combination with an aggregation over several neighbourhoods will increase this likelihood. In some circumstances, it may be impossible to change

uniqueness on a combined set of attributes (also referred to as an *equivalence class*) thereby necessitating the suppression of some attributes or of the record itself.

In order to assess the appropriateness of and requirement for a transform, a measurable uniqueness threshold should be identified. This is often reflected through the use of *k*-anonymity terminology. As described in the previous chapter, *k*-anonymity renders any given record with a set of attributes in a dataset indistinguishable from at least *k*-1 other records on all or a subset of those attributes. In this way, the level of anonymity afforded to the individual implicated by that record becomes quantifiable. Note that a *k*-anonymity threshold of  $k = 5$  means that any given record is indistinguishable from at least 4 other records on the selected attributes, or that the probability of correctly identifying that individual is 1 in 5 which is 20%.

Typically in the literature the reference for this assessment is the dataset itself; in other words, any given record in the dataset is indistinguishable from at least *k*-1 other records *within the same dataset*. However, this is not necessarily sufficient as the underlying population may also affect the effective uniqueness threshold – such as the example of the 97 year-old male given above. Furthermore, when considering the underlying population, uniqueness likelihood is generally inversely proportional to the population size; that is, as population size increases, the likelihood that an individual is unique on a given set of variables decreases [146]. It is, however, impractical to demand a uniqueness threshold of 0 (i.e. every single other individual in the comparison group shares the same values for the set of attributes of interest). This would result in extensive suppression and/or aggregation of values that would render the data useless for public health purposes. Some studies have suggested using population uniqueness thresholds of  $k=5$  and  $k=20$ , but no standards currently exist [146].

### 7.3. RE-IDENTIFICATION RISK

Re-Identification risk is the degree to which an individual can be identified from the dataset and is itself a multi-faceted concept. The underlying assumption in phrasing it as re-identification (as opposed to simply identification) is that the data have been “de-identified” and released, and an intruder then attempts to “re-identify” individuals.

Let us first address the relationship between identification risk and uniqueness. Within this context, different types of risk have been suggested. El Emam and Dankar describe *prosecutor* re-identification risk as an attempt to re-identify a record belonging to a specific individual known to be in the dataset, and *journalist* re-identification risk as an attempt to re-identify any individual simply for the sake of demonstrating being able to do so [147]. In the case of the former, the maximum probability of re-identification, as discussed in the previous section, is given by  $1/k$ , where  $k$  reflects the anonymity threshold within the dataset. In the case of the latter, the effective anonymity threshold is a function of a larger identifiable sample or population dataset accessible to the intruder. The same authors have also described a third type of risk which they refer to as *marketer* re-identification risk. This risk is described as an attempt to identify as many individuals as possible within a given dataset, also using a reference identifiable dataset [148]. Aggarwal and Yu have also published on these and other kinds of re-identification risk, though with different terminology [149].

It is vital to keep in mind that re-identification risk in the presence of additional information - be it in the form of knowledge or an existing repository - is not only a function of the equivalence class size but also of the overall uniqueness of the records. For example, let us assume that we know that John Doe, aged 32 years, has been tested for HIV and that a dataset containing patient gender, age and a binary HIV test result variable (positive or negative) is released. Because HIV status is perceived by most to be sensitive information, we transform the data to achieve 20-anonymity on age and sex by categorising age into 10-year intervals. If we look at the dataset and

find that it contains 25 males aged 30-39 years, then our equivalence class size is 25. In this case, we have improved on our re-identification probability of  $1/k$  since the probability of correctly identifying the specific record that belongs to John Doe is  $1/25$  or 0.04. Looking only at the equivalence class in this scenario, though, is quite useless, since not only does it not add anything, but it even tells us less than we already know (since we know his exact age)! If, however, of the 25 individuals in the equivalence class of interest, 20 tested positive, then we actually have a  $20/25$  or 0.8 probability of correctly inferring that John Doe tested HIV positive - a whopping 2000% difference!

Another aspect of re-identification more specific to location is related to the graphical release of the data. It has been shown that it is possible to use published maps of health events to identify individuals [11,150]. In such cases, while the transform may be sufficient to de-identify cases within a dataset, the visualisation of the cases in relation to their surrounding geography presents a re-identification risk.

Yet another re-identification risk of some location transforms is associated with repeated random perturbation. Repeated iterations of location transforms that employ controlled random spatial perturbations may increase the risk of re-identification simply because of the principles of central tendency and dispersion. This is particularly the case if the perturbed points are randomly placed within defined distances of the original without consideration of the population's spatial distribution patterns or underlying geographical features. As iterations of such transforms increase, so too does the risk of being able to identify the individuals concerned [16,144,151].

#### **7.4. MANAGING RE-IDENTIFICATION RISK**

As noted above, there are various considerations that impact re-identification risk, with uniqueness constituting just one of these. One method for managing re-identification risk through uniqueness is by controlling the geographic area population size (GAPS). Generally, the larger the GAPS, the less unique a given individual will be on a

controlled set of attributes. However, GAPS cut-offs are also variable, dependent on the number and nature of these attributes. To provide an empirically grounded basis for using GAPS cut-offs, a study using 2001 census data for Canada was conducted [58]. The study proposed a model for predicting GAPS cut-offs by aggregating nested geographies and assessing uniqueness on a combination of quasi-identifiers. The maximum number of possible combinations of these quasi-identifiers, referred to as *MaxCombs*, was used. The results demonstrated the variable nature of the GAPS cut-offs and suggested that for most of the Canadian population, individuals are anonymous on age and sex within the geography defined by the first three letters of Canadian postal codes. This geography is referred to as the Forward Sortation Area (FSA).

To further refine and confirm the findings of this study as related to the appropriateness of releasing information at the urban FSA level, a follow-up study was conducted using granular micro-data from the 2001 Canadian census [146]. Although these data represented a 20% sample of the Canadian population, permission to access and use them had to be granted through the Statistics Canada Research Data Centres (RDC) because of their granularity. Uniqueness thresholds of 5% and 20% were explored using the population size of the FSAs (FSA GAPS) and the *MaxCombs* for various quasi-identifiers. The models developed can be used to manage re-identification risk and provide recommendations on uniqueness threshold choice.

Since the micro-data were only available at the Census Tract level, however, a method was required to allow mapping of these Census Tracts to the corresponding FSA. The author devised such a method by creating a Canada grid relating the two geographies. The method was published with this study in *BMC Medical Informatics & Decision Making* in 2010 and is given below, as published.

#### 7.4.1. CREATING THE CANADA GRID

*Original contribution to co-authored study as published in BMC Medical Informatics & Decision Making (2010) [146]*

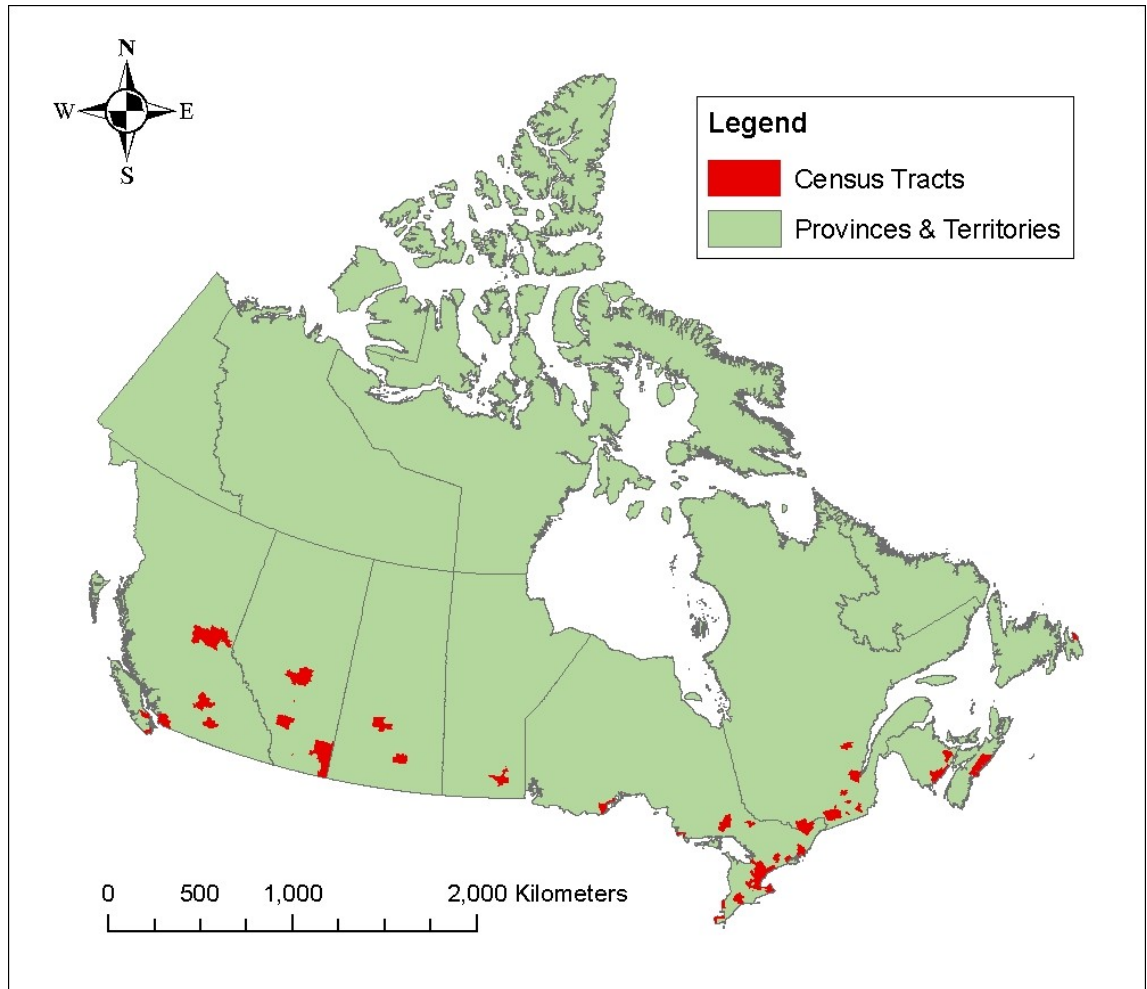
##### **Background**

The smallest geographic unit provided in the census micro-data file available through Statistics Canada's Research Data Centre (RDC) is the Census Tract (CT). CTs are only defined for census metropolitan areas and census agglomerations with urban core populations of at least 50,000 individuals. They are defined by Statistics Canada as "...small, relatively stable geographic areas that usually have a population of 2,500 to 8,000." [137] The 2001 census contained a total of 4,798 CTs distributed over 9 provinces (no CTs are defined for the Territories or the province of PEI; see Figure 4)

In order to compute re-identification risk by forward sortation area (FSA) in our current study, we needed to devise a method to estimate conversion between census and postal geography. A gridding methodology similar in nature to the Gridded Population of the World Project (GPW) [139] at the Center for International Earth Science Information Network at Columbia University [140] was utilized, allowing assignment of geography based on areal weighting using a population grid for Canada.



**Figure 4: Distribution of 2001 census tracts across Canada**



## Methods

Population-based weights were assigned to CT-FSA unions based on a created population grid for all of Canada. The grid cell size was one kilometre by one kilometre, and assigned populations were based on the 2001 census profile at the dissemination area level (DA). This is the smallest geography at which census profile information is released by Statistics Canada [45]. Similar to the PCCF+ product [134] from Statistics Canada, these population weights were then used to randomly assign census tracts to their associated FSAs. Details of the steps taken to create the population grid are described below.

Twenty six (26) complete grids of dimensions 1554 by 546 Kilometres were created using a script in ESRI's ArcMap 9.2 [152], as specified in Table 6. This created 848,484

one kilometre square cells per grid, for a total of 22,909,068 cells covering the Canadian landmass. The script for creating the grids was downloaded from the ESRI scripts Website [130].

Once the grids were created, the next task was to assign an estimated population to each cell. This was done using the Statistics Canada DA file [133]. First, all DA polygons identified as water (water code = 1) were removed. A new DA shape file containing only land DAs was created. DA boundaries were then dissolved so that DAs with disparate polygons were captured within one record. Areas and perimeters were summed for each polygon to give the total DA area and perimeter. This reduced the number of records from 62,015 to 52,924. The 2001 DA populations were then extracted from Statistics Canada's 2001 Census files using Beyond 20/20 [141]. Total population, as well as sex and age-stratified populations were extracted for all DAs across Canada, using four separate profile files (Western Canada and the Territories, Ontario, Quebec and Atlantic Canada). The DA variable name was renamed to DAUID to match the boundary file naming convention, and appropriate names were given to the population variables. Next, the 2001 DA population file was joined with the 2001 DA boundary file to create a 2001 Canada DA boundary file containing total and sex and age stratified populations.

**Table 6: Canadian grid development sections and attributes**

Grid Section	x	y	Rows	Cols.	# Cells	# Cells (DA-clipped)	# Cells (populated DA-clipped)
00	-2341699	310266	1554	546	848,484	147,282	95,225
01	-1795699	310266	1554	546	848,484	323,759	292,052
02	-1249699	310266	1554	546	848,484	400,335	352,048
03	-703699	310266	1554	546	848,484	421,104	252,417
04	-157699	310266	1554	546	848,484	442,583	112,863

**Table 6: Canadian grid development sections and attributes (continued)**

<b>Grid Section</b>	<b>x</b>	<b>y</b>	<b>Rows</b>	<b>Cols.</b>	<b># Cells</b>	<b># Cells (DA-clipped)</b>	<b># Cells (populated DA-clipped)</b>
05	388301	310266	1554	546	848,484	444,187	47,006
06	934301	310266	1554	546	848,484	588,000	220,587
07	1480301	310266	1554	546	848,484	514,762	202,006
08	2026301	310266	1554	546	848,484	222,848	139,035
09	2572301	310266	1554	546	848,484	79,825	30,635
10	-2341699	1864266	1554	546	848,484	490,304	181,644
11	-1795699	1864266	1554	546	848,484	843,129	253,796
12	-1249699	1864266	1554	546	848,484	753,391	84,386
13	-703699	1864266	1554	546	848,484	749,156	802
14	-157699	1864266	1554	546	848,484	563,822	1,239
15	388301	1864266	1554	546	848,484	192,569	1,005
16	934301	1864266	1554	546	848,484	587,718	1,420
17	1480301	1864266	1554	546	848,484	342,289	683
18	2026301	1864266	1554	546	848,484	220,305	48,694
19	2572301	1864266	1554	546	848,484	55,829	25,720
20	-2341699	3418266	1554	546	848,484	21,506	0
21	-1795699	3418266	1554	546	848,484	168,942	531
22	-1249699	3418266	1554	546	848,484	135,498	686
23	-703699	3418266	1554	546	848,484	229,560	0
24	-157699	3418266	1554	546	848,484	424,214	1,101
25	388301	3418266	1554	546	848,484	258,726	210
26	934301	3418266	1554	546	848,484	26,188	160
<b>TOTAL</b>					<b>22,909,068</b>	<b>9,647,831</b>	<b>2,345,951</b>

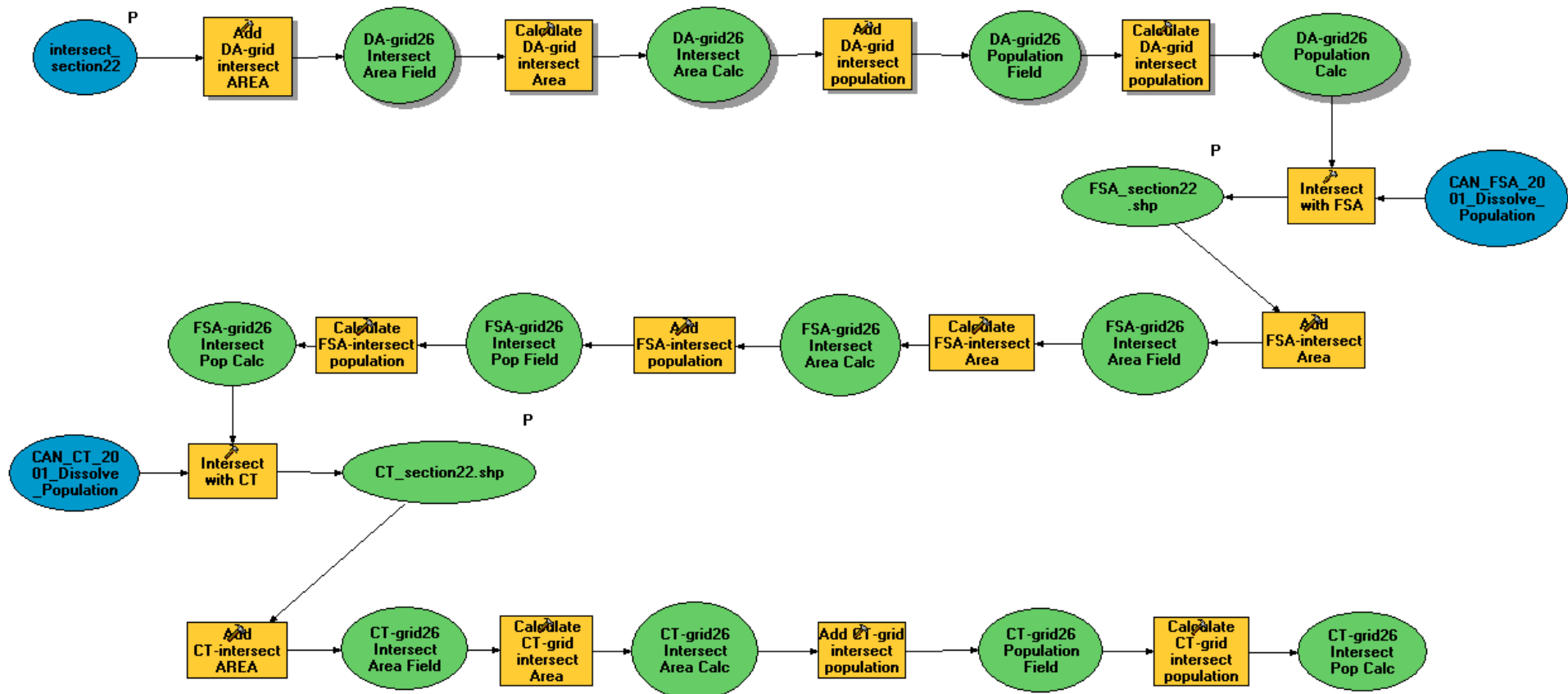
A “Select by attributes” function where population was not zero (0) was completed on the above file to create a new boundary file containing only DA polygons with reported populations. This further reduced the number of records to 49,153, creating a boundary file for non-water, populated DAs only. A “Select by location” function was completed on all 26 grids, for any cells that intersected the boundary file from the previous function. The resultant grids had a combined total cell count of 2,367,457.

A model (Figure 5) was created using the ArcGIS model builder, and run for each of the 26 grids, to create grid section intersects with 2001 DAs, FSAs and CTs. The model also calculated proportional grid sub-section areas and the corresponding population, based on underlying DA population and an assumption of uniform population distribution within each of the geographic areas.

A summary was done by each CT-FSA combination, to create unique CT-FSA records with the corresponding sum of the calculated grid-section populations. These summed populations were then divided by the total sum of the gridded-CT population to give the proportion of the population in each CT that lay within the corresponding FSA. In essence, this creates a population-based weight for each CT-FSA combination, allowing us to randomly assign any given record within a CT to its most likely (population-weighted) FSA.

A simplified hypothetical example of the end result is given in Table 7 and Figure 6. In this example, 64.07% of the population in CT16003 is found in FSA *K2S*, and 35.93% in FSA *K2T*. For CT 16004, 49.35% of its population is in *K2R*, 19.48% in *K2S* and 31.17% in *K2T*. This reduces the table to five rows, with a population-based weight for each unique CT-FSA combination. If, for example, there were then 28 records from the micro-data file falling in CT 16003, 18 (~65.86%) would be allocated to *K2S*, and 10 (~34.14%) to *K2T*.

Figure 5: ArcGIS model for building grid-Dissemination Area – Forward Sortation Area – Census Tract intersect polygons

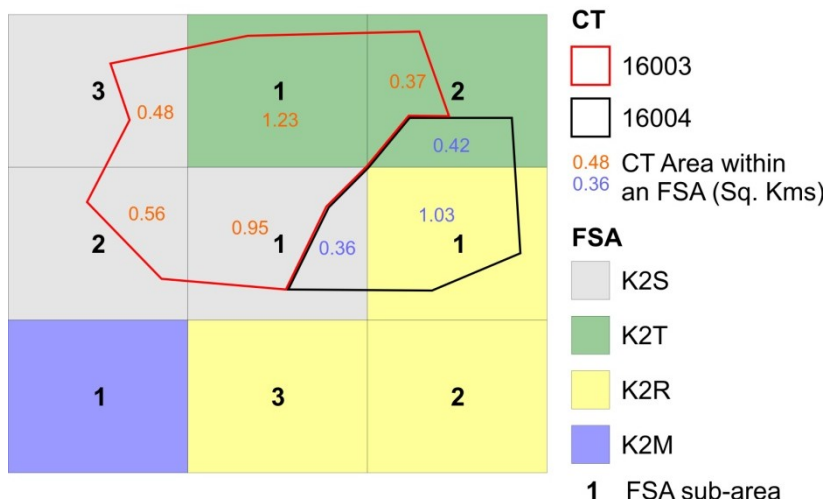


**Table 7: Simplified hypothetical example of the weighted association between Census Tracts and Forward Sortation Areas**

CT	FSA	FSA sub-area	FSA sub-area Pop Density (/Sq. Km.)	CT Area in FSA (Sq. Km.)	CT Pop	FSA Pop	Weight
16003	K2S-1	50	0.95	48	128	0.3750	
16003	K2S-2	25	0.56	14	128	0.1094	
16003	K2S-3	42	0.48	20	128	0.1563	
16003	K2T-1	20	1.23	25	128	0.1953	
16003	K2T-2	56	0.37	21	128	0.1641	
16004	K2R-1	37	1.03	38	77	0.4935	
16004	K2S-1	42	0.36	15	77	0.1948	
16004	K2T-2	56	0.42	24	77	0.3117	

CT is Census Tract; FSA is Forward Sortation Area; FSA sub-area is FSA sub-area; Pop is Population

**Figure 6: Example Census Tract - Forward Sortation Area sub-area overlay to illustrate the hypothetical example**



## Results

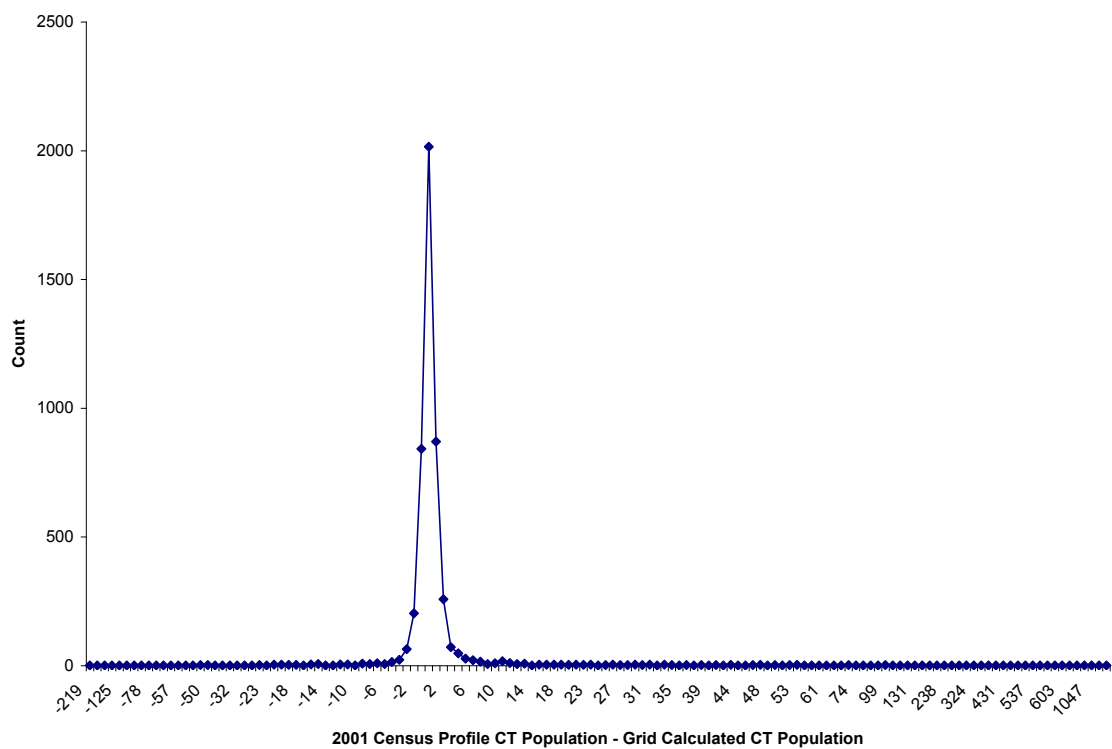
The CT population assignments based on the gridding methodology proved to be very similar to the 2001 Statistics Canada Census Tract population profile (Table 8).

The mean difference between the populations was 3.45 individuals, with a standard deviation of 48.96 individuals (median was 0). A graphical representation of the distribution of the population differences, by census tracts, is given in Figure 7.

**Table 8: Census Tract population comparison between created population grid and 2001 census profile**

	2001 Statistics Canada Population Profile Census Tract	Canada Population Grid Project Census Tract
Total n	4757	4757
Mean population	4413.99	4410.54
Standard Deviation	1911.77	1911.33
Minimum population	40	0
Median population	4290	4287
Maximum population	20635	20636

**Figure 7: Distribution of Census Tract Population Difference between Grid-Calculated Population and 2001 Census Profile**



Provincial analyses also showed a high concordance between the CT populations using the gridding methodology as compared to the 2001 Statistics Canada Census Tract population profile (Table 9). The greatest differences were in New Brunswick (mean difference = 6.97 individuals, standard deviation = 75.26 individuals) and Alberta (mean difference = 6.75 individuals, standard deviation = 81.67 individuals).

**Table 9: Provincial differences between Profile and grid Census Tract populations**

	NL	NS	NB	QC	ON	MB	SK	AB	BC
N	45	85	70	1246	2001	164	101	449	596
Mean	3.71	2.6	6.97	1.55	3.68	2.93	-1.18	6.75	4.79
Std Dev	12.01	19.37	75.26	26.19	51.38	27.14	37.86	81.67	51.38
Median	0	0	0	0	0	0	0	0	0

## Conclusions

The population grid created in this study provides a means for linking census geography to postal geography in Canada. While creating population grids in and of itself is not a novel idea, the created grid in this project allows the mapping of census geography to postal geography, based on population weights. The procedure assumes a uniform population distribution within the geography being used. However, since CTs only occur in highly populated urban areas, this was felt to be an appropriate assumption. A similar assumption would not hold in rural or less densely populated areas, and this technique would therefore not be appropriate. However, it could be utilized, and further refined, by incorporating additional information, such as ecumene areas, satellite imagery for residential and inhabited areas, address data, etc.



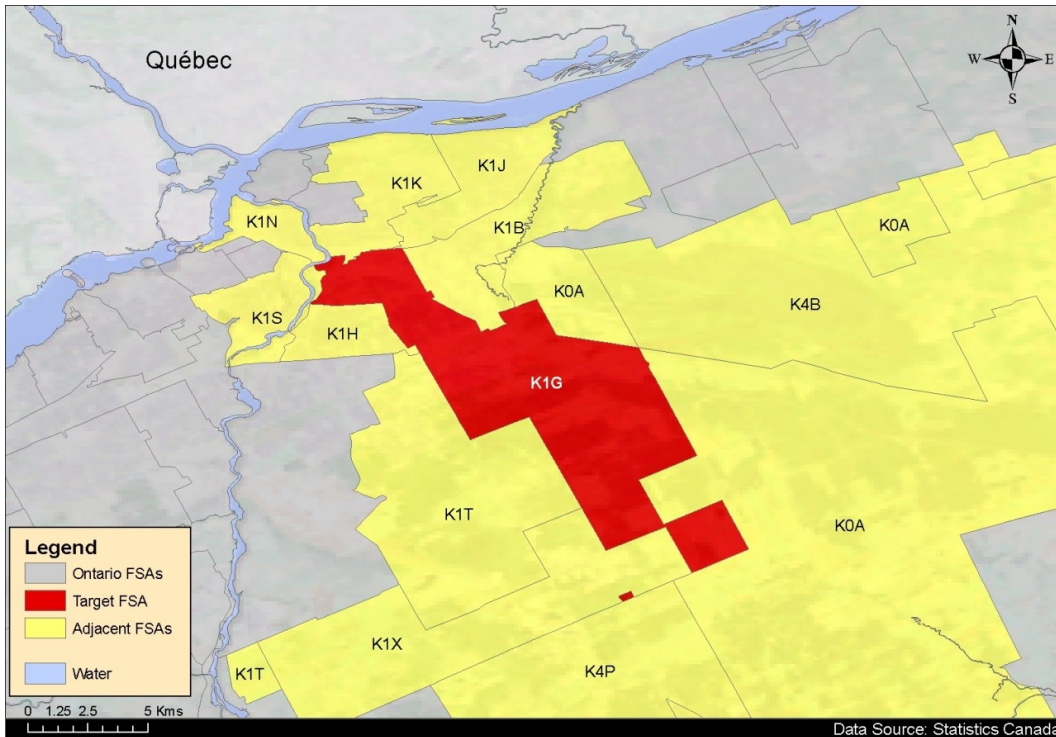
## 7.5. RETHINKING SPATIAL AGGREGATION

***Aspects of this section were submitted to the Public Health Agency of Canada in the form of a report as included in Appendix G (Volume II).***

While clinical health information at the individual level is typically associated with the patient's full residential address, it is seldom if ever available to public health practitioners. Instead, the patient location is often provided at a much less granular scale, such as the zip code in the United States, or the postal code or forward sortation area in Canada as described above. In these instances, all cases falling at the same geography are typically geocoded such that they are geographically represented at the same point - usually this is the geographic or weighted centroid of the polygon represented by that geography. In the two studies described above, aggregation was really a case of nested geographies – instead of postal codes, the larger geography of aggregated postal codes with the same first three characters (FSA) was used. However, what if aggregation is required across non-nested geographies? A study was therefore also conducted looking at aggregation across urban FSAs. The methodology, while illustrated with FSAs, can be applied to any geography, since as has already been demonstrated the urban FSA in Canada is generally sufficiently large for data release (though this depends, as has been previously mentioned, on the number and nature of attributes).

The algorithm aggregates adjacent geographies (Figure 8), so the first step in the study was to generate an adjacency matrix for Canadian FSAs. This indicates, for each FSA, all other adjacent FSAs in the same province. A first-order adjacency matrix was conducted by the author using the GIS software ArcMap 9.2.

**Figure 8: An example of a Forward Sortation Area “K1G” and those adjacent to it.**



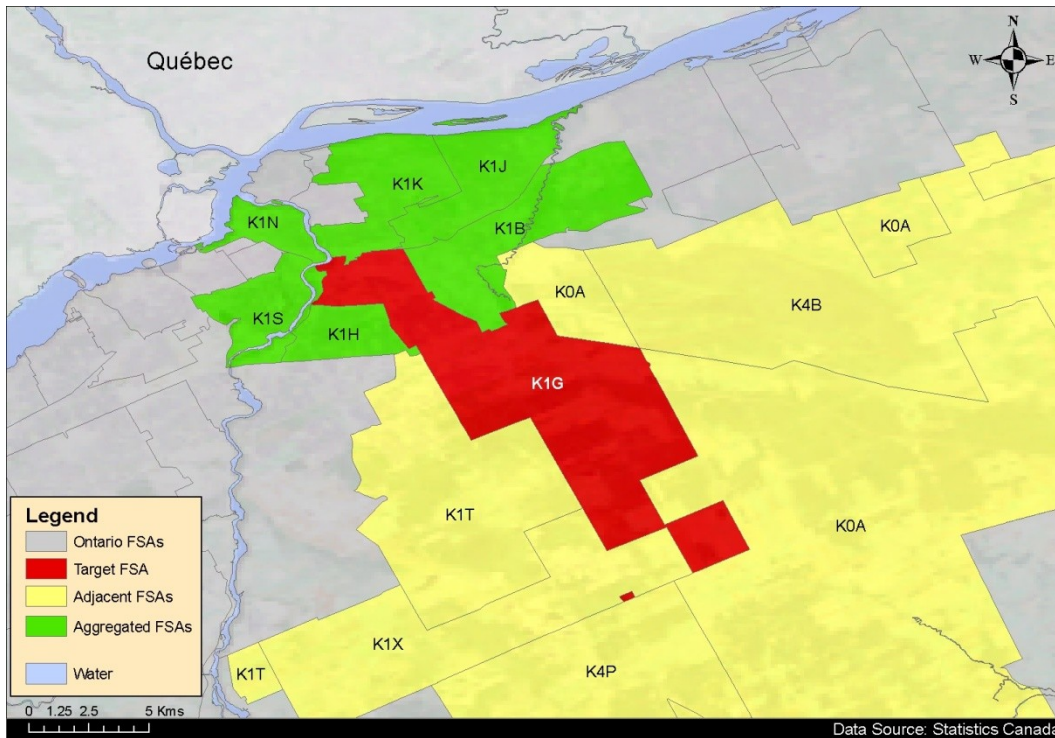
Note that one FSA can be a collection of multiple nested, contiguous or non-contiguous polygons

The aggregated areas also had to be minimally dispersed; in other words, distances between adjacent FSAs had to be minimised, thereby favouring aggregation in “clusters” as opposed to “long strips” (Figure 9). This was accomplished by defining a homogeneity metric which reflects the proportion of FSAs of interest that a given “eligible” FSA is adjacent to in a cluster formation. For example, consider the FSAs in Figure 10 around the capital city of Canada, Ottawa. Our primary FSA that we are aggregating to is K1G, and we have begun by combining it with the FSA denoted by K1N. These, then, are our FSAs to aggregate to. As the figure illustrates, the FSA denoted by K1K is adjacent to both of these, and is therefore adjacent to 100% of the FSAs we wish to aggregate to; it therefore has a homogeneity value of 1. The FSA denoted by K1X is only adjacent to K1G, and therefore is only adjacent to 50% of the FSAs we wish to aggregate to; it therefore has a homogeneity value of 0.5. By setting a minimum threshold homogeneity requirement, we can control the degree of connectivity of the aggregated geographies and therefore reduce spatial disparities.

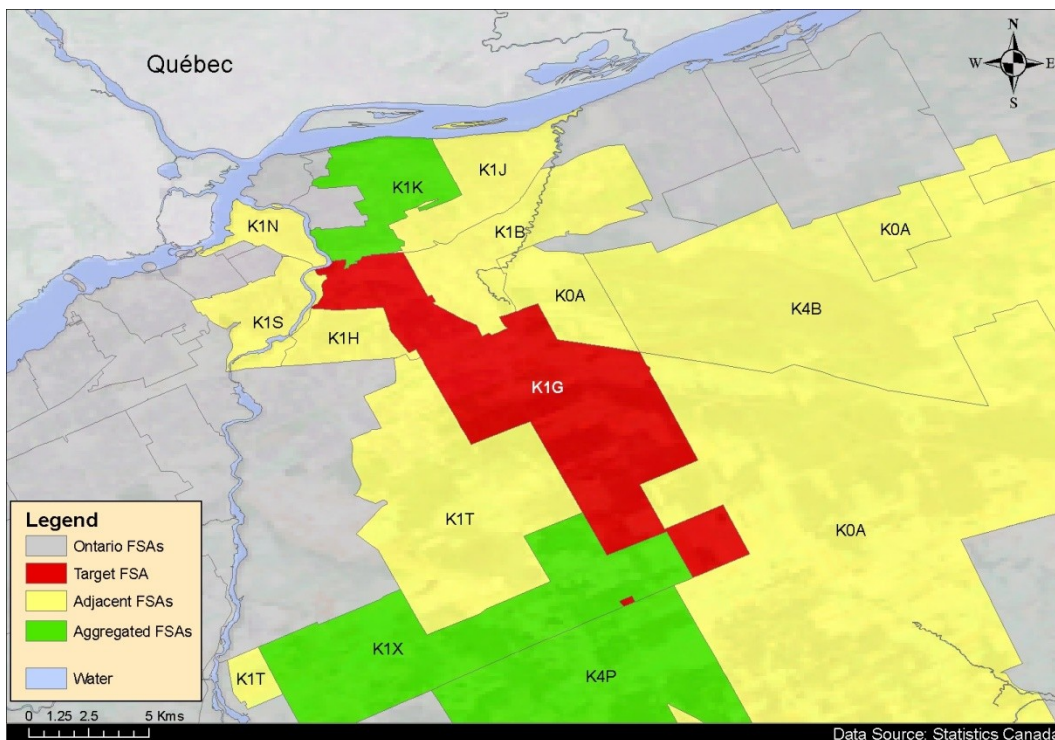
In addition to reducing spatial disparity, however, we must also minimise the amount of population “noise” added through aggregation. For example, given two FSAs eligible for aggregation with the same homogeneity value, how do we decide on the one to use? If one has a population of 20,000 and the other a population of 30,000 then aggregating the latter will add more noise to the end result. In order to adjust for this in scenarios where the homogeneity values are not the same, this value is therefore multiplied by the population size and the FSA with the lower result is given preference.

By using the MaxCombs to predict the acceptable population size cut-off (GAPS), this aggregation algorithm can then be used to provide a better aggregation of geographies to meet the required GAPS. The method improves on existing aggregation methods by incorporating information on the quasi-identifiers as given through MaxCombs, maximising the aggregation of spatially proximal geographies and minimising the amount of noise added to achieve the required GAPS.

**Figure 9: Aggregation options for Forward Sortation Area polygons adjacent to K1G: (a) “clustered”, minimizing distance between aggregated FSAs, and (b) “string”, where aggregation is stretched based on other parameters, irrespective of geography**

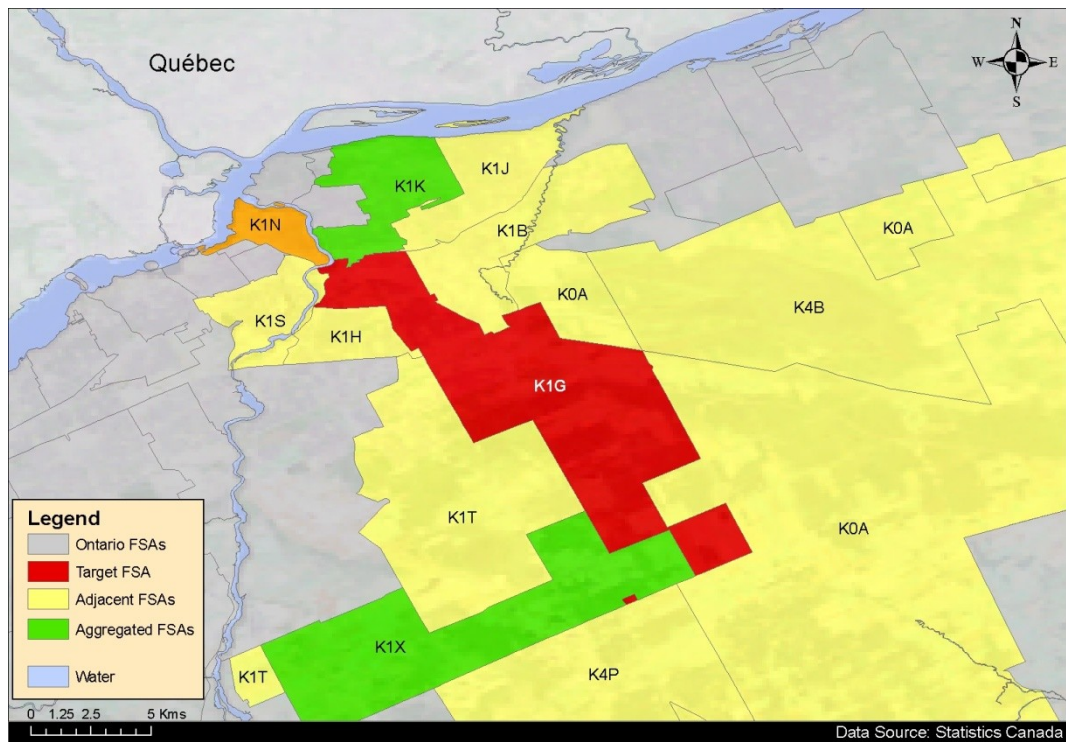


(a)



(b)

**Figure 10: Illustration of the “homogeneity metric” based on the adjacency of Forward Sortation Areas**



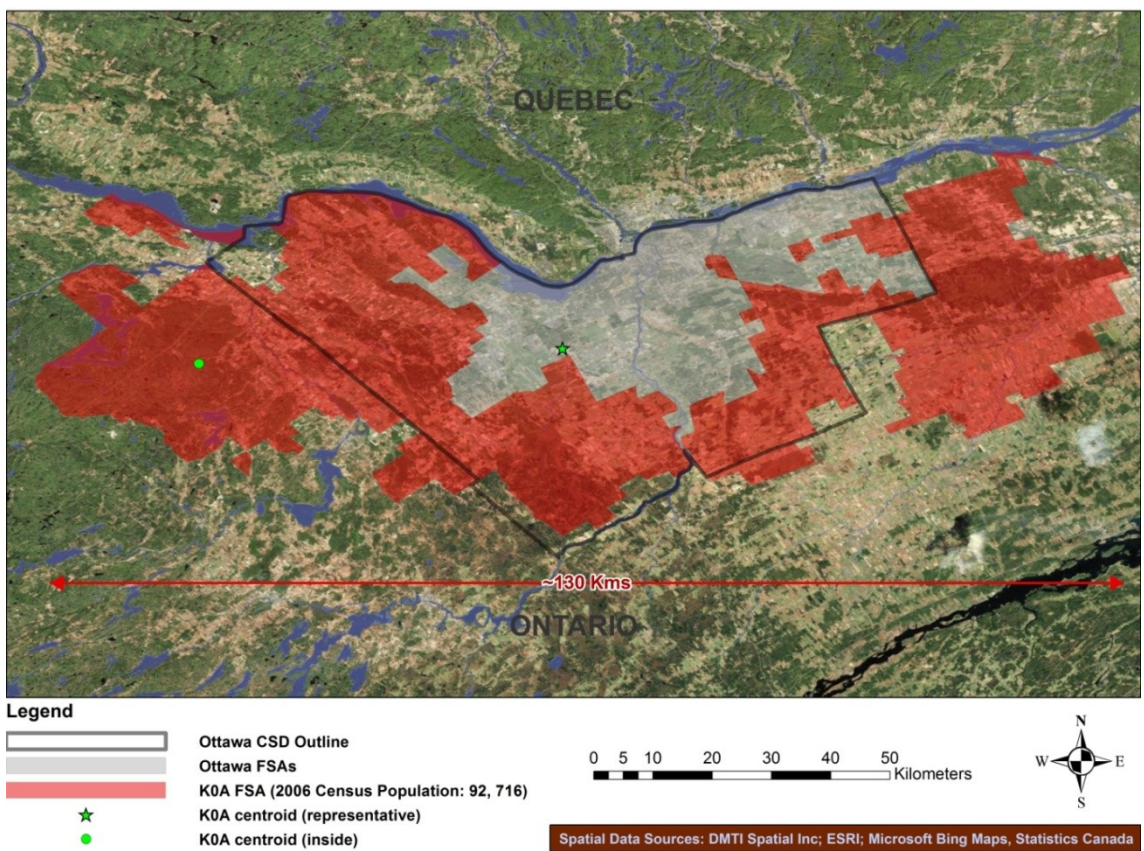
Considering two adjacent Forward Sortation Areas, K1G and K1N, this figure shows how K1K is adjacent to both, and therefore has a homogeneity metric of 1. K1X, however, is only adjacent to K1G and therefore has a homogeneity metric of 0.5

It is worth noting that this methodology is only described with FSA geographies for convenience because of the availability of postal codes. In urban areas, these are sufficiently small that they may, in some studies, be an appropriate unit of aggregation. However, consider the scenario in which toxic substances are released into the environment by a source located at a given point. It is highly unlikely for such a source to be located smack in the middle of an urban area, and for a study investigating cases presenting with a given related toxin, the residential location of the case relative to the source may be critical. Aggregating to the FSA in these cases may lead to erroneous conclusions. Figure 11 shows an obvious example, where not only does the rural FSA "K0A" in Ottawa, Ontario, Canada extend over 130 kilometres, but it also crosses into multiple census subdivisions. The 2006 census population of this FSA was 92,716 individuals. Note that the 130 kilometres is the shortest possible (geodesic) distance from one end to the other; minimum travel distance, for example, on the road network, will be at least this much, with travel through urban (i.e. likely high-traffic) areas of

Ottawa. Assuming a source of interest that lies at the east end of this FSA, aggregating cases to the FSA will place **all** of them either at the representative centroid shown by the green star, or, if the centroid is forced to fall inside the FSA polygon, the green circle - both shown in Figure 11. This effect can be mitigated by either using more granular areas (such as smaller census boundaries), creating one's own geographic areas or making use of the grid methodology described previously in this chapter. It is therefore crucial to consider the impact of aggregation and the context in which it is carried out.

Additional aggregation-related thoughts are discussed briefly in Chapter 12 of this study as there is potential for further enhancement and future implementation of spatial aggregation methods.

**Figure 11: Extent of the K0A Forward Sortation Area and the Ottawa Census Subdivision in Ontario, Canada (2006 Census)**



## 8. The Multidimensional Point Transform

*A condensed version of this chapter was submitted as an original research article and published ahead of print in *Methods of Information in Medicine* (2011) [153]*

### 8.1. BACKGROUND

Privacy, as related to personal and identifiable health information, has repeatedly been a subject of contention within public health practice and health research: the literature is littered with comments and complaints [65,154,155], surveys have sought to assess the perceptions of public health professionals and the general public [14,51,70,156,157], and both privacy advocates and public health professionals appeal to a vaguely painted patchwork of legislation [154,158]. The issue does not generally lie with direct identifiers such as name or an identifying number, but rather with attributes or variables that are not in themselves identifiable but that can be used in combination to re-identify individuals. These are referred to as *key attributes* [159] or *quasi-identifiers* [160]. As an example, age and sex are commonly used public health quasi-identifiers that have been characterised as having “high utility to an intruder” attempting to re-identify individuals from a dataset [146].

As public health methods advance with ever-evolving technology to better capture the entire context within which health events occur, requirements for privacy–protective methods also increase. In an attempt to address the privacy issue, algorithms for anonymisation and privacy enhancing techniques (PETs) have been proposed and implemented [161], and calls for public health professionals to challenge policies and lobby legislators have been made [65].

As has been described, anonymisation algorithms are often measured as a function of indistinguishable records (uniqueness) and re-identification risk (measured as a

probability of re-identification). The term *k-anonymisation* refers to the concept where every record becomes indistinguishable from  $k-1$  other records [8,42]. While no standards for acceptable anonymity thresholds have been established for public health,  $k$  values of 5 and 20 (representing re-identification probabilities of 20% and 5% respectively) have been suggested and used in the literature [69,146].

One particular area that has seen a dramatic increase in concern is location privacy, particularly given the increasingly recognised importance of spatial information in public health [44,107]. With the ubiquitous use of Global Positioning Systems (GPS), online mapping applications that provide high-resolution aerial images, and the increasing use of spatial intelligence in public health, location privacy is becoming increasingly contentious – perhaps more so than with other information technologies [107,158]. The realisation of these issues is not particularly novel. Over a decade ago, Armstrong *et al.* published a paper on various mathematical transformations to mask original point location [16]. The authors describe three main masking categories: those that transform across records, for example aggregating all records within administrative or political boundaries; those that transform attributes, such as categorising age, for example, or suppressing ethnicity; and those that displace records such as spatially moving a point from one location to another by adding noise to its coordinates. Of the methods described, random perturbation was found to perform best overall as measured by retention of pair-wise relations, event-geography relations, clusters, trends and directional relationships (anisotropies) [16]. Other public health studies – including this study – have therefore continued to build on this type of spatial transform [112,143-145] and a good overview can be found in [107].

In a classical random perturbation, a circle of radius  $r$  is drawn around the point to be masked such that sufficient population is captured to render the point anonymous, and the point is randomly displaced within the area. This is repeated for each point, and if  $r$  is sufficiently large (i.e. captures enough population), one ends up with a series of



points that are difficult to trace back to their original locations due to the stochastic nature of the transform. Not all random perturbations are created equal, however, and advances in their development and implementation have been slow. Ideally, the displacement as measured by the perturbation distance should be minimised, and generally, the more densely populated the area in which a point (case) falls, the less it has to be spatially perturbed to meet a desired anonymity threshold. Adjustments to random perturbation therefore create dynamic radii dependent for each point on its underlying population [143]. This "context-sensitive approach" can be further improved by stratifying on other attributes, such as age and sex, to give a more accurate displacement that minimises information loss [143]. More elaborate revisions of random perturbation have been developed in recent years, including the use of Linear Programming (LP) [144] and a "donut" method of geomasking [145]. However, all of the proposed versions of these transforms modify location-based information almost as an afterthought or secondary anonymisation technique, either assuming that all other identifying information – including important quasi-identifiers such as age and sex – has already been anonymised or stripped, or adjusting the transform accordingly for selected underlying demographics using generalised weighting schemes. Instead, what is needed is a transform that operates discreetly on multiple attributes, in concert with location as part of the overall anonymisation algorithm.

## **8.2. OBJECTIVES**

The current methodology refines the random perturbation approach by combining new and previously studied methods to propose a flexible, dynamic and customisable multidimensional point transform (MPT) acting on attribute data. In this context, attributes of interest – such as location, age, sex, education, etc. – are referred to as *dimensions* since they define the scope of the transform. Like previous context-sensitive studies [112,143,145], the approach presented is an adaptive geomask. However, unlike others, it allows these other dimensions to be incorporated into the

anonymisation algorithm directly based on custodian and user tolerances and requirements.

## 8.3. METHODS

### 8.3.1. ALGORITHM: OVERVIEW

The proposed algorithm is dependent on the availability of a base population (real or synthesised) matrix,  $A$ , of  $N$  individual records with  $Q$  attributes. The dimensions of interest must be elements of the attribute set, and given the spatial nature of the transform, must include a location attribute – ideally the geographic coordinates of the individual’s relevant address. Given a list of patients,  $B$ , from this base population  $A$ , the goal of the algorithm is to randomly "move" each patient in  $B$  within a maximum perturbation distance  $\Delta$ , while controlling on all dimensions of interest for a defined anonymity threshold,  $k$ . "Move" in this case means randomly selecting an alternate record from  $A$  to represent the patient; in this way, the *locations* are realistic and non-random, but the *selection* is random.

Consider the simple example where the controlled pre-selected dimensions of interest are location, age and sex (other dimensions can be added, provided they are elements of both datasets). In other words, the algorithm's function is to ensure that the anonymity level  $k$  is maintained based on the location, age and sex of each individual in the dataset and does so by comparing the number of other individuals within a given distance matching on age and sex as required. While ensuring that the anonymity threshold  $k$  is maintained based on these dimensions, the algorithm sequentially perturbs or masks them as required based on pre-defined conditions and perturbation tolerances. Location perturbation is measured as the distance moved from the original point, and its maximum tolerance is defined by  $\Delta$ ; the age perturbation tolerance allows the dimension to be categorised, for example in 1-year increments, up to a maximum number of categories; and the sex perturbation tolerance is binary, either requiring a perfect match on gender or not.

The acceptable anonymity threshold is defined by  $k$ . For example,  $k = 5$  means that a given patient is indistinguishable from at least  $k - 1 = 4$  other individuals within the selection area. So if, within 300 metres of a male patient aged 12 there are four other male patients aged 12, then randomly choosing any one of them including the original has a 20% chance of correctly identifying the patient. The maximum perturbation distance  $\Delta$  is the maximum acceptable threshold for spatial displacement. This does not mean that all eligible records for displacement will be up to  $\Delta$  away from the original point, only that this is as far as the algorithm is allowed to go to achieve the desired  $k$ . The actual maximum perturbation radius,  $R$ , will depend on the data and defined  $k$ .

Given the patient dataset  $B$  with  $j = 1$  to  $n$  patients, all patients in  $B$  are removed from  $A$  to give the complement non-patient base population,  $C$ . Removing the patient dataset individuals from the base population records at the onset of the algorithm has two key effects: it prevents selection of one patient in place of another, and it reduces re-identification risk by forcing  $k - 1$  to consist entirely of non-patients. Next, for each record in  $B$ , all records in  $C$  matching  $B_j$  on sex and age are isolated and the distance between each one and  $B_j$  is calculated. If fewer than  $k - 1$  matching records are found within  $\Delta$ , then the sex and age dimensions are perturbed (i.e. grouped or categorised), based on the pre-defined conditions and in parallel in both the case dataset and the population dataset, and the matching is re-done. This is repeated until at least  $k - 1$  matching records are found. If the algorithm is unable to reach the desired  $k$ -anonymity, then the record is non-transformable within the current requirements, is flagged as such, and the algorithm proceeds to the next record. Otherwise, the algorithm continues.

Of the matching records, the closest  $k - 1$  records are identified and a small random distance,  $\delta_r$ , is added to the farthest  $k - 1$  match distance,  $\delta_{\max}$ , to define the perturbation radius  $R$ . The addition of this random distance ensures inclusion of the

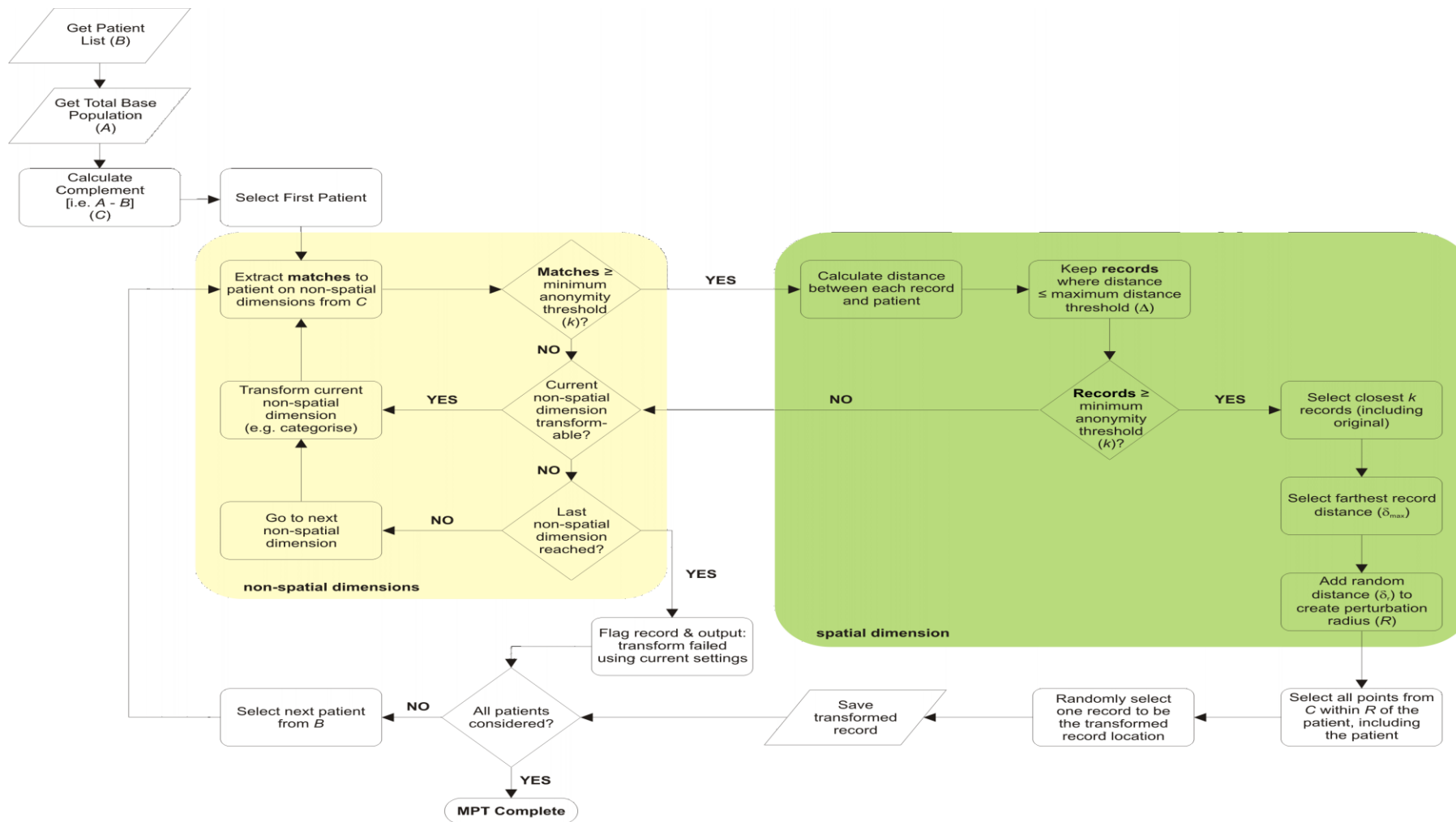
point used to set the farthest match distance in case any rounding occurs and guarantees a minimum  $k$ -anonymity. It also adds a small stochastic aspect that complicates re-identification of the original patient location. This is because not only does the selection of the transformed point become different with each run, but so also does the underlying pool from which the point is selected. A record within  $R$  of  $B_j$  is then randomly selected from  $C$ , and its location assigned as the perturbed point. This is repeated for the next record until all patient records in  $B$  have been transformed or flagged. The algorithm flow and components are illustrated in Figure 12.

### 8.3.2. DATA

Synthesised population data for New York County were acquired from the MIDAS project [162] by request. The dataset contained synthesised records at the individual level, with the dimensions of interest being age, sex and residential location (latitude and longitude in decimal degrees). For each record, latitude and longitude were converted from decimal degrees to radians prior to algorithm execution for use in extent and distance calculations.

New York county was specifically chosen as the study area to allow for comparisons with existing published methods - namely the results of the LP approach by Wieland *et al.* [8] - on distances required to achieve specified  $k$ -values when additional dimensions are taken into consideration. The two approaches are also similar in that they both seek to minimise perturbation distance and both rely on the presence of underlying spatially-referenced population data.

Figure 12: Multidimensional Point Transform flow



### 8.3.3. ALGORITHM: PRELIMINARY PROOF-OF-CONCEPT IMPLEMENTATION

Preliminary testing of the algorithm was completed using Monte Carlo simulations for various patient sizes. One thousand iterations were run for each of 25, 50, 100, 200 and 400 patient datasets, generated by randomly sampling records from the synthesised New York County population.

The controlled dimensions were sex, age and distance. The anonymity threshold  $k$  was set to 5 and the maximum perturbation distance  $\Delta$  to 1 kilometre. An exact match to sex was required (i.e. no perturbation allowed), and 5 levels of age categorisation were permitted (including exact age). Age categories were created by increasing the age range by one year for each successive level: for the first level, age range is 0, so it is the exact age; for level 2, the range is 1 to give age categories 0-1, 2-3, 4-5, etc. A simplified illustration of the implementation of the age categorisation is given in Figure 13. Note that level 5 matches the age categories generally used in census profiles and population surveys for age and sex stratified population counts such as the Canadian census [163], the United Kingdom census [164] and the American Community Survey [165].

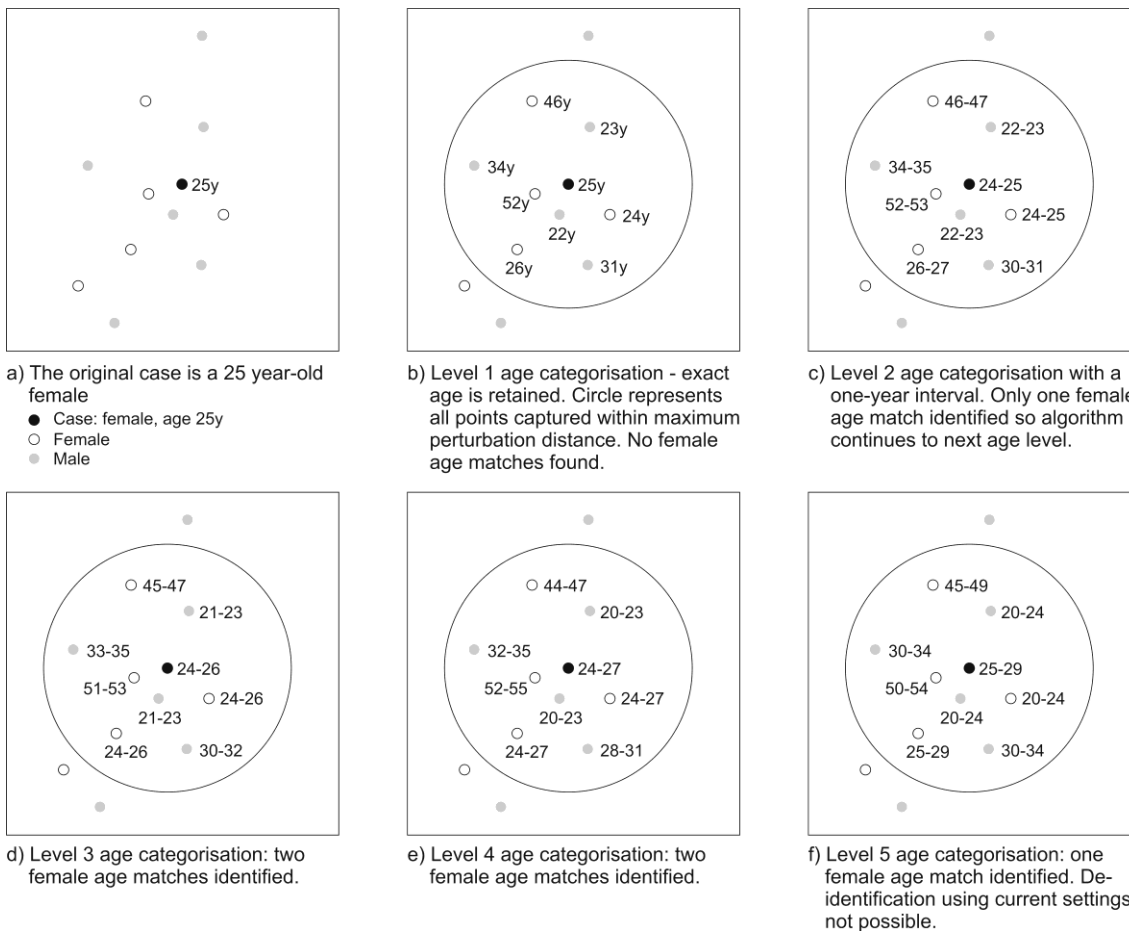
Extent-limiting steps were also added to the algorithm to improve computational performance. At the beginning of each iteration and after creation of the patient dataset, one kilometre was added to the maximum and minimum latitudes and longitudes of the patient dataset and used to constrain the extent of the base population (i.e. reducing the number of base records used in the algorithm, thereby improving performance). This method was also used when determining the eligible population for each record.

The small random distance,  $\delta_r$ , added to create  $R$  was restricted to a range of 1 and 10 metres to minimally impact geographic displacement. Distance was measured using the great circle formula.

Cumulative descriptive statistics (mean and standard deviations, as well as the minimum, maximum and median) were calculated for successive iterations to assess the effects of the transform on the perturbation distance. Analysis of the age dimension sought to identify the proportion of records requiring categorisation on age to achieve the required minimum  $k$  for each patient dataset size. The effect of adding the small random distance  $\delta_r$  on  $k$  was also described through descriptive statistics, as was the final perturbation radius. The displacement of the spatial mean of each patient dataset was also calculated in terms of perturbation distance (i.e. distance moved or displaced).

The algorithm was coded and run in SAS v9.1; the results were also analysed in SAS v9.1 and graphed using Microsoft Office 2007. The code was run on multiple dual-core machines in Random Access Memory (RAM) using RAMDisk software [166], as preliminary tests showed this to be much faster than mechanical hard drives.

**Figure 13: Simplified example of age categorisation using one-year intervals with 5 levels and k=5.**



Note that categorisation always starts at age 0 years (i.e. birth, consistent with census age strata)

## 8.4. RESULTS

In total, 775,000 records were randomly chosen from the synthesised New York population of 1,482,104 unique individuals and tested with this algorithm. These represented 601,790 unique individuals, thereby capturing 41% of the New York county population (Table 10).

Time taken to complete the algorithm ranged from about 5 minutes per iteration for the 25-patient dataset size, to just under two hours per iteration for the 400-patient dataset size.

The age dimension was seldom transformed, as summarised in Table 10. Only one record required an age-transform in each of the 50, 100 and 400-patient datasets,



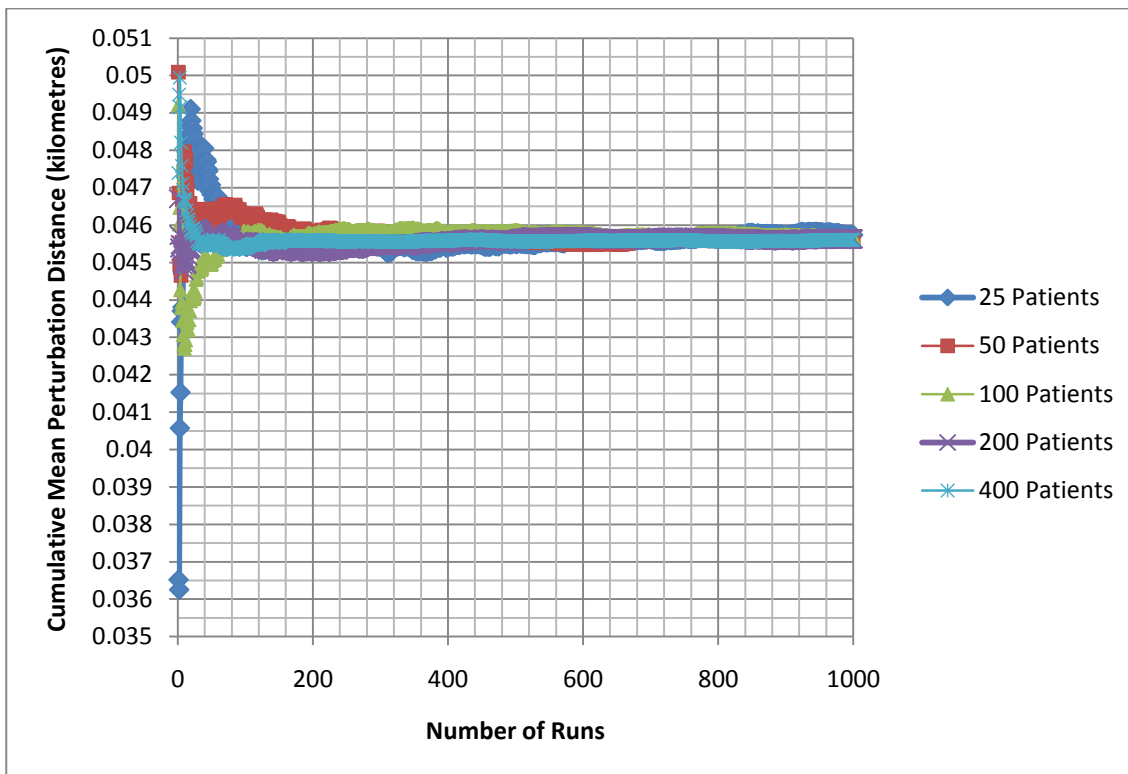
representing 0.0005% of the tested unique individuals. In the case of the 50-patient dataset this was an 83-year old male; for the 100-patient dataset an 88-year old male; and for the 400-patient dataset a 15-year old female.

The mean and median perturbation distances (46 metres and 39 metres respectively), as well as the mean and median perturbation radii (70 metres and 60 to 61 metres respectively), were consistent irrespective of the patient dataset size (Table 10). Cumulative means of the distance between the original and the transformed points are presented in Figure 14 for successive iterations, showing a plateau within 1 metre after less than 200 iterations.

The actual  $k$ -anonymity achieved across all runs averaged 5 individuals, matching on sex and age within the defined perturbation radius as prescribed by the pre-defined  $k$  requirement. Ignoring age and sex, the average number of individuals within the perturbation radius from which the random selection was made was just over 800 individuals. The change in the actual  $k$ -anonymity range was also just as dramatic when taking all dimensions into consideration versus only location, with a median of 6 in the former compared to almost 600 in the latter.

The overall spatial mean of the transformed points was within 5.6 metres of the original spatial mean across all runs, and was inversely related to the patient dataset size (Table 10).

**Figure 14: Mean cumulative perturbation distance for successive runs of the tested patient dataset sizes**



**Table 10: Results of the Multidimensional Point Transform algorithm with different patient dataset sizes for New York County**

SETTINGS	PATIENT DATASET SIZE ( $n$ )				
	25	50	100	200	400
Total records	25,000	50,000	100,000	200,000	400,000
$k$ -anonymity setting	5	5	5	5	5
Maximum $\Delta$ setting (kms)	1	1	1	1	1
Unique Individuals	24,569	49,191	96,056	187,101	348,947

**Table 10: Results of the Multidimensional Point Transform algorithm with different patient dataset sizes for New York County (continued)**

MEASURES	PATIENT DATASET SIZE ( <i>n</i> )				
	25	50	100	200	400
Age-Perturbed records	0	1	1	0	1
<b>Perturbation Distance (kms)</b>					
Mean	0.046	0.046	0.046	0.046	0.046
Standard Deviation	0.035	0.034	0.035	0.034	0.034
Minimum	0	0	0	0	0
Median	0.039	0.039	0.039	0.039	0.039
Maximum	0.975	0.979	0.914	0.980	0.992
<b>Perturbation Radius (kms)</b>					
Mean	0.070	0.070	0.070	0.070	0.070
Standard Deviation	0.042	0.042	0.042	0.042	0.042
Minimum	0.004	0.003	0.002	0.003	0.002
Median	0.060	0.060	0.061	0.061	0.061
Maximum	0.994	0.997	0.983	1.000	1.005

**Table 10: Results of the Multidimensional Point Transform algorithm with different patient dataset sizes for New York County (continued)**

	PATIENT DATASET SIZE ( $n$ )				
	25	50	100	200	400
<b>Actual <math>k</math>-anonymity level</b>					
Mean (rounded down)	5	5	5	5	5
Standard Deviation	1	1	1	1	1
Minimum	5	5	5	5	5
Median	6	6	6	6	6
Maximum	17	15	66	17	16
<b>Location Only <math>k</math>-anonymity</b>					
Mean (rounded down)	809	818	818	817	820
Standard Deviation	909	918	976	931	933
Minimum	19	19	15	19	7
Median	587	590	589	593	591
Maximum	38,122	43,782	68,449	66,100	58,965
<b>Spatial Mean Displacement (kms)</b>					
Mean	0.010	0.007	0.005	0.003	0.003
Standard Deviation	0.006	0.004	0.003	0.002	0.001
Minimum	0.0004	0.0002	0.0001	0.0002	0.0001
Median	0.009	0.007	0.005	0.003	0.002
Maximum	0.057	0.027	0.016	0.011	0.007

## 8.5. DISCUSSION

This study describes a multidimensional point transform (MPT) for anonymising data for public health use that includes location perturbation as a core component of the overall anonymisation algorithm. This novel approach provides the flexibility to allow dynamic and interdependent transformations on health data; by perturbing the location dimension in concert with other user-defined dimensions, the MPT offers a more comprehensive and valid anonymised dataset than existing proposed random perturbation transforms.

A preliminary test of the algorithm was completed to observe its functionality on a synthesised New York County population. Three dimensions were included in the transform: sex, age and location. The perturbation tolerance for the sex dimension was set to 0, indicating a perfect match requirement; the age dimension perturbation tolerance was set to a maximum of 5, allowing for the age to be perturbed by up to 4 years; and the geographical distance perturbation tolerance was set to a great circle distance of 1 Km, thereby allowing a maximum spatial displacement of up to 1 Km. The overarching anonymity threshold was set to 20% (i.e. within the one kilometre radius of any given case, at least 4 other records are indistinguishable from the case on the dimensions controlled for). Patient dataset sizes of 25, 50, 100, 200 and 400 patients were tested with 1,000 iterations each.

The quality of the transformed data was found to be high on the three transformed dimensions: since the sex perturbation tolerance was 0, there is no loss of information on the sex dimension; the age dimension was unaltered for the vast majority of runs (99.9995%); and the mean spatial displacement of the transformed records was 46 metres from the original patient point, irrespective of patient dataset size. Fifty percent of points being displaced across all runs were moved less than 40 metres. In addition, all transformed records were spatially accurate in that their locations represented actual household locations within the synthesised population.

Since New York County has a very high population density, it is not surprising that a  $k=5$  anonymity level was achieved within such a small distance and with no impact on the age dimension. To further illustrate the multi-dimensional aspect of the algorithm, *ad hoc* analyses were completed using 100 iterations of sample sizes of 25 cases with  $k=20$  and  $\Delta=50\text{m}$ , thereby testing a total of 2,500 records. With these tightened constraints, only 19 of the total records were transformed with no age-categorisation required; 148 were transformed with a level 2 age categorisation (i.e. a 1-year interval; recall level 1 is the exact age with no categorisation); 274 were transformed with a level 3 age categorisation; 292 were transformed with a level 4 age categorisation; and 290 were transformed with a level 5 age categorisation. The total successfully transformed records of the 2,500 was thus 1,023 (41%) with a mean perturbation distance of 31 metres. The remaining 1,477 records could not be transformed within 50 metres of the original point at a  $k=20$  level, and all 1,477 reached level 5 age-categorisation as expected. If this were a real scenario, the user at this point, in collaboration with the data custodian, would have to decide on the best approach; for example, from the user's end, additional levels could be added to the age categorisations, gender perturbation could be allowed or maximum perturbation distance could be increased. From the custodian end,  $k$  could be decreased, for example conditional on the user meeting specific security and data handling requirements as evaluated within an appropriate framework (see Part III). Note that, as described, increasing the distance threshold does not translate to a non-optimal distance transform because the minimum distance required to achieve the desired  $k$  is used for each record independently.

Previously described algorithms treat other record attributes separately from location, sometimes with a weighting to account for the generalised underlying population demographics. For example, Armstrong *et al.* assume that all other potentially identifying attributes such as health information, age, sex, and so on are sufficiently “non-individual specific” [16]. Even in previous studies where the random perturbation transforms have been carried out within the context of the underlying population, they

have not done so in parallel with other dimensions, but rather by applying generic weights. For example, Kwan *et al.* [112] acknowledge the arbitrary nature of their weight factors, and the fact that only one run was performed for each perturbation and analysis. They also state that the results are specific to the unique and particular combination of their underlying population and the case-data. Similarly, Cassa *et al.* [143] implemented a probabilistic Gaussian-skewed random perturbation transform based on the underlying age-based population density. This consideration allowed case points lying in less densely populated areas (based on census block groups) to be probabilistically perturbed over larger distances to meet required  $k$ -anonymity thresholds. As with the Kwan *et al.* study [112], this was done using generalised age and age-based population density weights, which the authors refer to as “multipliers”. Also as with previous studies,  $k$ -anonymity calculations were probabilistically based, without taking into consideration other individual dimensions (e.g. age and sex). In other words, a male individual aged  $x$  years who is moved within a certain distance that captures 1,000 other people can be considered to have a  $k$ -anonymity level of 1,000, even if he was the sole male aged  $x$  years within that area. One could reasonably argue that, given the general importance of and requirement for retention of these dimensions in public health practice and in light of the actual  $k$ -anonymity findings of this study, they must be included in any assessment of the anonymisation level of an algorithm. The other issue that this potentially raises is that, without including an adjustment for attribute-stratified population, everyone gets treated in the same way such that anonymisation is based on the underlying overall population density and not in concert with the individual's relevant attributes. Consider, for example, two cases in the same rural area of 200 people, one 35 years old and the other 92 years old. Using a non-stratified population-specific algorithm with a  $k$ -anonymity requirement of  $k=20$ , for example, would anonymise both within the rural area. However, while it is possible - and possibly probable - that there are at least twenty 35-year olds in the area, it is highly unlikely that there are at least twenty 92-year olds. The MPT algorithm described in this study resolves these issues.

Building on geographical masks, Wieland *et al.* [144] used linear programming to perturb patient points given a maximum re-identification probability such that the perturbation distance is minimised in New York county. As with this approach, the MPT algorithm seeks to minimise perturbation distance and relies on an existing spatial population to do so. However, it accomplishes this while simultaneously transforming on multiple dimensions (in this case, sex, age and location), thereby further minimizing information loss. This has a demonstrably significant impact on the implications of the transform. For example, Wieland *et al.* [144] suggest that aggregating to zip code in New York county yields a  $k$  value of about 884 and a corresponding perturbation distance of 519 metres; the LP method they describe has an associated perturbation distance of only 3.3 metres for the same  $k$ , although it is based on moving patients grouped by census blocks. As shown in the results for this transform, a similar  $k$ -value acting on non-grouped individual points requires a perturbation of, on average, 46 metres. However the **actual**  $k$  when taking the other dimensions into consideration drops dramatically to only 5.

More recently, Hampton *et al.* [145] published on a "donut method" of geomasking which simply added a minimum perturbation distance to existing context-sensitive approaches [143]. The authors argue that allowing a case to be perturbed to its original location presents a re-identification risk, since an intruder will know that a few individuals may still be correctly identified. The study also prevents cases from being perturbed outside of their original administrative boundaries, arguing that demographic characteristics of the boundaries are often significant for public health purposes, though also dependent on the "research environment". This is generally true: the relevance of the demographics depends on the scale of the area of study and the study's intent. The authors also suggest that their approach is adaptive because it not only adjusts for the underlying population density, but also for minimum and maximum  $k$ -anonymity - whereas other random perturbation methods are only "semiadaptive" because they fail to be bound by a minimum anonymity constraint. However, one could



just as easily argue that the donut method as described is also only semi-adaptive because it fails to consider the details of the population demographics, such as age and sex. Given the aggregate use of underlying population density, the authors' suggestion that other dimensions such as age and sex could be incorporated into the algorithm would have to rely on weighting mechanisms similar to those previously described [112,143]. The donut-algorithm also does not address the possibility of randomly generating a point in a residentially-improbable or impossible location, such as a river or park, and is subject to a re-identification risk associated with multiple iterations [150,167]. In contrast, the MPT resolves many of these issues, and can easily exclude original patient locations, incorporate a “donut-like” effect if desired, and retain points within defined geographies.

#### **8.5.1. RE-IDENTIFICATION RISK**

One of the most important noted issues with random perturbation algorithms thus far is the fact that repeated iterations on the same dataset increase the likelihood of re-identification [150,167]. The MPT in the current study reduces this risk in two ways: (1) by using the actual underlying population distribution and (2) by incorporating the random small distance,  $\delta_r$ , to create the maximum selection area. The use of the actual underlying population distribution and its corresponding geographic points, as opposed to randomly placed points, avoids inaccurate or unrealistic placement of a point and can add a skew to the point pattern. In other words, repeated random point generation within a circle of radius  $r$  will have a uniform distribution, the mean of which will approximate the original patient point. However, repeated random selection from pre-defined points within a changing radius will have varying distributions based on the locations of those points, which may be skewed. The addition of the random and variable perturbation radius  $\delta_r$  also has an added stochastic factor that potentially creates a slightly different selection pool with every run. Therefore, the mean location of repeated iterations will depend on the spatial distribution of the underlying population

and the variability of  $\delta_r$ , and will not necessarily approximate the original location unless the population is uniformly distributed around the patient.

A possible re-identification weakness of the MPT lies in the prevention of selection of one case as a transform of another. For example, given two cases of identical age and sex within  $R$  of one another, it could be discovered that each is excluded from the transformed options for the other in favour of a more distant point, allowing potential re-identification of both original points. Removing the preventative selection criterion can resolve this, though it may also allow re-identification since repeated iterations will result in case location selection twice as many times as others. Yet another potential for re-identification exists if the parameter settings for perturbation tolerances and thresholds used are known, though this would require extensive time and computing power since an intruder would have to re-create all possible selections. The multidimensional nature of the transform helps complicate re-identification efforts – the more dimensions are permitted to be perturbed, and/or the greater the perturbation allowance, the more difficult re-identification becomes – while exact dimension matching is mitigated by the user-defined anonymity threshold.

There are a number of key issues that make a comparison between the current algorithm and standard  $k$ -anonymity methods difficult, most notably the "reference" dataset used to anonymise the records. In many cases of standard  $k$ -anonymity methods, records are anonymised *within* the dataset with methods proposed to optimise the anonymisation process [168]. Therefore, the case data are their own "reference", such that, once completed, cases are indistinguishable from one another. In the MPT, data are anonymised *across* datasets; the reference dataset in this case, is the underlying population, such that cases are indistinguishable within the population. This also means that the two approaches essentially deal with different types of re-identification risk.

Another important difference is that the MPT was developed with a specific focus on the inclusion of granular spatial information. This has profoundly significant implications for the issue raised above. In the case of many standard  $k$ -anonymity algorithms, exact residential coordinates are unusable since they are essentially direct identifiers - unless they represent apartment buildings. The standard approach would be to either strip them or aggregate them, for example to postal or zip code. Once aggregated, use of  $k$ -anonymity algorithms is possible. However, this introduces two new issues: (1) the loss of granularity, context and analytical accuracy inherent to aggregation, as illustrated in the scientific literature and referenced in previous sections, and (2) potentially increased risk of re-identification. To expand on the second issue, let us assume that the points have been aggregated to the postal or zip code level, and a  $k$ -anonymity algorithm such as the Optimal Lattice Anonymisation algorithm [168] has been implemented. If the value of  $k$  is larger than the population with the anonymised values of the smallest postal or zip code area, then the risk of re-identification is necessarily increased. For example, if 5 records are indistinguishable from one another in area  $A$  as males aged 10-15, then it is true that we have successfully achieved a  $k$ -anonymity of 5 within the case dataset. However, if there are only 5 males aged 10-15 years in area  $A$  anyway, then the actual risk of re-identification is 100% - it does not matter that I cannot tell which case is which, the fact is I now know they are all cases and can re-identify all of them as such. Similarly if there are 10 such individuals in area  $A$ , my effective population-based  $k$  value is only 2 - a 50% probability of correct re-identification. Therefore, in order for this to effectively work, the underlying areal population matching on the variables or dimensions being anonymised on must be at least  $k^2$ .

The MPT, on the other hand, was specifically developed to allow anonymisation on the most granular level of spatial data, though as described, it can also work on aggregated spatial data if that is all that is available. Since it allows for anonymisation on multiple dimensions as well beyond just location, it is suggested as an improvement

over existing geomasking algorithms where spatial information is important (and the case is made that such information is becoming increasingly relevant to and critical for public health practice). Because it anonymises based on the underlying population, it avoids the second issue raised above. However, it should be noted that as described, it does not anonymise *within* the case data itself. Therefore, re-identification may be possible if an intruder knows that an individual is in the dataset and knows the individual's attributes. So, for example, there may be 100 females in the population aged 48 years within the acceptable distance threshold of a 48 year-old female case, but if she is the only 48-year old case and an intruder knows that Ms. Jane Doe is 48 years old and is in this case dataset, then her record will be easily identified and any additional data will be revealed. Hence the different re-identification risks addressed by the different approaches.

Because the reference datasets are necessarily different in the two scenarios described, performance metrics are therefore also different. With algorithms anonymising a case dataset relative to its own records, performance will be related to the size of the case dataset since that is the only dataset of consequence. However, with the MPT, performance is a multiplicative function of the size of *all* datasets used to anonymise; at minimum with granular data such as those used in the current methodology, this will be the product of the number of individuals in the case data and the number of individuals in the underlying population who are not cases. The MPT can be configured to anonymise a dataset relative to itself if so desired as noted briefly below under Strengths. This approach, however, was not tested and is not the focus of the current study since in the context of location privacy it only works with generalised or aggregated locations.

One of the previously mentioned issues of re-identification is particularly important in public health practice and can be rephrased in common public health terminology: meaningful privacy preservation is a function of prevalence or incidence. To further

stress this important point, consider 50 HIV cases in a population of 1,000. Random selection of any one individual has a 5% chance of correctly identifying an HIV case. If all information were stripped for the 50 cases except for their location, each individual would be otherwise unidentifiable from 999 other individuals (i.e.  $k=1,000$ ). However, while this has achieved the maximum possible  $k$  for this population, it still remains that correctly identifying an individual as having HIV has an effective  $k$ -value of only 20 because the prevalence is known.

### 8.5.2. LIMITATIONS

The MPT relies on the presence of an underlying base population containing the same dimensions as those required by the data-user. In the current implementation, for every age-sex combination in the patient dataset, there must be at least  $k-1$  other identical age-sex combination individuals in the non-patient dataset. As the number of dimensions controlled for increases, the number of available matches decreases, thereby potentially necessitating dimensional compromises which can be controlled by increasing the allowable perturbation of the individual dimensions. The MPT allows exploration of the optimum context-specific combinations for appropriate data release and use.

The current algorithm also makes use of a synthesised population. Its validity, therefore, depends on how well the synthesised data mirror reality *on all dimensions of importance*, if a synthesised base population is used. This is discussed in further detail below.

Some issues were encountered that impacted overall performance, including periodic file locks, competing background applications, power outages, and system resources. Although using RAM allowed faster completion, future implementations may be limited by the amount available for allocation and machine specifications. Other performance-enhancing factors may include use of solid state drives, multithreading and multiple

processors, and coding and implementation within an environment other than SAS, which is input/output intensive (i.e. it frequently reads and writes data, which affects performance). Performance will also be a function of the underlying population matrix size; the greater the number of dimensions and the larger the population, the longer the algorithm will take and the more resources it will require.

No amount of masking, de-identification or anonymisation can prevent the misuse of data. For example, information on patients with HIV can be misused irrespective of the scale at which it is released, though the ramifications may vary with that scale. So, in the extreme example where the algorithm is used to perform a multi-dimensional transform with a stringent  $k$ -anonymity requirement that perturbs points across the entire extent of the base population, as in the case of the HIV example, it will still show the prevalence of HIV in that base population. The release of data must therefore always be based on other important considerations, such as the level of trust in the user, the purpose and scientific or applied merit for which the data will be used, implemented security measures and so on. Such considerations form a framework from within which data-sharing decisions can be made and this is the focus of Part III of this study.

### **8.5.3. STRENGTHS**

Strengths of the MPT algorithm are its powerful flexibility and customisability, easily allowing criteria to be set on appropriate dimensions relevant to both the study and the target population as opposed to irrelevant or arbitrary political and administrative boundaries. For example, in the current implementation, every individual is associated with five age classes. Depending on the user requesting the information and the intended use of the data (both important aspects to consider within a data-sharing framework), a minimum age class can be set. In this way, information deemed more sensitive might only be released if age is categorised within the appropriate classes. If the base population file uses only age classes based on census information then it can

still be used. This would also allow different ages to be classed differently based on the population distribution within the region of interest. For example, given a population where the majority is between the ages of 18 and 65 years, one may choose to create age categories from birth to 18 years old, as well as over age 65 years, but keep exact ages intact for those between the ages of 19 and 64 years inclusive (given their higher numbers). One method of achieving this can also be to apply the transform on the base population using itself as the reference, on the age dimension only. This will ensure that within the defined threshold for anonymity and distance, the age dimension is categorised appropriately without transforming the location information. This new, transformed population can then be used as the base population for patient-list transforms.

Another advantage of the MPT is its use of a granular base population which can be assigned to increasingly coarser geography. For example, given only zip codes or postcodes for the patient dataset, points can still be approximated using the base population and other dimensions provided in the patient record and any given record would still have to meet the required anonymity threshold on all the dimensions within the underlying granular population. In such cases, a match on the dimensions within the zip code would be transformed according to the algorithm, using all the individual points around it in the population - the age and gender issue would be accounted for because the individual points are still being used to assess anonymity level, and the algorithm would not be subject to the modifiable areal unit problem as previously described. If the distance threshold used remains within the area defined by the zip code then the MPT will modify age and gender according to the specifications given it in order to meet the required  $k$ -anonymity. Therefore, a 32 year old female patient in postcode X1X1X1 can be assigned to any point within that postcode matching on age and sex within the allowable perturbations as part of the MPT algorithm. This will introduce a maximum error approximated by the sum of the maximum distance between residential points within the postcode area and the perturbation radius  $R$ . In

spite of this error, this would still be a better approach than current perturbation methods or use of the area's centroid since it still adjusts for the desired attributes to achieve the appropriate anonymity requirement and uses real locations to prevent unrealistic or impossible placement. Since the patient dataset is at a coarser geographic scale, the method would also further confound potential re-identification, with each run yielding different results based on a different point of origin.

The MPT also does not apply blanket rules to the entire patient dataset; rather, it anonymises each record individually for its own optimum transform. The merits of such a local transform as opposed to a global one are context specific; one can easily argue that, from a practical aspect, as long as the recoding is done consistently, local recoding is actually better since it allows one to satisfy the anonymisation requirements while releasing maximum data without being impacted by a few extremes or outliers - leaving it up to the user to decide whether or not to then apply some or all of the recoding more globally if informed of the recoding hierarchy. There is debate as to the better approach (in terms of local versus global recoding), but in the current case, the local recoding allows release of the data with the best possible configuration. It also allows for more sophisticated integration of contextual information, facilitating comparison and calculation *across* datasets. For example, given a dataset containing the locations of schools in the extent of the base population, the algorithm can easily include distance to the nearest school as an added dimension. So if, for example, one wanted to apply the algorithm to include preservation of the relative spatial distance to schools, a maximum spatial displacement of each record from its closest school can be specified in combination with the other relevant dimensions. The incredible power and flexibility this affords can also be reflected by implementing dimension-specific rules. For example, the algorithm can be modified to apply different rules to different records based on the values of one or more dimensions; if the case is a male, for example, allow this level of perturbation to these dimensions, if female then allow a different level to the same or different dimensions.



As has been mentioned, the algorithm is also not independent of the underlying geography. Because it uses pre-established locations for the random selection to meet the required anonymity threshold, knowledge of the existence of non-inhabitable regions or features will not increase re-identification potential (a noted issue with some of the current random perturbation techniques [16,144]). These factors allow the spatial aspect of the transform itself to be bound by multiple, contextually appropriate rules defined by the user.

If specific dimensions are not known *a priori*, such as education and income, an areal dimension can be added as part of the control to allow retention of the patient within the specified political or administrative boundary if desired. For example in the case of census boundaries, including a requirement for an exact match on the boundaries' codes will ensure that the case remains within that census area or flag it as not-anonymisable if the requirement cannot be reached with the perturbation rules. This allows the flexibility to use as little or as much data as are available to achieve optimum results. The advantage to including additional dimensions beyond administrative or political boundaries is the incorporation of actual contextual variables as opposed to potentially artificially-related areal units.

As mentioned, New York county is extremely population dense making it relatively easy to achieve reasonable anonymity with very little spatial displacement, even when multiple dimensions are considered. As the *ad hoc* analyses show, however, the MPT allows users and custodians to identify this and modify the parameters in order to achieve acceptable results. In this case, for example, the custodian may agree to lower the *k*-value if acceptable or pending certain requirements on the part of the user (e.g. use restrictions, security requirements, etc.). Conversely or simultaneously, the user may accept additional perturbations (i.e. of sex or age) or increased spatial

perturbation. The same decisions would have to be made for a sparsely populated rural area; either way, population density does not impair the MPT. By allowing the user and custodian to have complete control over the various aspects of the transform, including the appropriate or acceptable anonymity threshold, the MPT provides a powerful, flexible and truly adaptive “user-sufficient mask” [16] which minimises divergence from the original dataset on a case-by-case contextually sensitive basis.

#### **8.5.4. USING SYNTHESISED POPULATIONS**

Health data are most valuable and informative in their most granular form, and developing a transform that works on individual point-level data at the address level is highly beneficial. However, such a transform would require knowledge of the underlying population – also at the individual point-address level. Although available through population registries, these data are themselves subject to privacy and confidentiality restraints, and are therefore generally not accessible for public health use. Instead, public health practitioners rely on aggregated census data to infer various population demographics. This is where synthesised populations may play a role. For example, a “synthesised, geospatially explicit” US population based on the year 2000 census has been generated to facilitate agent-based infectious disease modelling for the Modelling of Infectious Disease Agents Study (MIDAS) [169]. This population “correctly and appropriately” describes the age and sex demographics by household, and accurately reflects the actual US population. Details on the methodology and population characteristics have been published [169].

As previously stated, since the MPT makes use of a synthesised population, its validity depends on how well the synthesised data mirror reality on the dimensions of interest. Since the population is based on the year 2000 census, it may inadequately reflect population demographics for earlier or more recent studies. However, given the recurring nature of the census, algorithms used to build the synthesised populations can be re-run to generate new and relevant populations with each census year [169]. A

synthesised population may also be invaluable in exploring the relationships between perturbation distance and a variety of quasi-identifiers as illustrated through this study. Their use also allows for the creation of realistic, non-circular disease clusters for investigation – an issue that impacts other studies in this field [145].

Synthesised populations for the US and several other countries have been produced for MIDAS and are available by request. These populations were developed for epidemiological modelling, not for de-identification algorithms, further highlighting their general utility in public health. As such, the development of representative synthetic populations would be highly beneficial for public health practice in general. Indeed, development of a synthetic 2010 US population is currently underway by MIDAS scientists, as are tools to allow researchers to generate custom populations based on demographic variables of interest [162].

#### **8.5.5. ALGORITHM REFINEMENT**

Further refinement of the MPT would allow the user to set priority levels for the various dimensions. In the current example, the priority is given to age; age is perturbed only if the anonymisation threshold is not met within the prescribed maximum distance. Instead, the algorithm can be modified to prioritise minimum distance moved within a maximum age perturbation (i.e. the algorithm could begin with the maximum age perturbation to minimise distance and work backwards to achieve the optimum result). This provides maximum flexibility in exploring the optimal transform for a given dataset and context, as minimising changes in one dimension will necessarily impact the effect of other dimensions.

As an example, assume our dimensions of interest are distance, age, sex and race with decreasing priority assignment. In this case, the MPT as illustrated in Figure 12 will first search for  $k-1$  exact matches on age, sex and race. In the absence of meeting this requirement, it will generalise race within the defined generalisation threshold and

look for  $k-1$  exact matches on age, sex and generalised race. Assuming it still fails, it will then generalise sex, and look for  $k-1$  exact matches on age and generalised sex and race. And so on. Based on the current design, it will only move on to generalising the next dimension once it has reached the maximum designated generalisation of the previous dimensions with failure to identify  $k-1$  matches in the population, since the loop is intra-dimensional. The loop can also easily be changed to allow several dimensional generalisations within an iteration - i.e. across dimensions. In this case, and using the same example, race would be transformed to its first generalisable level, followed by sex if required, then age; assuming failure, it would then loop back and generalise race to its next level, etc. The intent is to minimise loss on those dimensions deemed by the user to be more important to retain closer to their original value, as opposed to finding an overall perceived "optimal" solution, while minimising spatial disturbance (i.e. distance perturbed).

The MPT settings can also be informed by other research in this area. For example, the maximum number of combinations (MaxCombs, as previously described) of variables of interest is a good predictor of uniqueness [58,146] and can be used to determine appropriate "geographic area population size" (GAPS; also previously described) [58] for privacy preservation. This can be used to inform preliminary decisions on setting  $k$  and  $\Delta$  for the MPT; for example, one can begin by setting  $\Delta$  to the approximate mean radius of the census geography most closely corresponding to a calculated GAPS cut-off. MaxCombs can also be used to inform the dimensional categorisation levels, particularly since it is dependent solely on the number of response categories and not the types of the quasi-identifiers.

It has been shown that  $k$ -anonymity can, in some cases, be "over-protective", particularly for smaller sampling fractions, resulting in unnecessary information loss [147]. The current methodology helps reduce such information loss by incorporating the relevant dimensions directly into the anonymisation algorithm, allowing the user to set

permissible categorisation and priority levels, and performing local recoding (i.e. allowing observations to have different and overlapping response intervals [147]). Appropriate  $k$  values should be a function of the user and the use of the data, as well as the governance structures in place; for example, a much higher  $k$  would likely be appropriate for publicly-available data or researchers with low security measures in place while lower  $k$  values may be appropriate for studies that have a high impact on population health, or trusted researchers with strong security measures in place. Some general criteria for setting this threshold have been proposed [146] and should be incorporated into a more comprehensive framework for the disclosure of data as described in Part III of this study.

Preliminary MPT testing was conducted on three dimensions: sex, age and location. Additional dimensions, larger patient datasets and different base populations with varying population densities should also be explored, as should the effects on common spatial statistics used in public health. Since random perturbation techniques generally increase Type II error probability (e.g. cluster dilution) and do not affect Type I error probability [16], further studies on appropriate thresholds and applications of this algorithm are required in various contexts and with different base populations. Additional analyses quantifying the relationship between the anonymity level achieved and the distance displaced for specific contexts and base populations can also serve as part of a framework for assessing appropriate uses. Currently, privacy legislation applies to "identifiable individuals". However, with the growing literature around anonymisation, one can now ask the question "at what  $k$ -value does an individual cease to become identifiable under the legislation"? Acceptable anonymity thresholds therefore need to be set and standardised, and the legislation needs to be revised to better reflect this in privacy definitions.

Sophisticated software agents [15,16] could be used to combine the ingredients required (e.g. the base population from a municipal population registry, the health data

from the custodian, and the user requirements) and return an appropriately and optimally transformed dataset (or null result, if no adequate transform is feasible given the data and user specifications). This allows the user to explore analyses that may only become evident after visual exploration of the data's distribution. A graphical user "front" would be highly beneficial for this purpose, and an image of such an interface is suggested and described in Chapter 11.

## 8.6. CONCLUSIONS

The multidimensional point transform proposed in this study works concertedly on multiple attributes, *including the spatial attributes*, to give a more complete and appropriate transform that builds location privacy into the anonymisation model from the beginning. Unlike previous studies, this algorithm does not leave other attributes "untouched", but it does result in a transformed matrix with the same dimensions of the original matrix [16].

The ideal transform preserves the confidential and private nature of individual health records, as well as the geographic integrity of the data, to facilitate public health practice [16]. The optimal approach depends not only on the purpose for the data use and the acceptable disclosure risk [16], but also on the characteristics of the data. Acceptable disclosure risk by the custodian is also a multifaceted consideration based on acceptable anonymity thresholds, trust in the user, adequate security measures, and so on. However, such algorithms cannot substitute for secure and ethical conduct, and a framework for the appropriate disclosure, use and dissemination of data containing personal identifiable information is required [154]. There are also instances in which the release and use of identifiable information in public health are essential [170], and consideration must be made within a developed framework to allow for such cases. The proposed algorithm in this study presents a multidimensional approach that allows one to tweak and optimise the trade-offs for any given dataset and purpose, presenting a necessary component of the much-needed public health framework.

**PART III  
GUIDANCE**





## 9. Towards A Conceptual Framework

*The more important the subject and the closer it cuts to the bone of our hopes and needs, the more we are likely to err in establishing a framework for analysis.*

**Stephen Jay Gould**

### 9.1. INTRODUCTION

As has been discussed and exemplified through the body of literature, various perturbations and transforms have been developed and explored as a means of addressing location-privacy issues in public health and other fields. However, these are not all created equal and their application is context specific. Given a particular scenario, how does a public health professional or custodian decide on whether a transformation should be applied for a given objective? How would a custodian and a data recipient assess the privacy concerns around data release and use? In the context of health research, ethics boards often have discretion over whether or not a custodian can release data to the researcher, but how does the research ethics board assess the appropriateness of data release from a privacy perspective? What factors should be considered in the implementation of a given transform and the use of its outputs? In attempting to address these questions, it quickly became apparent that a framework with a broader and more general approach to data-sharing within the sphere of public health is required.

The purpose of this framework is to provide guidance to the public health community on appropriate data release assessment from a privacy perspective, stemming from the current research on location-privacy. By consolidating a large body of literature and various recommendations across multiple disciplines, this workable framework presents five interrelated domains, each with various dimensions, to describe and evaluate the risks associated with data-sharing and its anonymisation in public health practice. It is important to note that the application of the framework would benefit greatly from the development and adoption of universally accepted standards,

definitions and structures, many of which are currently lacking for many of the components. This does not detract from the framework itself, which can be used as a guide to help inform the development of such standards, but only from its universal and consistent application. Like any tool, what one gets out of it depends on what one puts into it.

The motivation for this framework initially came from the required guidance to the public health community on the appropriate disclosure of location-sensitive information as documented in the literature [70,154]. However, as mentioned, the requirement for a broader approach was quickly recognised. The development of such a new approach is also supported by the Committee on Health Research and the Privacy of Health Information in the United States as evidenced in their 2009 book published by the National Academy of Science [68]. As such, while the overall framework and methods used themselves are generic, the results and their exact applications will be context-specific, based on scenario-defined parameters. In other words, the framework presented herein forms an operational base or foundation from which specific public health scenarios and contexts can then be extruded. Its application, therefore, will vary from scenario to scenario, and will be largely defined by the user's settings of various parameters relating to relevant aspects of the scenario. In this way, it lends itself to a dynamic model that allows scoring and weighting to be iteratively modified to achieve a "best fit" scenario, providing for a powerful, flexible and adaptable structure for decision making.

## **9.2. FRAMEWORK OVERVIEW**

In their recommendations [68], the Committee on Health Research and the Privacy of Health Information in the United States endorsed a goal-oriented approach to developing guidelines that enabled appropriate decision-making as opposed to prescriptive regulations; instead of defining permissible and rigid activities which may

not apply in all situations, the focus must be on providing adaptable and enabling governance structures to facilitate required public health activities. It is in this spirit that this framework has been developed, based on five core *domains* which in turn are each composed of four *dimensions*. Guidance for the assessment of each dimension can be augmented through a series of proposed measures, which could be translated into a "Yes/No" checklist for ease of use and reduction of bias. The overall structure of the framework is illustrated in Figure 15 and the flow is provided in Figure 16. The collection of domains, dimensions and criteria is based on what has been reported in the public health literature, as well as the ethical and legal literature surrounding privacy and general data use and disclosure, a number of surveys and best practice guidelines, and the author's personal experiences and interactions within the public health communities in Canada, the United States and the United Kingdom, all of which are referenced throughout this study. Not only does the implementation of such a framework provide invaluable guidance on aspects to consider in the data-sharing decision-making process and how to assess them, but in doing so it also allows for an evidence-base for the decision making process. This allows users of the framework to concisely and clearly identify the justification for either sharing the data or not, the requirements and conditions for data-sharing, and issues that must be addressed in order to facilitate the process if required.

The domains of the framework address the **recipient** of the data, the **data** being requested, the **purpose** for which the data are being requested or shared, the implications of any **transformations** performed on the data, and the intended **output** or eventual dissemination of the data beyond the recipient. Note that the **output** domain is a super-domain assessed on the other four domains, but contributing to the overall assessment for data release. A summary of these domains and their corresponding dimensions is given in Table 11.

It is important to note that a **subject** domain (referring to the individuals to whom the data pertain) has not been included since the framework is based on the absence of individual control-based principles as previously discussed. The framework provides guidance on the non-mandatory release of identifiable or de-identified data for purposes other than those for which it was originally collected (as previously mentioned, this is sometimes referred to as secondary use) for public health practice.

**Figure 15: Structure of the proposed public health data-sharing framework showing domains and their corresponding dimensions**

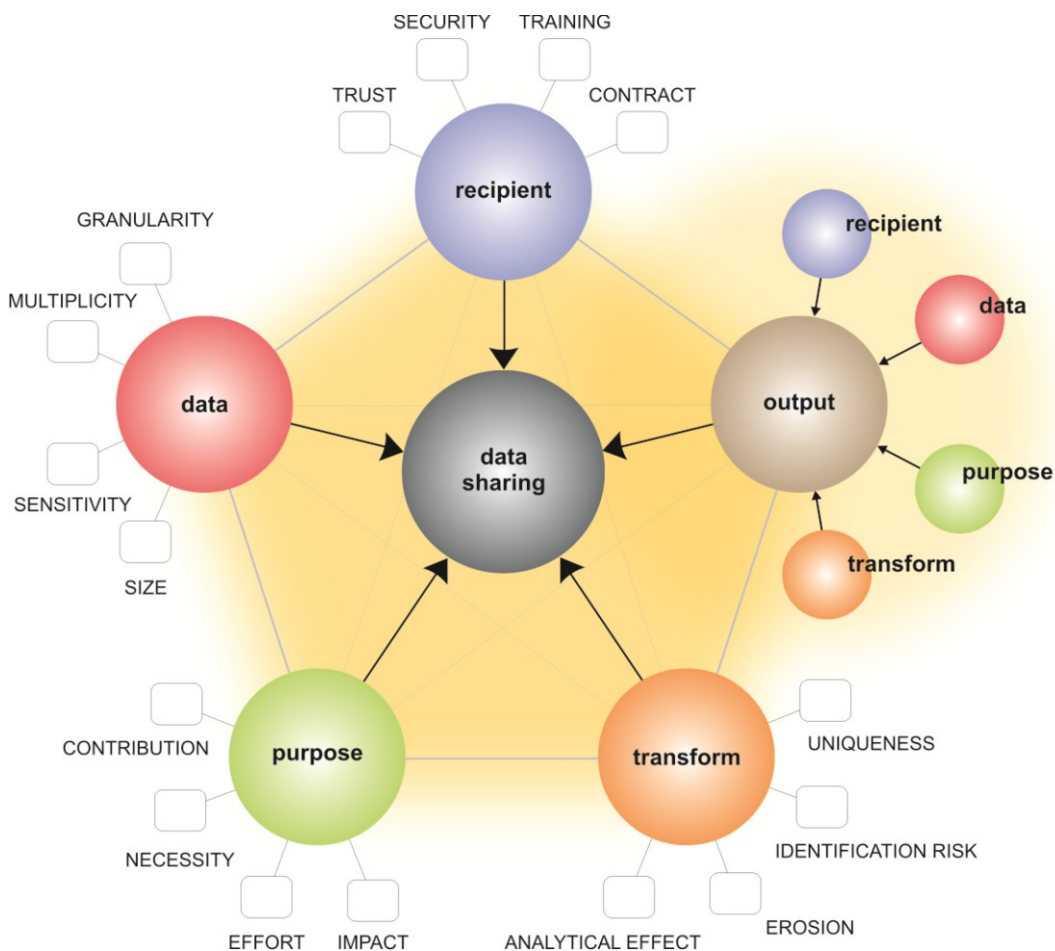
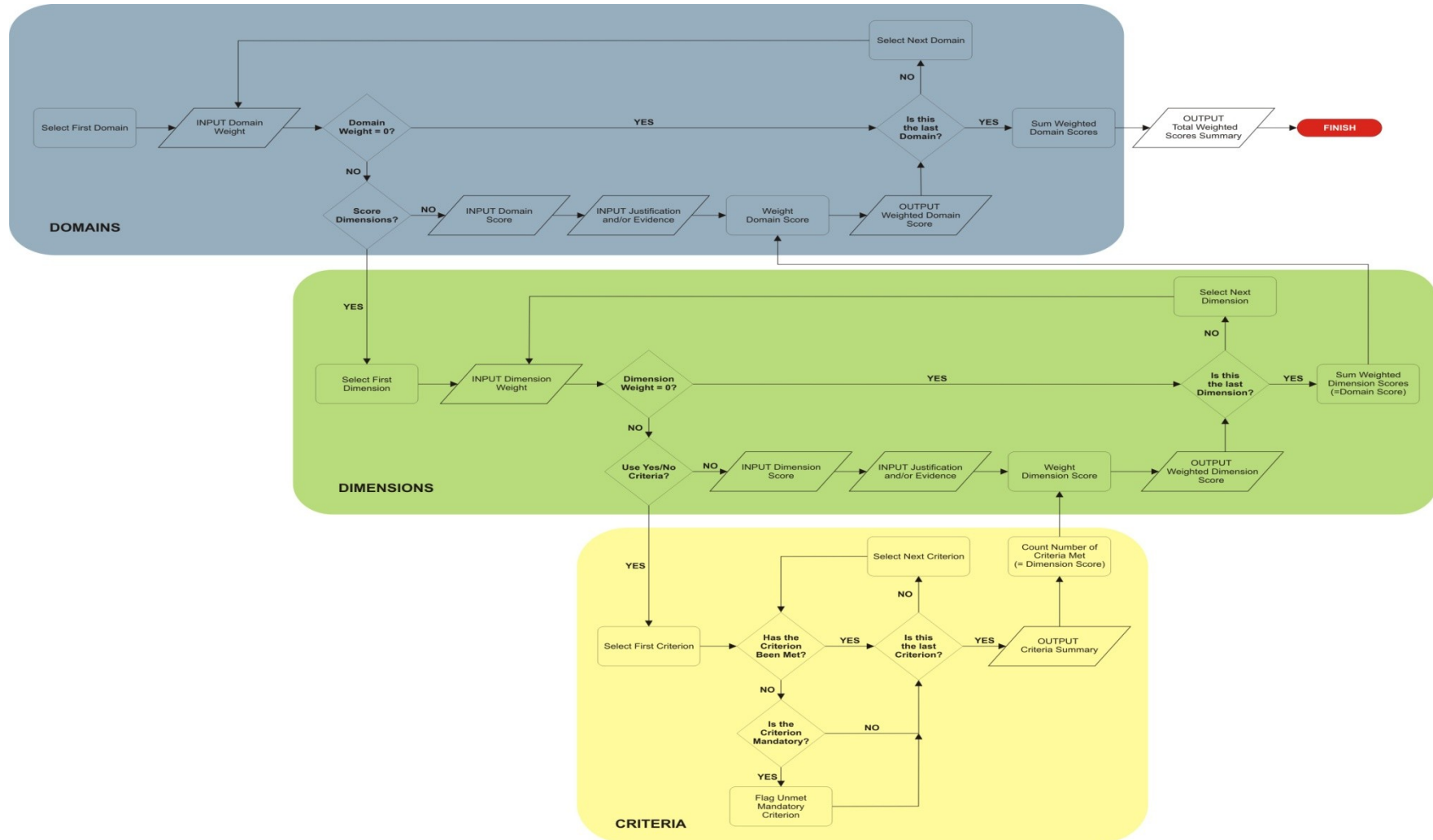


Figure 16: Flow of the proposed public health data-sharing framework



**Table 11: Outline of the proposed framework domains and their corresponding dimensions**

Domain	Dimension	Description
<b>RECIPIENT</b>		
	TRUST	The degree to which the provider of the data is confident that the recipient will handle them in an agreed-upon appropriate, ethical and professional manner.
	SECURITY	The policies and safeguards enforced and implemented by the recipient to ensure the secure handling of the data and to minimise and mitigate potential breaches
	TRAINING	The recipient's level of awareness and training related to data-sharing and use, personal data, privacy, confidentiality and security.
	CONTRACT	The implementation of an enforceable agreement between the data provider or custodian and the data recipient.
<b>DATA</b>		
	GRANULARITY	The specificity of a each variable
	MULTIPLICITY	The number of variables in the dataset and/or recurring records for a given individual
	SENSITIVITY	Measured by the impact of public access to the data on the individual or communities
	SIZE	The size of the dataset relative to the size of the population at risk or population of interest represented by the data

**Table 11: Outline of the proposed framework domains and their corresponding dimensions (continued)**

Domain	Dimension	Description
<b>PURPOSE</b>		
	CONTRIBUTION	The actual or potential relevance and application of the purpose for which the data are requested to the goals of public health practice
	NECESSITY	The relevance of and requirement for each requested data element to achieve the purpose
	EFFORT	The effort required to share the data including consideration of the appropriate legislation of the relevant jurisdictions
	IMPACT	The potential costs and benefits associated with data-sharing
<b>TRANSFORM</b>		
	UNIQUENESS	The degree to which an individual in a given population is unique on a combination of attributes
	IDENTIFICATION RISK	The ease of identifying an individual, either from the data alone or in combination with other information
	EROSION	The loss of Information within the dataset (for example loss of complete values and changes to data values) as well as in relation to other contextual or environmental features
	ANALYTICAL EFFECT	The quality and validity of the knowledge gleaned from the data through analysis

**Table 11: Outline of the proposed framework domains and their corresponding dimensions (continued)**

Domain	Dimension	Description
<b>OUTPUT</b>		
	"Super-Domain"	Re-assess on previous four domains in the context of the required output

The framework does not provide measures for situations in which release of information is mandatory or legally required. It should be reiterated that in general, legislation allows the release of identifiable information in the interests of public safety, or for research and/or statistical purposes, in the absence of consent [154]. Also note that once information and data have been de-identified, they are no longer considered personal information, and therefore neither fall under privacy legislation nor carry any legislated requirement for consent.

Descriptions of the framework domains, dimensions and suggested criteria are given in the remainder of this section along with some proposed measures for the implementation of the framework.

### 9.3. THE RECIPIENT

The *recipient* domain refers to the individual(s) receiving the data but more specifically and importantly, the individual(s) who will be *responsible* for the data. In some cases, this may be the custodian of the data (for example when being given custodianship), though in most cases it will likely be an individual or group of individuals (e.g. researchers, epidemiologists, etc.) requesting data from a custodian for a specific purpose. This can also include “agents of the custodian”, who use the data on behalf of the custodian for secondary purposes [146]. Custodians can still use this framework to assess the adequacy of their own environment as well.



In scenarios where there are multiple recipients, it is up to the framework user to decide whether to assess each recipient individually, one recipient as a representative of the others (and therefore on whom the responsibility for the data lies as far as that relationship is concerned), the “weakest link” on each dimension, or the group collectively as an entity or organisation:

1. Each recipient individually: while this is the most accurate method, it is also the most time-consuming. However, it can also allow each individual to be assessed once and given a domain-specific identity-based score, which can then be used for that specific individual in future contexts.
2. One representative recipient: in most cases, this will likely be an individual- and role-based decision – for example, where the chosen representative is the principal investigator on a research project, or the designated data custodian, etc. In this case, the recipient’s roles and responsibilities must be clearly defined, and the composite score becomes a context and domain-specific identity- and role-based score.
3. The “weakest link”: the benefit of this approach is that it uses the most conservative measure for each dimension; however, in doing so, it may inappropriately withhold data from a group of recipients. This method therefore uses dimension-specific identity-based scores.
4. The group as an entity: In this case, each dimension is scored based on the qualities attributed to the entire group or organisation, therefore becoming a group-identity-based score.

### **Identification**

The recipient should be clearly identified along with name, role and contact information.

## Dimensions

Assessment of the domain is based on the dimensions of *Trust*, *Security*, *Training* and *Contract* and therefore includes identity- and role-based components. Note that an identity-based measure is simply one that is based on the identity of the individual (e.g. reputation, trustworthiness, history, etc.) whereas a role-based measure is one based on the individual's role (e.g. the principal investigator of a research project, a medical officer of health, a quarantine officer, a primary care-giver, etc.).

### 9.3.1. TRUST

*Trust is the lubrication that makes it possible for organizations to work*

**Warren G. Bennis**  
**Distinguished Professor**  
**University of Southern California**

Trust is a critical component of our daily social and professional relationships and has been differentiated into different types under different paradigms [129,171]. For example, the past decade or so has witnessed incredible growth in the “e-phenomenon”. Be it e-mail, e-commerce, e-banking, e-mapping, or e-health, this phenomenon has brought with it increased convenience, time-savings, and vast amounts of accessible information. Inevitably, this has also resulted in an increase in personal information intrusion. In response, commercial entities have adopted privacy-related standards, identified on their sites through “seals”. It has been suggested that having the privacy policies and governance structures in place – in this case, as represented by these seals – builds consumer trust and therefore facilitates the sharing of information [172]. In this case, it is suggested that the mere presence of privacy policies is an antecedent to trust, which in turn is an antecedent to information-sharing. If only it were that simple!

The issue in the above argument is that in many cases, the trust is based on a false sense of assurance; there is no guarantee that the proclaimed privacy policy will be adhered to, or that any proposed enforcement mechanisms exist; only the word of the

commercial entity. Trust is therefore a more complex dynamic, dependent on a combination of the policies presented to the consumer, as well as the consumer's knowledge of the entity and its reputation (including, if applicable, the requirements of the relevant "seal" program). Indeed, surveys on the public's perceptions of personal information use have shown differences in trust ratings based on the *identity* of the data-recipient, particularly when government institutions are implicated [124,125,128,173]. In primary health care, trust between care-providers is crucial, and often assumed; the Direct Project [174] in the United States, which aims to provide standards and guidance to facilitate the sharing of health information between health providers, is entirely built on the assumption of pre-existing trust between parties [175]. However, trust can be a considerable problem in the context of public health, since much of public health is a governmental mandate. It comes as no surprise therefore that building trust is a critical component for data-sharing in public health practice. Since the notion of privacy is heavily dependent on trust, it is suggested that the presence of a strong privacy governance structure is not by itself sufficient, though it is certainly a requirement as reflected throughout this framework.

Within this framework, *trust* refers to a generalised trust in the recipient to manage and use the data in an ethical, responsible, secure and accountable fashion. Care must be taken to assess this independently of technical capacity to do so, which more appropriately falls under the *security* dimension. The *trust* dimension is therefore purely identity-based. Analogous to an assessment at a job interview, the recipient is assessed on "personal suitability traits" for handling and safeguarding the data being requested. Although this can be somewhat subjective in nature – particularly if the parties involved have no history together or known incidents – several measures are proposed to help assess the *trust* dimension:

MOTIVES: Not to be confused with purpose (addressed under its own domain), motive has more to do with the recipient's underlying drivers and reasons for

requesting the data as opposed to their intended use. Considerations to help assess motive include whether or not the recipient has anything to gain from acquiring the data or lose from not acquiring them, though it should be noted that potential gains and losses do not in and of themselves imply motive. If a potential does exist, it should be addressed directly with the recipient and mitigating conditions be included in a signed agreement between the parties involved (this is captured in the contract dimension). Gains are not limited to the more commonly thought of commercial and financial gains, but also extend to reputation gains, criminal gains, and even malicious “gains” such as the embarrassment or defamation of an individual, group or organisation, including the data custodian. Likewise, examples of losses to the recipient may include commercial, financial or reputational losses.

**REPUTATION:** The reputation of the recipient captures a multitude of relevant assessment criteria. Is the recipient competent? Is the recipient experienced in data use? What is the level of experience? How mature is the recipient in his or her respective field? Does the recipient have a history of and reputation for ethical handling of personal information? What is the recipient’s track record / previous performance?

**CHARACTER:** Perhaps the most difficult to assess, character assessment here does not refer to commonly conducted personality assessments. Rather, it addresses whether or not the recipient has demonstrated integrity, responsibility and good ethical judgment in his or her career and professional and social interactions, and is tightly knit with motives and reputation.

**IMPACT:** Do personal gains from breaching trust outweigh the costs? In other words, is it in the recipient's best interests to safeguard the data and the data provider's trust?

HISTORY: Has the user provided data to this recipient in the past? If so, were there any issues encountered that compromised trust? Has the recipient had any breach incidents in the past? If so, what was the effect, and what was done to prevent similar future occurrences? Have appropriate and corrective or remedial actions been taken / implemented to ensure no recurrence? Was the incident handled transparently and appropriately?

ACCOUNTABILITY: Defined as the "state of being answerable for decisions and actions" [101], accountability is a recurring principle and theme in privacy management and governance models and legislation [60,61,176]. Methods for assessing accountability include whether or not the recipient has data-privacy policies in place to govern the collection, use and disposal of personal information, and if so, whether such policies are regularly reviewed and updated, clear and exhaustive, recent and relevant, and enforceable. If they are enforceable, how is this accomplished? It would be beneficial to have such policies reviewed, approved and endorsed by an independent authority, such as an Office of the Privacy or Information Commissioner. Does the recipient report to another individual? Who ensures that the recipient remains compliant with policies and agreements? In addition to the accountability models and structures the recipient(s) may have within their own organisation, accountability to the data custodian is also important. It would therefore also be beneficial to address whether or not the recipient will report back to the custodian and if so, how often and on what. This is best captured through a signed agreement between the parties involved as described under the *contract* dimension.

TRANSPARENCY: This speaks to the recipient's open disclosure of policies, practices and outcomes to external parties, including the public. In their 2006 report on the attitudes and expectations of Canadians on privacy and the use of their personal information for health research [51], the Canadian Policy

Research Networks reported that the public identified transparency as a critical part of accountability, which in turn could lead to increased trust and therefore increased acceptance of the use of their information. Transparency should be included within the recipient's relevant policies (captured in the *security* dimension) but has been included here since it has been implicated as a significant contributor to the development of trust.

### 9.3.2. SECURITY

*The only truly secure system is one that is powered off, cast in a block of concrete and sealed in a lead-lined room with armed guards – and even then, I have my doubts.*

**Gene Spafford**  
**Professor of Computer Science**  
**Purdue University**

Information security refers to the policies and safeguards put in place to ensure the data's integrity, confidentiality and appropriate management. It is necessitated by examples of loss, theft and inappropriate use, and is thus an ever evolving critical component to the success and survival of any enterprise. Standards organisations such as the International Organization for Standardization (ISO) [177] have developed various security standards (e.g. ISO 27K), forming the foundations for enterprise information security management systems and establishing certification methodology, and guidelines have been published by international and national entities.

Given its personal nature, it comes as no surprise that health information features prominently in these standards and guidelines. In some cases, it is addressed within the broader framework, such as in the supplementary ISO 27799:2008 standards [178], the US Department of Homeland Security Privacy Office guidelines [179], the EU Data Protection Directive [80], and the OECD Information Security Guidelines [180]. In others, it is afforded its own dedicated guidelines by an appropriate body, such as the NHS code of practice for information security management in the UK [181] or addressed within broad health research guidelines such as the Canadian Institutes of

Health Research best practices for protecting privacy in health research [182]. In spite of these inconsistent approaches, one thing is clear: information security is critical to health information management.

In this framework, the *security* dimension is used to refer to the policies and physical and technical safeguards in place to ensure ethical management of the data. This includes its secure storage, controlled access, use and dissemination, and appropriate measures to minimise and mitigate breaches. Suggested measures to aid in the assessment of the recipient's *security* dimension include the following:

**POLICIES:** Does the recipient have a privacy and information security policy in place? Are the recipient's information security policies and practices accredited by a recognised and trusted organisation? Are these policies transparent (tied in to the trust dimension above)? Do the policies provide an accountability framework? The assessment of such policies is exemplified through the standards for information governance against which one is measured for compliance by the NHS in the UK [183].

**STORAGE:** Where and how the data will be stored is a fundamental security issue and addresses both physical and technical storage issues. For example, will they be stored on a locked medium only, such as a USB key or a standalone computer, or a networked system? Consider, for example, the difference between encrypted storage on a non-networked password-protected computer in a locked, restricted access room and a shared drive on a server accessible to all individuals within an organisation. The former may not be necessary, but the physical and technical security differences are clear.

**AUTHENTICATION:** Who will be allowed access to the data, and how will authentication be enforced? Ideally, only recipients should be authenticated

unless they are given explicit, documented and regulated power to delegate or transfer responsibility for the data. Authentication can also be assigned different levels. For example, one level may be simple read access, another might be read and write access, and yet another might include the ability to transfer or transport the data.

**ACCESS:** Once an individual is authenticated, how will he or she be able to access the data? This is related to the storage measures above; for example, access to the facility, access to workstations, access to the data media and so on.

**USE:** Are appropriate measures in place to ensure that the data are used appropriately and within any parameters agreed upon? Are controls in place to ensure the data are used for the specified purpose and in compliance with any terms agreed upon?

**TRANSFER:** Methods of transferring data can present various security issues in their own right. For example, compare the transfer of data over an open wireless network with that over a secured network or actual physical transfer of a storage medium. Included in this are issues around how the data are transmitted or provided to the recipient by the provider in the first place.

**ENFORCEMENT:** Are there established procedures and appropriate resources for enforcing the security measures put in place? Are breach consequences and mitigating responses acceptable?

**DESTRUCTION:** are the means for destroying the data once the purpose has been accomplished acceptable?



### 9.3.3. TRAINING

*Excellence is an art won by training and habituation.*

**Will Durant (summing up some of Aristotle's ideas)**

**The Story of Philosophy: The lives and opinions of the world's greatest philosophers**

The *training* dimension captures professional qualifications and memberships, relevant training and experience. Recommendations for educating professional communities on privacy issues and handling of data abound in the literature and published organisational guidelines [94,179]. The entity providing the data can define how credentials are evaluated based on its mandate and operational requirements. Data custodians may develop privacy training sessions specific to their data holdings and requirements, and request that all applicants undergo these prior to being granted access to the data. However, a better approach would be to have standardised training developed by an office of a privacy or information commissioner where such an office exists (and establishing such an office where it does not).

It is suggested that training be measured as a function of the following:

**AWARENESS:** Recipients should be aware of the right to privacy and its importance to individual identity and autonomy, and should be able to identify what is personal information. Awareness training should also capture the sensitivity of personal data, particularly health data, and the privileged responsibility that comes with access.

**LEGISLATION:** Recipients should be aware of the relevant legislation in their jurisdiction, as well as the data custodian's jurisdiction, and provided access to tools that facilitate this awareness. A suggested example of such a tool has been designed and created and is described in Chapter 11.

USE: Training on proper and acceptable use of personal information, including enforceable security protocols for access, use and storage.

DISSEMINATION: Training on proper and acceptable dissemination of personal information, if at all, the methods and formats in which it can be disseminated, the audience to whom it can be disseminated and methods for documenting and safeguarding dissemination.

BREACHES: standard procedures to be followed in the event that a breach is suspected or confirmed, and the responsibilities of the various parties that may be involved, including how, when and where to report a breach, and the role(s) of privacy and information commissioners.

It is important to keep in mind that the function of the training is not to saturate the individuals with privacy information, but rather give them the awareness and tools required to make responsible and privacy-respecting decisions. Any custodian-specific policy requirements should be clearly outlined in the contract (see below) and agreed upon by the recipient prior to data release.

#### **9.3.4. CONTRACT**

*What usually comes first is the contract*  
**Benjamin Disraeli**

This dimension simply indicates the comprehensiveness of a signed data-sharing agreement between the parties involved, governing standard operating procedures and best practices irrespective of the mechanism used for implementation (e.g. in the case of government organisations, for example, this may be through a Memorandum of Understanding (MOU)) or, as is the case with the NHS in the UK, assurances as assessed through the Information Governance Toolkit [184]. Unlike Disraeli's quote, however, the contract should actually come last - once all other components of this framework have been assessed leading to the necessary inclusions within it as

described below (framework domains appear in bold type where relevant). It is proposed that a contract include as many of the following considerations regarding the data being released as possible in order to be considered comprehensive:

**RECIPIENT:** The recipient of the data should be explicitly and uniquely identified.

This allows clear identification of the responsible individual(s) and thus also provides a level of accountability.

**CONTENT:** This should reflect the data domain, detailing the specifics of the data including fields or variables being provided to the recipient and the total number of records, as well as any transforms carried out to alter the data.

**PURPOSE:** A statement on the scope of work for which the data will be used, as reflected in the purpose domain. This may be for a specific project or set of projects only, projects and any related activities (“consistent use”), or the generic activities of the recipient.

**POTENTIAL GAINS:** This should outline any potential gains to the recipient, along with details on how they will be addressed. Potential gains to the provider should also be considered.

**RETENTION:** An indication of the length of time for which the recipient may keep the data, and should therefore reflect the purpose. Indefinite durations should be acceptable, provided they are justified in writing in the contract, and agreed upon by both parties. (purpose)

**DESTRUCTION:** Guidelines on how data will be deleted once the duration has expired (recipient)

**STORAGE:** Details on how and where the data may be stored, including any jurisdictional restrictions (recipient)

**AUTHENTICATION:** Guidance on requirements for accessing the data and levels of authentication if appropriate. For example, will all members of the research team have equal access authority, or will this be assigned at the discretion of the recipient? Will the data provider need to approve such authority on an individual basis? This should also take into consideration any cross-jurisdictional issues (for example where recipients are in different legal jurisdictions, if relevant, such as an individual in Canada collaborating with another in the United Kingdom.) (recipient)

**ACCESS:** Details on how the data will be accessed once authentication is successful. Will this be via a password on a networked computer, for example? (recipient)

**POLICIES:** Any custodian-specific policies not already covered in the previous sections and to which the recipient is expected to adhere should be clearly outlined. (recipient)

**DISCLOSURE:** confidentiality clauses governing the recipient's disclosure of the records provided, as well as agreement on the conditions and formats under which disclosure is acceptable (e.g. for publication; what can be disclosed, at what scale, and in what format). This should reflect the output domain.

**ACCOUNTABILITY:** Requirements for the recipient to inform the custodian of progress, breaches, etc. (recipient)

BREACH: Identification of what would constitute a breach and how it will be handled should one occur. (recipient)

ENFORCEMENT: What are the consequences to the recipient if a breach were to occur with the current data? How will this be enforced? (recipient)

It should be reiterated that the assessment of the *contract* dimension is based on the comprehensiveness of the agreement, not on an assessment of the components listed above which are more appropriately assessed under the corresponding domain. For example, two agreements identical on all components listed except for access, where one is stricter than the other, would be assessed identically on this dimension. The stricter access controls, however, would be captured in the assessment of the *security* dimension of the **recipient** domain.

#### 9.4. THE DATA

Much of the literature and legislation seek to identify specific variables that can either alone, or in combination with others, identify individuals (the latter often being referred to as "quasi-identifiers"). However, assessing the privacy-risk of data-sharing based simply on the presence or absence of specific identifiers or quasi-identifiers is overly simplistic and potentially unnecessarily prohibitive. Arguably, any data about an individual are personal and fall along a contextually influenced continuum of "identifiability" - the ability to identify a specific individual. Identifiability is a function of how specific the data are, as well as the amount of data involved. The latter can be in the form of multiple records for the same individual, or simply a combination of different attributes (captured through variables). In addition to identifiability, one must also consider the potential negative impact or consequences for individuals and communities from the release of this data to the recipient, as well from inadvertent release to others. The more sensitive the information, or the more readily it can be combined with other information, the greater the potential for negative impact. At times, however, this may also translate to greater value for or be justified by the *purpose*.

In combination with the dimensions of this domain, consideration must also be given to the ease with which one can link a given dataset with another. The higher the granularity and/or the multiplicity, the more information there is to inform accurate linking to other datasets. This then creates somewhat of a "feedback loop", since linking datasets would in turn increase multiplicity and, depending on the contents of the linked data, granularity. In doing so, this increases identifiability, which can be an issue if the linkage involves de-identified sensitive information or results in a dataset with higher representation in a smaller population.

### **Identification**

The name(s) of the dataset(s) being considered for sharing should be clearly identified along with a data dictionary and metadata for each one. The dictionary should identify each variable or field in the dataset, along with a description of what it captures and how it is coded or categorised if applicable. The metadata does not need to be exhaustive, though it is good practice to include metadata with any data release. At minimum, it should include the date and version, if applicable, of the dataset and identify the custodian, the sampling frame (or underlying population to which the data pertain) and the geographic extent captured by the sampling frame (e.g. name of neighbourhood, city, county, country, etc. as applicable). Another useful descriptor is the methodology employed for the data collection.

Once the dataset has been deemed appropriate for sharing, the metadata should also include the assessments from this domain. If the data have been modified (transformed) to facilitate sharing, the metadata should include the assessments from this domain pre- and post- transform.

## Dimensions

The **data** domain characterises the nature of the data being assessed for sharing based on the tightly interwoven dimensions of *Granularity*, *Multiplicity*, *Sensitivity* and *Size* as defined below.

### 9.4.1. GRANULARITY

*Every man's life ends the same way. It is only the details of how he lived and how he died that distinguish one man from another.*

**Ernest Hemingway**

Granularity is essentially the level of detail in the data. It affects the specificity of a single variable for a given record, and hence the probability it can be attributed to a specific individual. Thinking spatially, this becomes an issue of scale. For example, given only a street address of a patient and nothing more, the probability of correctly identifying the patient is  $1/i$  where  $i$  is the number of individuals at that address. This is more granular than neighbourhood, which in turn is more granular than city, etc. A more common example in public health practice is the categorisation of age; the greater the categorisation span, the lower the granularity. Therefore as granularity increases, so too does identifiability.

Changes in granularity must match the purpose and requirements of the intended use; increasing granularity to render a dataset anonymous or de-identified only makes sense if doing so will not compromise the purpose for which it will be used. It is therefore important to note that granularity can have a significant impact on the utility of the data; an issue addressed in the erosion dimension of the **transform** domain.

### 9.4.2. MULTIPLICITY

*Multiplicity: (1b) the number of components in a system; (2) a great number*

**Merriam Webster-Dictionary Online**

Multiplicity refers to the number of variables, either explicit or implicit, in the dataset, as well as the recurrence of individuals within the data. It affects the specificity of

combined variables for a given record and across multiple records if applicable, and thus the probability that together they can be attributed to a specific individual. The greater the multiplicity, the more unique individuals can become on data combinations and the greater the identifiability.

Explicit variables are self-explanatory: they are simply present in the data. Implicit variables are those that are characterised or implied by the dataset itself. For example, geography can be an implicit variable if it is known that a list of patients are all within a given area; sex can be an implicit variable in a dataset of patients with prostate cancer (which is therefore implicitly all male).

Recurrence of individuals within a dataset is akin to additional variables if the records are not identical (and assuming the recurrence is not an error). For example, a patient may recur in the dataset with multiple symptoms and one record per symptom. Or, recurrence can be due to a temporal factor - for example, multiple hospital visits or diagnoses, or a longitudinal study. In either case, such a "long" data format can often be converted to a "wide" format such that each individual only has one record, but all the data are reflected as added variables - hence the inclusion within the multiplicity dimension.

### 9.4.3. SENSITIVITY

*The sensitivity of men to small matters, and their indifference to great ones, indicates a strange inversion*

**Blaise Pascal**

There is no clear definition for what constitutes sensitive information. Some have taken a blanket approach, simply defining health or medical information as being sensitive, and thereby applying blanket rules and regulations [185-187]. Others have taken a more relative stance, defining it in terms of the potential consequences on individuals, organisations and communities: harm, distress, embarrassment, discrimination, inconvenience or financial loss [179,188]. Still others have mashed the two together,





it can also impact such things as insurance rates and even job eligibility [193]. Therefore, instead of attempting to define sensitive information in simple absolute terms, it is more appropriate to return to the relative perspective and define it in terms of potential consequences based on context.

As mentioned, the relative approach is sometimes couched within the context of some form of unauthorised access. However, while data sensitivity is indeed context-specific, this context is independent of method of access. Data that are not sensitive in the hands of an authorised user do not suddenly become sensitive if inappropriately revealed or if they fall into unauthorised hands. Let us therefore be clear and consistent; in keeping with the definition, data sensitivity is based on the potential consequences of their misuse, be it by an authorised or unauthorised individual. Within this relative approach, we simply consider the potential consequences.

Within this framework, then, sensitive information is defined to be any information which can be used to cause harm, distress, embarrassment, discrimination, inconvenience or financial loss to an individual or community. In this context, community is used broadly to refer to a group of individuals homogenous on at least one characteristic (e.g. belief, ethnicity, neighbourhood, habit, social media groups, profession, organisation, etc.). While assessment on these risks can inform the sensitivity of the data, however, sensitivity must be viewed within the broader and more comprehensive context outlined in this framework. Sensitivity is not a reason to withhold data if the purpose justifies their use within an appropriate ethical governance framework.

#### **9.4.4. SIZE**

The number of records in a dataset and the underlying population are both important because together they reflect the proportion of at-risk individuals or population of interest who are captured by the data. This is therefore dependent on both the size of

the dataset (i.e. number of different individuals it captures) and the size of the underlying relevant population. This relationship presents two different types of associated identification risk.

The first type of risk is perhaps the most frequently addressed in the literature and is the ability to identify individuals in the dataset. More directly related to the underlying population size, it is often mitigated through data suppression or areal aggregation and is referred to as the geographic area population size (GAPS) [58]. Generally, the smaller the GAPS, the fewer the number of individuals with common attributes and therefore the greater the identifiability ("generally" because this may not be the case in deliberately more homogenous communities, such as nursing homes). For example, consider a dataset containing a single male aged 78 years. The probability of identifying this individual if it is known that the data are from a small rural community is much greater than if they are from a large urban centre or even a nursing home where it is more likely that there will be more males ages 78 years. This risk is therefore also intertwined with the granularity and multiplicity dimensions; if instead of 78 years we were to provide a less granular age, such as 70-80 years, the size of the "eligible" underlying population pool increases and identifiability therefore decreases. However if we then add a variable such as ethnicity, we increase multiplicity and the size of the eligible underlying population pool decreases.

GAPS cut-offs can be estimated by using a composite "score" that considers both the granularity and multiplicity of the dataset. The maximum number of different possible combinations of the variables within a dataset is taken to be the product of the number of possible values for each one. This product is referred to as the *MaxCombs* (maximum number of combinations) [58,146]. The higher the *MaxCombs* value, the more "parsed" the data, and therefore the greater the potential for identifying an individual and the greater the GAPS cut-off. I have co-authored on this methodology;

please see Chapter 7 for more details and Appendix G (Volume II) for the relevant publication.

The second type of risk is less obvious and is the ability to assign an attribute or characteristic to any given individual or group without necessarily identifying the individual. Consider, for example, a dataset for Canada of females aged 10 to 14 years vaccinated for Human Papilloma Virus (HPV) in 2006, containing 3,900 records (the year 2006 is used for convenience and illustrative purposes as it allows for use of the 2006 Census for Canada). This count represents about 0.38% of Canada's female population aged 10-14 years and in the absence of additional identifying information it is impossible to point to any randomly selected 10-14 year old female and confidently state that she has been vaccinated. However, if instead of Canada, this dataset were for the town of Ajax in Durham, Ontario, Canada, it would actually represent almost 100% of the population of interest. One can then quite confidently infer that any given female aged 10-14 years in Ajax has received the HPV vaccine. While in this particular example the inference may be inconsequential, it illustrates the potential impact of the dataset size relative to the population it represents, particularly if the data were deemed to be of a sensitive nature.

## 9.5. THE PURPOSE

A recurring principle and theme in privacy legislation and indeed many of the proposed management models is that the data collected, stored and used be the minimum that is required to accomplish the purpose for which they are collected [60,61,80,176,182,194]. In many instances of public health, this requirement is violated as usage is typically for secondary purposes. The purpose, therefore, must merit this "violation" and must consequently be properly documented and assessed. This increases transparency which feeds back into the **recipient** domain.

## Identification

A clearly articulated statement describing the purpose for which the data are being requested should be included.

## Dimensions

The **purpose** domain is assessed using the dimensions of *contribution*, *necessity*, *effort* and *impact*.

### 9.5.1. CONTRIBUTION

As previously described, there is no standard definition or list of public health functions, activities or services, and the definition of public health practice is similarly inconsistent. What everyone does seem to be in general agreement about, however, is that the goal of the functions, services and activities is the same. Since our shrinking world also demands a global perspective, it is therefore suggested that the Public Health Agency of Canada's inclusion of a "healthy world" be adopted within the goal statement [30], which can therefore be phrased as follows:

*The goal of public health practice is to ensure healthy individuals and communities in a healthy world.*

This dimension therefore seeks to measure the relevance and scope of the purpose for which the data are requested to the goal as stated above.

### 9.5.2. NECESSITY

The *necessity* dimension addresses whether or not the purpose can be accomplished in the absence of the data under consideration for sharing and is directly related to the dimensions of the **data** domain. This is impacted by both the nature of the data fields, as well as their granularity. For example, age is often a necessary field for public health use, but it may not always be necessary in its most granular form (i.e. exact age). In some cases, distinctions between age groups may be adequate and little to no benefit is derived from additional detail. This dimension needs to be assessed for every field

being considered for sharing in the data, and may therefore impact dimensions and measures within the **data** domain such as granularity and the MaxCombs measure.

To avoid ambiguity, it is suggested that the assessment of this dimension take on a binary form; a simple Yes/No response for each field under consideration. Either the output is necessary or it is not as decided by the custodian or an external body and justified by the recipient. As with all components of this framework, dimensions and domains are intended to work together to inform the data-sharing process.

### **9.5.3. EFFORT**

This dimension captures the ease with which the data can be shared, and takes into consideration the two key barriers identified by practitioners in the previously conducted survey: legislation and bureaucracy.

Naturally, the purpose for which the data will be used must be legal. Although this may seem to be a somewhat obvious and simple requirement, assessment of this dimension can be quite complex due to two key factors: our ever-shrinking world and the blazing speed of technological advancement. Whether it is the mobility and resilience of pathogens in an increasingly connected and mobile world, the continued evolution of technological development with global environmental impacts, or the ability to reach millions across the globe with humanitarian aid and health-promoting propaganda within mere minutes of an event, it is clear that health events have no regard for man-made boundaries and jurisdictions and the pace of technology far exceeds that of legislative reform. Cross-jurisdictional collaboration and data-sharing is no longer a "nice-to-have" but a "must-have". Just as individuals function within societies, societies function within nations, which in turn function within a much larger global construct. Good public health must therefore be practiced within a global paradigm, and the increasing roles of and collaboration between nations and global organisations such as the World Health Organisation (WHO), the Pan-American Health

Organisation (PAHO) and the European Centre for Disease Prevention and Control (ECDC) are evidence of this. This recognition is exemplified through the vision statement of the Public Health Agency of Canada: "Healthier Canadians and Communities in a *healthier world*" [30] (italics added). It is also succinctly expressed on the home page of the Association of Schools of Public Health Website: "Global health is public health; Public health is your health; You are only as healthy as the world you live in" [195]. However, in the words of Shakespeare's *Hamlet*, "there's the rub": this global approach is significantly complicated by the legal landscape.

As previously discussed, privacy legislation in general permits the use of personal information in the interests of public safety or for research purposes, though there is no universal or standard application [154]. The solution, it is proposed, is not to introduce additional constraining legislation, but rather enabling guidelines for ethical professional conduct and consequences for unethical use. In assessing this dimension, the users should determine whether the purpose has any legal ramifications within all jurisdictions implicated, and whether any current legal issues or legislation prevent the distribution and/or use of the data, or if much effort is required to obtain legal permission and draw up the necessary legal documentation (if required). Cross-jurisdictional data-sharing should therefore also be assessed in terms of its practicability. In the end, these mechanisms must allow users to feel comfortable being accountable for their use of the data as opposed to fearing potential consequences due to misconceptions and misunderstanding of these mechanisms. In some cases, as has been previously mentioned, the sharing of the data may be made legally obligatory by the purpose. However even in such cases, it seems that the fear of privacy implications and the related issues around risk, trust and other dimensions addressed within this framework can outweigh considerations for public safety and mandatory reporting obligations [196].

The bureaucracy aspect is related to the legality aspect; generally, the greater the number of implicated jurisdictions the greater the bureaucracy involved. It will also be related to the processes and policies in place by the parties involved.

This dimension therefore rates the effort required to implement the data-sharing agreement or process and is, to a large degree, countered by the contribution dimension. If contribution is low but effort is high, the parties involved may wish to revisit the purpose and assess whether a different approach is required, or if it merits the required effort.

#### **9.5.4. IMPACT**

Impact refers to the benefits and risks related to both, the approval and the rejection of the data-sharing request. These must be considered in terms of multiple aspects: the individuals to whom the data pertain (we shall refer to them simply as the individuals), the general public or population (i.e. societal impact), specific communities, the custodian, the recipient, and public health knowledge. The most difficult of these to reconcile are the individual and societal impacts since they revolve around individualistic versus communitarian philosophies. These approaches have been the subject of much research and debate, but in the end, they amount to subjective philosophies with vague and non-committal legislative backing. The *impact* will be largely determined by the *intent*; research, for example, will have more of a positive impact on contribution to knowledge than a public health service, for example. An impact matrix summarising some of the benefits and risks to each aspect is suggested in Table 12 both for data disclosure and non-disclosure.



**Table 12: Impact matrix showing examples of the benefits and risks of disclosure and non-disclosure of data to implicated groups**

		<b>DISCLOSURE</b>	<b>NON-DISCLOSURE</b>
<b>INDIVIDUAL</b>	<b>RISKS</b>	<ul style="list-style-type: none"> <li>• No control over personal information</li> <li>• Potential for embarrassment, stigmatisation, discrimination, financial loss and refusal of services</li> </ul>	<ul style="list-style-type: none"> <li>• Compromised public health system</li> <li>• Inefficient public health services</li> <li>• Delayed or untimely response</li> <li>• Increased potential for illness</li> <li>• Compromised diagnostic and response functions</li> <li>• Safety at risk</li> </ul>
	<b>BENEFITS</b>	<ul style="list-style-type: none"> <li>• Access to improved services for health protection</li> <li>• Increased awareness and education</li> <li>• Well-informed promotion initiatives</li> <li>• Improved health through enhanced prevention &amp; control strategies</li> <li>• Engagement in public health</li> <li>• Contribution to community and population health</li> <li>• Contribution to knowledge</li> </ul>	<ul style="list-style-type: none"> <li>• No risk of breach or inappropriate disclosure</li> </ul>

**Table 12: Impact matrix showing examples of the benefits and risks of disclosure and non-disclosure of data to implicated groups (continued)**

		<b>DISCLOSURE</b>	<b>NON-DISCLOSURE</b>
<b>SOCIETY</b>	<b>RISKS</b>	<ul style="list-style-type: none"> <li>• Generalised mistrust in the event of a breach</li> </ul>	<ul style="list-style-type: none"> <li>• Compromised public health system</li> <li>• Inefficient public health services</li> <li>• Delayed or untimely response</li> <li>• Increased potential for illness</li> <li>• Compromised diagnostic and response functions</li> <li>• Safety at risk</li> </ul>
	<b>BENEFITS</b>	<ul style="list-style-type: none"> <li>• Improved health through enhanced prevention &amp; control strategies</li> <li>• Contribution to population and global health</li> <li>• Contribution to knowledge</li> </ul>	<ul style="list-style-type: none"> <li>• None identified (unless one believes that ignorance is bliss!)</li> </ul>

**Table 12: Impact matrix showing examples of the benefits and risks of disclosure and non-disclosure of data to implicated groups (continued)**

		<b>DISCLOSURE</b>	<b>NON-DISCLOSURE</b>
<b>COMMUNITY</b>	<b>RISKS</b>	<ul style="list-style-type: none"> <li>• Potential for embarrassment, stigmatisation, discrimination, financial loss and refusal of services</li> </ul>	<ul style="list-style-type: none"> <li>• Compromised public health system</li> <li>• Inefficient public health services</li> <li>• Delayed or untimely response</li> <li>• Increased potential for illness</li> <li>• Compromised diagnostic and response functions</li> <li>• Safety at risk</li> </ul>
	<b>BENEFITS</b>	<ul style="list-style-type: none"> <li>• Contribution to community, population and global health</li> <li>• Contribution to knowledge</li> </ul>	<ul style="list-style-type: none"> <li>• No risk of breach or inappropriate disclosure</li> </ul>

**Table 12: Impact matrix showing examples of the benefits and risks of disclosure and non-disclosure of data to implicated groups (continued)**

		<b>DISCLOSURE</b>	<b>NON-DISCLOSURE</b>
<b>CUSTODIAN</b>	<b>RISKS</b>	<ul style="list-style-type: none"> <li>• Liability due to limited control over released information</li> <li>• Potential backlash from individuals to whom the data pertains</li> <li>• Potential loss of trust, credibility &amp; financial implications in the event of a breach</li> </ul>	<ul style="list-style-type: none"> <li>• Loss of reputation (unless non-disclosure is justified)</li> <li>• Impact on individuals, society and communities</li> <li>• May be perceived as an obstacle to knowledge and public health practice</li> <li>• Potential breach of legislation, public trust or ethical duty</li> </ul>
	<b>BENEFITS</b>	<ul style="list-style-type: none"> <li>• Building and fostering of partnerships and networks</li> <li>• Contribution to individual, population, community and global health</li> <li>• Contribution to knowledge</li> </ul>	<ul style="list-style-type: none"> <li>• No risk of breach or inappropriate disclosure</li> <li>• Potential legal or ethical justification</li> </ul>

**Table 12: Impact matrix showing examples of the benefits and risks of disclosure and non-disclosure of data to implicated groups (continued)**

		<b>DISCLOSURE</b>	<b>NON-DISCLOSURE</b>
<b>RECIPIENT</b>	<b>RISKS</b>	<ul style="list-style-type: none"> <li>• Implementation of governance structure and protective measures</li> <li>• Potential loss of trust, credibility &amp; financial implications in the event of a breach</li> <li>• May clash with existing policies; political tensions</li> </ul>	<ul style="list-style-type: none"> <li>• Inability to accomplish purpose</li> <li>• Potential for inappropriate or un-informed decisions or public health actions</li> <li>• Decision-based evidence-making</li> </ul>
	<b>BENEFITS</b>	<ul style="list-style-type: none"> <li>• Ability to conduct desired activity</li> <li>• Contribution to individual, population, community and global health</li> <li>• Contribution to knowledge</li> <li>• Evidence to inform decision-making, policies and positive reform</li> <li>• Evidence-based decision-making</li> </ul>	<ul style="list-style-type: none"> <li>• None identified</li> </ul>

**Table 12: Impact matrix showing examples of the benefits and risks of disclosure and non-disclosure of data to implicated groups (continued)**

		<b>DISCLOSURE</b>	<b>NON-DISCLOSURE</b>
<b>KNOWLEDGE</b>	<b>RISKS</b>	<ul style="list-style-type: none"> <li>• May clash with existing policies; political tensions</li> <li>• Resistance to change</li> </ul>	<ul style="list-style-type: none"> <li>• Lack of information and evidence; no advancement, improvement or contribution of generalisable knowledge</li> </ul>
	<b>BENEFITS</b>	<ul style="list-style-type: none"> <li>• Contribution to individual, population, community and global health</li> <li>• Contribution to knowledge</li> <li>• Improved methodology</li> <li>• Affect positive change</li> <li>• Build evidence-base</li> </ul>	<ul style="list-style-type: none"> <li>• None identified</li> </ul>

## 9.6. THE TRANSFORM

As described in Part II, several tools and methods have been developed to anonymise textual data in public health, such as pseudonymisation, k-anonymity techniques and data suppression. However, there is very little in the way of location-anonymising tools and methods are limited, focusing largely on adaptive and semi-adaptive geomasking techniques. Within the current context, this framework domain focuses specifically on the assessment of location-transforms and their implication to the spatial dimension of the data.

### Identification

The name and brief description of the transform (if one is conducted) should be clearly stated, along with reference to its methodology. The parameters and settings of the transform should also be included.

## Dimensions

The **transform** domain assesses the data on four dimensions: *Uniqueness*, *Identification Risk*, *Erosion* and *Analytical Effect*. These dimensions reflect the effects commonly discussed in the literature, and also consequently addressed through the novel Multidimensional Point Transform (MPT) algorithm proposed by the author (please see Part II; manuscript has been accepted for publication [153]). All four dimensions require a comparison with the data prior to the application of a transform since they seek to assess the transform's effect.

### 9.6.1. UNIQUENESS

The *uniqueness* dimension reflects the degree to which an individual in a given population is unique on the combination of attributes released through the data requested (this is described in more detail in Part II of this study)

Within this framework, acceptable uniqueness thresholds must be defined by the data custodian prior to data release and should be a function of the data contents and the closely linked concept of re-identifiability as given below. A transform that is applied on the data – whether it is suppression, aggregation, or a more comprehensive multidimensional transform such as the one proposed by the author – must be able to estimate the effective anonymity level and operate in accordance with the established threshold requirements. However consideration must also be given to the utility of the transformed data (this is addressed by the last two dimensions of this domain).

The measure for this dimension, therefore, is how well a given transform can not only estimate the anonymity level of each record, but more importantly modify the data to meet the required threshold. Also, the transform should be able to allow the user to specify whether the anonymity is to be achieved within the dataset only, within the underlying population, or within both. If the latter two, the transform should also be able to perform the necessary operations using the underlying population if available (ideal)

or an appropriately synthesised population (acceptable). More details on and implementation of these aspects of location-transforms are given in Part II and the MPT manuscript [153].

### 9.6.2. IDENTIFICATION RISK

Within this framework, the terms identification risk and re-identification risk are considered synonymous since the underlying issue is the same - the ability to identify an individual. Identification risk is itself a multi-faceted concept as briefly described in the **data** domain and is closely related to uniqueness, as well as the nature of the transform and the way in which the data are released (as discussed in the **output** domain).

Within this dimension of the **transform** domain of this framework, the probability of being able to identify an individual using a given transform must be assessed based on the risk tolerance of the custodian. This must be done in the context of various types of identification risks as outlined in Chapter 7, and the necessary metrics must therefore be agreed upon for assessment. An appropriate transform must therefore include measures or estimates to mitigate the appropriate identification risk. The MPT proposed by this author mitigates several of these risks as described in Part II and the forthcoming publication [153]

### 9.6.3. EROSION

Erosion refers to the loss of Information both within the dataset (for example loss of complete values and changes to data values) as well as in relation to other contextual or environmental features. A description of how each attribute will be or has been transformed should therefore be provided to the recipient to assess whether or not the data will remain useful for the purpose for which they were requested. Transforms can impact the granularity of the data (e.g. by grouping attributes into categories), the size of the data (e.g. by suppressing records that do not meet uniqueness requirements)



and the multiplicity of the data (also through suppression, for example). This will ultimately affect the data's utility.

From a location privacy perspective, a fundamental geographical principle is that features that are closer together geographically are more related than those that are more distant. Therefore, the farther a record is displaced from its original location, the greater the erosion of its relationship to the geography, the environment and surrounding patients. Consider, for example, a study looking at environmental risk factors for patients with a respiratory condition. The farther any given patient is moved from their original location, the less valid the attributions to surrounding factors which may have otherwise been significant, such as vegetation, air pollution, the presence of industrial buildings, exposure to car fumes based on proximity to a busy road, etc.

Therefore, a dynamic and adaptive transform that allows users to define priorities for the erosion of various attributes relative to their own values as well as within the context of other attributes and features is ideal. To date, the only such transform that allows such an adaptive approach while including the spatial dimension by design in the overall anonymisation algorithm is the MPT [153].

#### **9.6.4. ANALYTICAL EFFECT**

Closely related to the dimension of erosion is the resulting analytical effect. For any given transformation of the data, the impact on analyses should also be assessed. For example, there is sufficient evidence to show the detrimental impact of aggregation on cluster detection [9,10,135]; aggregation also suffers from the commonly referred to MAUP: modifiable areal unit problem which has been previously discussed. The effects on simple descriptive spatial statistics is also important, and in some cases can be minimised by user-defined parameters to control the transform's algorithm. For example, in the MPT, the spatial mean of a set of points will be directly related to the allowable perturbation distance. However, this in turn will be influenced by other user-

defined settings for the algorithm [153]. From a location-privacy perspective, it is desirable to achieve a minimum geographic distance perturbation from the original point to reduce the analytical impact. However, the ideal transform will allow the user to control the degree of perturbation to each field or variable individually to minimise negative analytical effects specific to the purpose. For example, it may be more important to allow a slightly larger geographic distance transform (i.e. compromise the minimum geographic distance requirement) in favour of a reduced age transform (i.e. minimum age-perturbation) if age is more important for the analysis or has a larger effect on a dependent variable.

This dimension therefore seeks to assess and minimise the effects of a transform on the quality and validity of the information inferred or acquired through the data. The purpose is to reflect, as closely as possible, the information that would have otherwise been gleaned from the original, granular and non-transformed data. Once again, the MPT's flexible and adaptive methodology provides a powerful foundation on which to build [153].

## 9.7. THE OUTPUT

*Better health is not a science problem, it's an information problem*

**Thomas Goetz**

The **output** domain assesses the release, if any, of the final product created by the recipient. In most cases, this will likely be a summary of the recipient's findings, but it is, in essence, a "super-domain" since it is simply another instance of data-sharing. **Output** reflects the "so what" of the purpose - how is the information derived from the use of the data going to be disseminated, if at all, and to whom. This may itself be a specified component of the purpose (e.g. if the purpose is to inform a policy maker on the impact of an environmental exposure to cancer risk then once the recipient has completed the analysis, the findings will then have to be output to the appropriate policy maker). As indicated in the **recipient** domain, it is strongly suggested that disclosure requirements in a contract be based on the assessment of this domain

which should therefore be assessed and agreed upon by all parties involved in the data-sharing process.

**Output** is a critical domain, as it can be the least controllable and therefore have the most profound and tangible of consequences. It can provide the information and evidence required to fuel forces for positive change and improved health, or it can destroy reputations and individual and organisational livelihood. A map published in a local newspaper showing appropriately transformed locations of cases with schizophrenia may be acceptable, identifying areas in need of resources or support and commenting on mental health issues and the stresses that contribute to them. However if the locations are not appropriately transformed and can be reverse-engineered to identify individuals, the implications are far-reaching; not only will this impact the patients themselves, but it may also alienate neighbourhoods, generate a societal response and diminish the credibility of those involved.

Since the **output** is essentially, as stated, a form of "data-sharing", its assessment is based on the same four domains described above, with some minor nuances. It should be noted that the recipient does not become the custodian of the data once the data are shared, and the purpose for which the data-sharing is initiated may necessitate a particular type of output. It is therefore very important to include **output** as a domain because it is a necessary consideration in the initial data-sharing assessment between the custodian and the recipient. This also allows the **output** to be assessed holistically and independently of the original data-sharing assessment, ,

### **Recipient**

The **recipient** domain in this case reflects the audience and therefore focuses on who will have access to the **output**. The dimension must be assessed in conjunction with the **purpose** for the **output** and the **data** being output. The audience may be the general public, senior management, other researchers, clinicians, etc. In many cases,

the audience will likely be the general public; in a scoring scheme where a value of zero is used to indicate non-existence (e.g. no trust, no security, no training and no contract), the general public would score 0 on all four dimensions giving an overall **output recipient** score of 0. This would therefore necessitate a high score in other domains as appropriate and as illustrated and discussed in the next chapter.

## **Data**

The data to be released must be assessed in the context of the **recipient**, and in conjunction with the dimensions of the **transform** and **purpose** domains. As with the **output recipient** domain, the **output data** domain will also be assigned a very low score if the audience is the general public.

## **Purpose**

Again, assessment of the **output purpose** will depend heavily on the audience. For example, it may be necessary to output more granular and specific information for a core function of public health practice during an emergency than during routine surveillance using the same data. This would differ still from the relevant information being released to the public for a health promotion campaign.

## **Transform**

The **output transform** domain is the means by which risks associated with the **output** mitigated, and in this case addresses the specific **output** medium. Mediums can include written reports, oral presentations, posters, graphs, tables, maps, newscasts, etc. and must be considered in combination with the other domains. As previously mentioned, for example, publicly published maps can be used to identify individuals [11,150]. This may be appropriate for a quarantine team (audience) during a highly contagious outbreak (necessity) but not for public dissemination in a local newspaper.

## 10. From Conceptual to Concrete

### 10.1. INTRODUCTION

The framework as presented is built on five core domains, of which one, the **output** domain, is a "super-domain" requiring re-assessment on the other four domains. Together, these domains broadly capture the areas to be taken into consideration when evaluating the appropriateness of releasing or using public health data. The overall assessment and recommendation arising from the application of the framework is based on the combined and weighted scoring of these five domains.

Each domain is in turn built on a set of four dimensions, each also with its own scoring scheme and weight. The score for each of the domains is therefore a composite of the domain's weighted dimension scores. This gives a total of 16 dimensions, scored as a set for each data-sharing instance, of which the initial sharing request is one, and the output domain is a second such instance. Therefore, any given data-sharing assessment will involve scoring of at least two instances of the dimension set. These dimensions are not to be confused with those described in the multidimensional point transform presented in Part II.

As has been mentioned, the domains are interconnected, since they necessarily affect one another. The user should be given the ability to customise the scoring and weighting based on need and context within the framework's implementation. Consequently, situation-specific scoring schemes can be defined allowing the resulting recommendations to flexibly morph with contextual needs:

Dimension Weight:	For each dimension, the user can indicate the dimension's relative importance, which will be used to weight it in the domain's composite score.
-------------------	---

Domain Weight: For each domain, the user can indicate the domain's relative importance, which will be used to weight it in the overall computed score.

Acceptable Thresholds: Requirements to classify and assess the acceptability of weighted scores for the various dimensions and overall domains.

There are therefore two aspects to this framework and each of its components: the qualitative rationale, given by the actual content and criteria captured by the domains and dimensions, and the corresponding quantification that gives relative tangible results to help assess the former, issue guidance and inform decisions. The results of implementing the framework in this way may also identify changes that the custodian and/or recipient may benefit from. To better understand the scoring relationships between the various domains and dimensions and illustrate how the framework can be used to guide a data-sharing decision and learn from it, let us consider an implementation scenario.

## **10.2. SCENARIO SETUP**

Suppose that your division is looking to explore the relationship between geographical neighbourhood and the prevalence of sexually transmitted infections (STIs) in order to inform health promotion and sexual education campaigns. You are the senior advisor and principal investigator, and therefore have agreed to assume responsibility for your division. You therefore request an STI database from the appropriate health custodian in your jurisdiction, containing patient age, sex, address, primary diagnosis, number of partners, alcohol use, drug use, occupation and marital status. Together with the custodian, you complete a profile and assessment using the framework.

## **Weighting**

For each dimension and domain, the custodian has implemented a consistent weighting scheme from 1 to 5; the higher the weight, the more heavily the dimension or domain is weighted (signifying higher ascribed importance). A weight of 0 is possible, however not preferred since ideally, every dimension and domain in the framework should be utilised. However, the ability to set a weight of 0 allows the user(s) to customise the framework to specific scenarios. The custodian has provided the following attributions to each weight:

1. The dimension or domain is of very low significance to the custodian's determination of whether or not to share the data in a given context, and the assigned score is therefore used as is, without being weighted.
2. The dimension or domain is of some significance, but relatively low.
3. The dimension or domain is of reasonable significance
4. The dimension or domain is of high significance
5. The dimension or domain is of critical significance.

The custodian has therefore pre-assigned weights to the various dimensions and domains as described below. Note that dimensions within a domain can have the same weight if the custodian ascribes the same significance level to them. Furthermore, the weights are relative to one another and the effect is compounded by the assigned scoring scheme as explained below.

## Scoring

For each dimension and domain, the custodian has implemented a *rating* scheme of low, medium and high. The ratings are then quantified as scores: 1 where the dimension is rated as low, 2 for a medium rating, and 3 for a high rating. Since overall higher scores favour data-sharing, the scoring scheme is reversed where appropriate (this is clarified through the scenario). The

**Table 13: Example of a custodian weighted-score matrix**

		WEIGHT				
		1	2	3	4	5
SCORE	1	1	2	3	4	5
	2	2	4	6	8	10
	3	3	6	9	12	15

custodian has also implemented a score of 0 where a rating cannot be given (for example there is insufficient information or the dimension is completely absent).

The weighted score for any given dimension is calculated as the product of the dimension's score and weight. Therefore, as illustrated in the custodian's weighted score matrix (Table 13), a dimension of reasonable significance (weight=3) with a low rating (score = 1) will have the same weighted score (3) as a dimension of very low significance (weight = 1) and a high rating (score = 3).

The overall domain score is calculated as:

$$S \equiv \frac{\sum_{d=1}^n (s_d \times w_d)}{\text{maxwscore}}$$

Where  $S$  is the domain score,  $s$  is the dimension score,  $d$  is the dimension (i.e. first, second, third or fourth dimension),  $n$  is the number of dimensions (4),  $w$  is the dimension weight and  $\text{maxwscore}$  is the maximum possible weighted score for the domain calculated as

$$\text{maxwscore} \equiv \sum_{d=1}^n (\text{max}s_d \times w_d)$$



Where *maxs* is the maximum score that dimension *d* can have – in this case the contract dimension has a maximum score of 1, whereas the other three dimensions have a maximum score of 3 each.

As will be clear by the end of the scenario, the calculations and settings, along with their relationships between and within domains, are easy to change and define by the user, adding to the power and flexibility of the framework's implementation.

### **Thresholds**

The custodian has pre-determined that data-sharing will depend on an overall scoring profile where a score of 85% or higher is considered low risk (i.e. safe to share data), 70% to 85% considered medium risk, and anything under 70% to be high risk (data-sharing not acceptable).

#### **10.2.1. RECIPIENT DOMAIN**

##### **Identification**

Primary Contact:	Pat Smith
Role:	Principal Investigator / Public Health Promotion Advisor
Credentials:	MD, MHSc
Training:	No relevant privacy training
Contact:	STI Health Health Promotion Division 123 Imasample Drive, Imunreal 1-888-STI-HELP x. 123
Secondary Contact:	Alex Jones
Role:	Section Manager
Credentials:	MHSc, MBA

Training: No relevant privacy training

Contact: STI Health  
Health Promotion Division  
123 Imasample Drive, Imunreal  
1-888-STI-HELP x. 111

Other Contacts: Drew White; PhD; Senior Analyst  
Tracy Green; MSc; Health Promotion Advisor

### Rating

**Trust:** You and the custodian have had no previous interactions and are only known to one another through your respective organisational memberships. The lack of a relationship makes you both reluctant to rate this dimension as high, however your respective credentials, expertise, job requirements, lack of historical privacy breaches and motives as identified through your purpose allow the custodian to feel comfortable rating this dimension as Medium (score = 2).

**Security:** You provide detailed documentation of the policies and security measures implemented by your organisation; since you work for STI Health, it comes as no surprise to the custodian that these measures are in place. As a result, the security dimension is rated as high (score = 3)

**Training:** As indicated in your identification profile, you have had no relevant privacy training. Scoring on this dimension is therefore quite straightforward and you are rated as low (score = 1).

**Contract:** The custodian is used to these sorts of requests and has therefore implemented a detailed agreement to which you must agree prior to being approved for data-sharing. You have reviewed the agreement and after some discussion and

clarification with the custodian are prepared to sign it. In this case, the custodian has chosen to score this dimension on a binary scale – either 0 for no, or 1 for yes. The custodian is pleased that the requirements for this dimension have been fulfilled, and rates it as yes, a contract is (or will be) present (score = 1).

### **Weighting**

The custodian reasons that in situations where there has been no previous interaction with the recipient, the security dimension is the most important, greatly outweighing the rest, followed by the contract dimension, then the trust dimension, which in turn is more important to the custodian than whether or not you have had relevant privacy training. The trust dimension is therefore given a relative weight of 2, the security dimension a much higher weight of 5, the training dimension a weight of 1, and the contract dimension a relative weight of 4.

### **Scoring**

In this domain, a high score on any dimension is a better score for the domain overall.

Therefore, the overall score for this domain is calculated as follows:

$$S = \frac{(2 \times 2) + (3 \times 5) + (1 \times 1) + (1 \times 4)}{(3 \times 2) + (3 \times 5) + (3 \times 1) + (1 \times 4)} = \frac{23}{28} = 0.86$$

Your **recipient** domain score is therefore 86%; based on the custodian’s thresholds, this rates you as “high”, and therefore “low risk”.

## **10.2.2. DATA DOMAIN**

### **Identification**

Dataset Name: Imunreal STI Data

Dataset Date: July 1, 2011

Custodian: Imunreal Public Health

Patient Count:	704
Geographic Extent:	City of Imunreal
Extent Population:	41,000
Fields Requested:	Age (non categorised; range 21 – 71 years)
	Sex (male / female)
	Address (Street level)
	Primary Diagnosis (STI diagnosis: Chlamydia / Gonorrhoea / Syphilis / Candidiasis / Hepatitis B / Herpes / HIV / HPV)
	Number of partners (range 0 – 4)
	Alcohol use (binary: Yes / No)
	Drug use (binary: Yes/No)
	<b>Occupation</b> (free text)
	Marital Status (single / married / common law / divorced / separated / widowed)

## Rating

**Granularity:** Given that you are requesting street-level address, the custodian immediately rates this dimension as high (score = 1)

**Multiplicity:** Each record in the dataset represents a unique individual, so from that perspective multiplicity is low. However, given the number of variables you are requesting, the custodian feels that multiplicity is bordering on high. After some discussion, you agree to drop the occupation field as you reason that, given that it is free text, it is likely not as useful to your purpose as you had hoped. The custodian decides to rate this dimension as medium (score = 2).

**Sensitivity:** There is not even a slight pause on this one, and the dimension gets rated as high (score = 1).

**Size:** To put this in perspective, you and the custodian agree to look at two values: the proportion of individuals in Immunreal that this dataset represents, and the calculated maxcombs value. The proportion is the quotient of the patient count to the extent population, and is therefore quite low: 0.017. This is expected, since you already knew that the prevalence of STIs in Immunreal was around 2%. The maxcombs value is calculated as the total number of combinations given by the data fields. Age has 51 possible values; sex has 2; address has 704; primary diagnosis has 8; number of partners has 5; alcohol use has 2; drug use has 2; and marital status has 6. The product of these gives a maxcombs value of 68,935,680, which is considerably higher than the population of Immunreal, causing the custodian great concern – the high maxcombs value suggests a high likelihood that each individual in the dataset is unique in the population on their combination of variable values. The custodian therefore suggests that you consider implementing a location transform and rates this dimension as high (score = 1).

### **Weighting**

The custodian has determined that granularity and sensitivity are of high significance and substantially more important than multiplicity and size. Therefore, granularity and sensitivity are assigned a weight of 4 each, whereas multiplicity and size are each assigned a weight of 2.

### **Scoring**

In this domain, a high score on any dimension is actually a worse score for the domain overall, and therefore to maintain consistency, the scoring for the rating is reversed. In other words, “high” is re-assigned a score of 1, and “low” a score of 3 so that the score is “normalised” – the higher the score, the better the rating. Therefore, the overall score for this domain is calculated to be 0.39.

Your **data** domain score is therefore 39%; based on the custodian's thresholds, this rates poorly, and therefore as "high risk". Given the nature of the data, and the fact that we have not transformed it in any way, this is not surprising.

### 10.2.3. PURPOSE DOMAIN

#### Identification

The purpose of this study is to explore the relationship between geographical neighbourhood and the prevalence of sexually transmitted infections (STIs) in order to inform health promotion and sexual education campaigns. More specifically, campaigns will focus on the relationships between sexually transmitted infections and each of alcohol use, drug use, and sexual habits as reflected by the number of partners. The gender, age, address and marital status of cases will help inform different methods and target groups within defined neighbourhoods in order to tailor the campaigns to the specific population demographics.

#### Rating

**Contribution:** Although the prevalence of STIs in Imunreal is less than 2%, you and the custodian both agree strongly that continued informed promotional and educational campaigns would be greatly beneficial to Imunreal. The custodian therefore rates this dimension as high (score = 3).

**Necessity:** You outlined the way you would be using each variable in your purpose statements, so you have started off on the right foot. The custodian quizzes you on the necessity of exact age, as opposed to age groups, as well as whether or not you really need marital status, but in the end the custodian agrees to rate this dimension as "high" provided you add some rationale to your purpose for requiring all the variables requested. You agree (score = 3).

**Effort:** Since you and the custodian are both in the same jurisdiction, this is somewhat simple. Neither of you has read through the entire relevant legislation, but after consulting the relevant sections as outline in the Public Health Guide to Privacy Legislation prototype at <http://www.personplacetime.org/tools> and the legislation links on the site, you rate the effort as low. Because low effort is good, the scoring for this dimension is reversed as was the case with the **data** domain (score = 3).

**Impact:** To assess this dimension, the custodian decides to assign a composite score based only on the rating of the risks and benefits of sharing the data. The reasoning, the custodian argues, is that the risks of disclosing the data reflect the benefits of non-disclosure. So if, for example, there is a breach risk associated with disclosure, this translates to an absence of that risk associated with non-disclosure, and rating both would then cancel the other out or compound the effect, depending on the scoring scheme used. Therefore, only disclosure benefits and risks are rated.

**Disclosure benefits:** Because of the importance of the contribution as agreed on by you and the custodian, the custodian decides to rate the impact of disclosure benefits as "medium". However, you manage to argue that this is a significant issue, and show the custodian some literature indicating that the prevalence of STIs in Imunreal has actually increased slightly over the past few years. This gives more credence to your argument, and the custodian revises the rating to "high" (score = 3).

**Disclosure risks:** Given the high sensitivity of the data, as well as its granularity, the high risk associated with its size and the potential serious consequences of a breach, the custodian rates the risk associated with data-sharing as high (score = 3).

## Weighting

The custodian and you agree that contribution is a critical dimension of the purpose domain, followed by impact as a highly significant dimension, necessity as being of reasonable significance and finally effort as being of some significance. Contribution is therefore given a weight of 5, impact a weight of 4, necessity a weight of 3 and effort a weight of 2.

## Scoring

In this domain, high scores on contribution, necessity and beneficial impact of disclosure favour data-sharing, whereas high scores on effort and risks of disclosure discourage it. In order to remain consistent with the overall scheme where higher scores favour data-sharing, scoring of effort and the impact of disclosure need further consideration. As indicated in the effort dimension, the scoring is reversed; it is rated as low and therefore receives a score of 3.

The impact dimension, however, is given a composite score based on the combined disclosure benefit and risk ratings. Furthermore, the custodian has decided to weight the risk-impact score based on the security dimension score for the **recipient** domain.

The custodian reasons that if no security-adjustments are made, then a high benefit score would be negated by a high risk score, irrespective of the security measures in place by the recipient. Therefore, the custodian decides to adjust this by multiplying the benefit score by the security dimension score, and subtracting the risks score. Therefore, when the security is low (score = 1), then if both benefits and risks are rated as high, both receive a score of 3 and therefore balance one another out, giving an overall composite impact score of 0. If security is medium or high (score = 2 or 3 respectively) then the security-weighted impact of the benefits will increase (will become 6 or 9 respectively) resulting in a positive difference. Since a higher score favours data-sharing, this will be the case. However, if the benefits are rated as low



(score = 1) and the risks are high (score = 3), then if security is low (score = 1), the security-weighted impact score will be -2. When you consider that this can be compounded by the weight assigned to the impact, this can potentially indicate an extremely high risk (again, recall that a higher score favours data-sharing) In this scenario, this is the only dimension that can receive a negative score, which may be appropriate – if weighted highly and there is a high risk, then this will certainly skew the results and flag it as such.

Therefore, the overall score for this domain is 1. Note that in the calculation of maxscore for impact, the maximum impact score is achieved with high benefit and low risk, and is therefore  $3-1=2$  (not 3), multiplied by the security dimension score (in this case 3). Also, in this scenario, rating impact risk as unknown will assign it a score of 0, thereby favouring data-sharing; alternatively, it can be assigned a score of 4 to guarantee a conservative approach if the risks are not identified and all calculations will have to be adjusted accordingly.

Your **purpose** domain score is therefore 100% - a perfect score!

#### **10.2.4. TRANSFORM DOMAIN**

##### **Identification**

In spite of the custodian's suggestion, you feel that any transform you apply to the data would have a negative effect on achieving your purpose. Specifically, you argue that any perturbation in the location may result in a change in neighbourhood, and that the granularity of the other variables is equally as important. After some conversation with the custodian you agree to revisit this dimension should the final verdict be unfavourable.

## Rating

**Uniqueness:** Since exact address is provided without a transform, along with various other quasi-identifiers, the custodian immediately rates this dimension as "high". Again, because this does not favour data-sharing, the dimension is assigned a low score (score = 1).

**Identification Risk:** In the absence of a transform and given the granularity size of the data, the custodian rates this dimension as "high". As with uniqueness, it is therefore assigned a low score (score = 1).

**Erosion:** Since no transform is being conducted, there is no erosion and the dimension is rated as low. The custodian has decided that sharing more precise data is more favourable, and that this would be balanced out by the uniqueness and identification risk dimensions - in other words, the most favourable scenario is one in which all dimensions in this domain are low. Therefore, a low rating on this domain is given a high score to favour data-sharing (score = 3)

**Analytical Effect:** as with erosion, this is rated as low (score =3)

## Weighting

The custodian has decided that uniqueness and identification risk are critical, whereas erosion and analytical effect are of reasonable significance. The former two are therefore assigned a weight of 5, and the latter two a weight of 3.

## Scoring

Your **transform** domain score is 58%; based on the custodian's thresholds, this rates your transform as "low", and therefore "high risk" (expected given that no transform is carried out and therefore consistent with the scoring of the **data** domain).

### 10.2.5. OUTPUT DOMAIN

#### Identification

You stated in your purpose that your end goal is to inform health promotion and sexual education campaigns. To do this, you plan on working with your colleagues within your health promotion division to create and release various reports to the public. You will also be engaging an education specialist, a designer and your communications division, all of whom will be assessed as members of the public. The public will only have access to high-level generic results at the neighbourhood level. Reports and campaigns will include aggregated data, figures, maps and tables. You therefore proceed to rate and score this "super-domain" on each of the previous domains and their dimensions:

#### Rating

**Recipient:** The audience is the public. This domain is therefore rated low on all dimensions, and the contract dimension is given a "no" response. All dimensions are given a weight of 1.

The final score of the **recipient** domain within the **output** super-domain is low at 30%.

**Data:** The data to be released will be summarised by no more than 5 age groups to be determined by the findings, as well as gender differences. In agreement with the custodian, there will be no fewer than 20 age- and gender-stratified individuals within each neighbourhood (i.e.  $k=20$ ). Address will not be released in any format. Primary diagnosis will only be related to generic statements for the whole of Imunreal (i.e. no neighbourhood-specific diagnoses). The number of partners will be reflected through educational campaigns based on the findings across neighbourhoods, as will inclusion of alcohol and drug-related content, though with increased targeting of problematic neighbourhoods. Marital status will be used to inform STI counselling centres using the overall findings across neighbourhoods. No point-maps will be created - only

choropleth maps, to visually demonstrate differences, if any, between neighbourhoods on the above factors. Similarly with graphs and tables.

As with the request assessment, the custodian has determined that granularity and sensitivity are of high significance and substantially more important than multiplicity and size. Therefore, granularity and sensitivity are assigned a weight of 4 each, whereas multiplicity and size are each assigned a weight of 2.

Based on this information, all dimensions within the **Data** domain are rated as low. The final score of the **data** domain within the **output** super-domain is therefore high at 100%.

**Purpose:** The purpose within the **output** domain is essentially to reduce the impact and incidence of STIs in the population of Imunreal, inform educational campaigns in schools, community centres and neighbourhoods, and provide support to STI counsellors. As such, contribution, necessity and impact in terms of benefits are rated as high, and the effort and risks are rated as low.

The custodian has decided that contribution is critical, as is impact, and therefore gives each of these a weight of 5. Necessity and effort are weighted as 4 and 2 as before.

The final score of the **purpose** domain within the **output** super-domain is high at 100%

**Transform:** Since the data being released are obviously different from those being requested by the recipient, one or more transforms will have to be carried out on the data to ensure the minimum  $k = 20$  requirement is met using the appropriate age categories. By meeting the minimum  $k$  requirement, the custodian is happy to rate uniqueness and identification risk as low, although this also means that erosion and analytical effect are high.

As before, the custodian has decided that uniqueness and identification risk are critical. However, since the data are not being released for further analysis, erosion and analytical effect are of very low significance. The former two are therefore assigned a weight of 5, and the latter two a weight of 1.

The final score of the **transform** domain within the **output** super-domain is high at 89%.

### **Weighting**

Within the **output** super-domain, the custodian has decided that the **recipient** domain is not very significant, since it is the general public, however the remaining three domains of **data**, **purpose** and **transform** are of high significance. Therefore the former is given a weight of 1, whereas the latter are each given a weight of 4.

### **Scoring**

The scoring for the **output** super-domain is done in exactly the same way as for each domain, except in this case, the domains become the dimensions in the formula. Therefore, the final **output** score is high at 90%.

### **TOTAL SCORE**

Now that each domain has been scored, it is time to calculate the overall score to see whether or not the request can be fulfilled and the data shared. The final score is based on weighted domain scores.

### **Weighting**

The custodian recognises that in most cases, the data domain will score quite poorly simply because of the nature of STI data. Therefore, the custodian has implemented a domain weighting scheme where the three most important domains are the recipient,

the purpose and the output, considered to be critical to the data-sharing decision. The reasoning is that if the recipient can demonstrate the necessary requirements for safeguarding the data, the purpose merits its use and the output poses no risk, then this is more important than the fact that the data, by its very nature, lends itself poorly to data-sharing, and the transform dimension would therefore have to dramatically alter the data in order to facilitate their release. The custodian has not assigned these latter two domains a weight of zero, however, which means that they will still be considered in the final scoring; instead, they are considered to be of reasonable significance. The domain weighting scheme, therefore, is as follows: the recipient, purpose and output domains are each given a weight of 5, whereas the data and transform domains are given a weight of 3.

### **Scoring**

Using the above weights as applied to the score from each domain gives a total score of 80%. This is below the custodian's threshold of 85% for low-risk data-sharing, but within the "medium risk" range. The recommendation therefore is to review low-scoring domains and assess. For example, the recipient could agree to implement a transform, or omit one or more variables (e.g. marital status) which may therefore increase the score.

### **10.3. SCORING BIAS**

As may have been evident in the scenario, scoring of the dimensions can be quite subjective. Different raters may consistently favour lower or higher ratings, may intentionally give specific ratings to favour a desired outcome, or simply have different thresholds and understanding of the ratings that can easily lead to inter-rater variability. One way to reduce such biases was mentioned in Chapter 9: rate each dimension on a checklist of predefined "Yes/No" measures, which reduce subjectivity. Potential checklist measures were suggested for some of the dimensions in Chapter 9. Another way to reduce bias is to require a rationale for each rating, as suggested in the above

scenario, along with having the custodian and the recipient jointly involved in completing the assessment. The nature of the request and the results of the assessment could even be made publicly available, as is common with scientific grant applications. Given the appropriate governance structure, these various checks create a level of accountability for the rater(s) that would also help reduce bias. In any case, testing and evaluation of the selected rating scheme is required by users of the framework, who would benefit from the adoption of organisational standards and possible training on its consistent implementation. The weighting and scoring schemes would in turn be dependent on the organisation's policies for information governance and data sharing.

#### **10.4. SUMMARY AND CONCLUSIONS**

The scenario as presented above is a simple example of how the framework may be used in a very flexible and customisable implementation to help assess the decision of whether or not to share data in response to a data-sharing request. The domain and dimension weights, scores and thresholds are all individually customisable, allowing context-specific settings. Furthermore, interactive relationships can be built in, as illustrated by factoring the security dimension into the impact-risks dimension scoring. The implementation also allows the custodian and the recipient to quickly and easily identify the dimensions and/or domains that require additional changes in order to meet the required thresholds. Along with each domain and dimension, both the recipient and the custodian have the responsibility of documenting the requirements, the scoring and weighting justification or rationale, and the steps taken to address the requirements. These, along with the final results, can then be used to justify the data-release to a research ethics committee or review board if required.





**PART IV  
MOVING FORWARD**



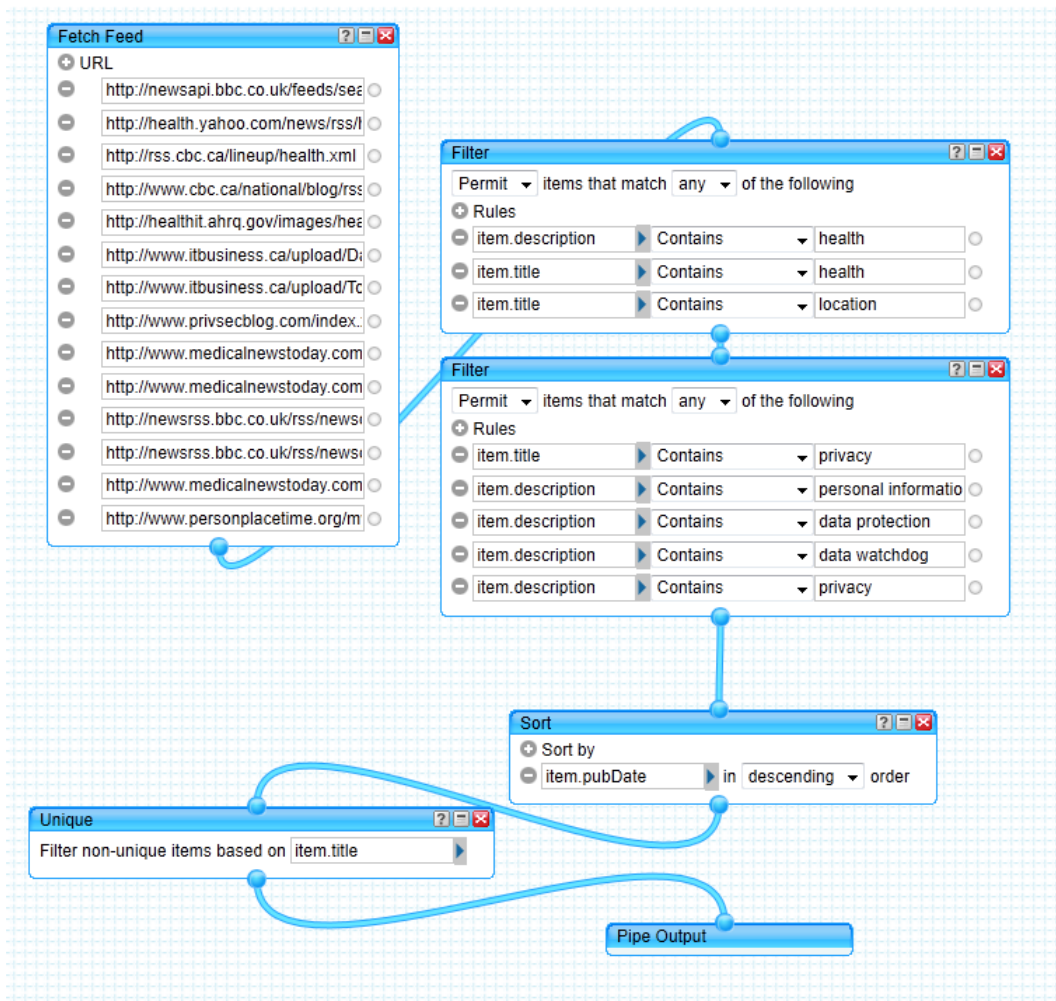
## 11. Facilitating Practice Through Tools

### 11.1. INTRODUCTION

Much of the content of Part I through Part III has much utility beyond this written tome; indeed, as indicated by many of the survey responses outlined in Part I, this research is much needed by the public health community. Information, however, is rather useless in isolation; it is meant to inform, and must therefore be made available for consumption.

To facilitate the sharing of information and knowledge, a research Website was created at study inception and can be found at <http://www.personplacetime.org>. One of the key ingredients on the home page of the site is the inclusion of an automated news-feed on privacy-related news items in public health from around the world. The feed was created by using Yahoo Pipes, which is a Web-based tool that allows the user to aggregate and mashup content from around the Web [197]. This was then integrated into the research site. The pipe incorporates updates from a variety of Web sources, filters on specific keywords in the description and title, sorts by date, checks for duplicates and outputs the resulting unique news items; a flow schematic as created in pipes is shown in Figure 18 and the result as displayed on the research home page is shown in Figure 19. This is a useful way of staying abreast of the most recent news and updates related to the areas of interest while maintaining a historical thread.

**Figure 18: Yahoo Pipe created to automate RSS feed collection from several Website on privacy and health headlines**



In addition to using the Web to disseminate information, it is beneficial to develop intuitive and user-friendly tools to help guide public health practitioners in the relevant aspects of the research. While the development of such tools was not the focus of this study, some examples and prototypes were none-the-less designed as "proof of concept" models. These are presented below and are all available through the research site at <http://www.personplacetime.org/tools>.

**Figure 19: Screenshot of the Website homepage showing the results of the Yahoo Pipe for the news feed mashup**



For questions or comments regarding this site, please send me an [email](mailto:Philip.AbdelMalik@personplacetime.org)  
©Philip AbdelMalik  
Site Last Updated: 03-Jul-2011

## 11.2. DEMISTIFYING THE LEGISLATION

In Part I we reviewed the current legislative landscape as it pertains to privacy and public health practice, and identified the public health professional community's perceptions. Indeed, according to the Office of the Privacy Commissioner of Canada, it is a lack of understanding of legislation that leads to the perception that privacy laws impede or compromise safety and security [113]. As the Office acknowledges:

*Departments have been taken to task by our Office for disclosing personal information when they should not have, so it is not surprising that they might often err on the side of caution...There is a presumption in favour of non-disclosure unless there are compelling arguments to the contrary. [113]*

We also argued for the requirement for a global approach to public health, which therefore necessitates a capacity for data-sharing on a global scale, which in turn necessitates cross-jurisdictional legislation or legislative reform. In addition, within the framework presented in Part III, we identified a requirement for the purpose to be consistent with legislative requirements in the relevant jurisdictions. Yet legislation is often difficult to read and understand, and for the public health professional, allaying data-sharing fears over legal obligations is important for facilitating the flow of information.

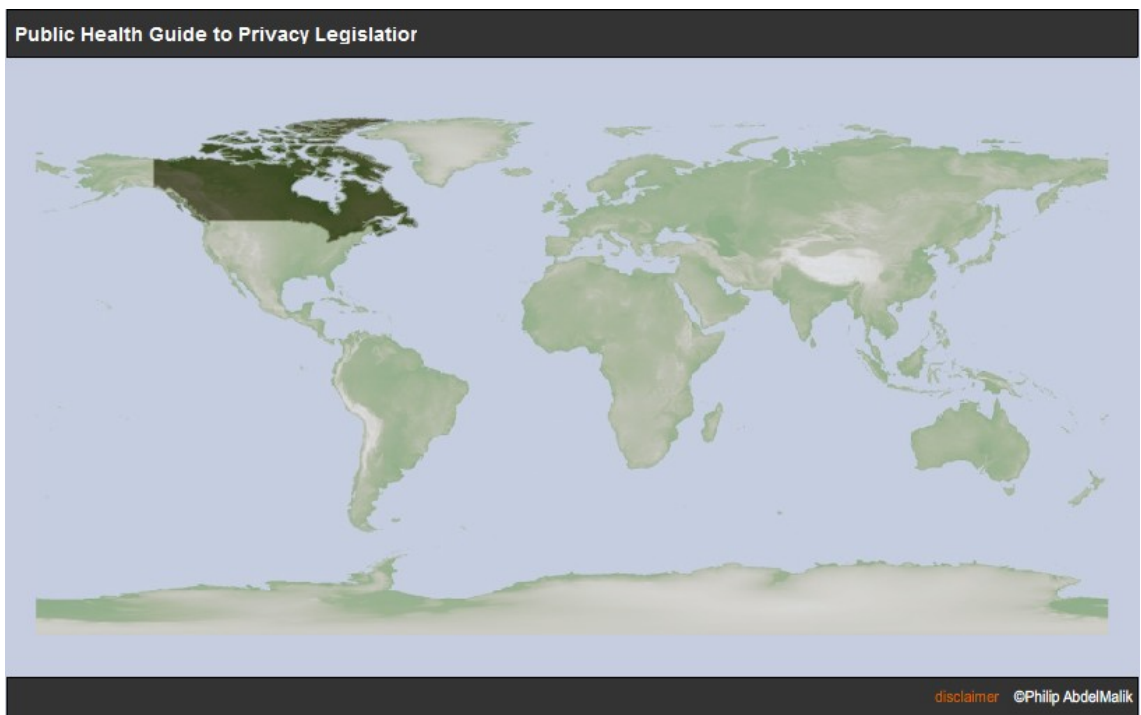
To avoid "erring on the side of caution" and facilitate this flow, it would be useful to have a user-friendly tool that clearly summarises legislation as it pertains to privacy and public health issues for practitioners to consult. An example of such a tool was developed by the author as a flash. As a prototype only, the tool currently contains information for Canadian federal legislation and provincial legislation specific to British Columbia. Only legislation relevant to the issues around privacy and public health is covered. A quick walk-through of the tool's legislative summaries is provided below.

On accepting the terms presented in the disclaimer (Figure 20), the user is then taken to a world map to select the jurisdiction of interest. Available countries appear darker than the others, and currently only Canada is selectable (Figure 21).

**Figure 20: Splash screen of the demo tool designed as a public health guide to privacy legislation (available through the study Website)**



**Figure 21: World map in the public health guide to privacy legislation; selectable countries - currently Canada only - appear darker**




Upon selection, the user is then shown a summary of the relevant legislative components (Figure 22). On the left, a summary of the jurisdiction is shown, along with

a map of selectable sub-jurisdictions, if appropriate. In the case of Canada, for example, the nation is composed of ten provinces and three territories, each of which has its own legislation. These are reflected in the map of Canada on the left-hand side of the screen (Figure 22).

**Figure 22: Summary of relevant privacy legislation for Canada as shown in the public health guide for privacy legislation tool**

**Public Health Guide to Privacy Legislation - CANADA**
FEDERAL

Canada is made up of ten provinces and three territories, each of which is split up into local health regions. Federally, there are two Acts governing Privacy: *The Privacy Act*, which governs federal government bodies, and the Personal Information Protection and Electronic Documents Act (PIPEDA), governing private sector bodies. Provinces are subject to PIPEDA, unless they have legislation that has been deemed sufficiently similar.



ALBERTA	NUNAVUT
BRITISH COLUMBIA	ONTARIO
MANITOBA	PRINCE EDWARD ISLAND
NEW BRUNSWICK	QUEBEC
NEWFOUNDLAND & LABRADOR	SASKATCHEWAN
NORTHWEST TERRITORIES	YUKON TERRITORY
NOVA SCOTIA	

	Statistics Act	Health Act	PIPEDA	Privacy Act
Reference:	R.S. 1985, c.S-19	R.S.C., 1985, c.C-6	S.C. 2000, c. 5	R.S.C. 1985, c.P-21
Scope:	Statistics Canada Chief Statistician	Provinces	Private Sector	Federal Government
Exemptions:	None Identified		Federal Departments See Section 7	Various library & museum items
PID Definition:	None	None	Vague; Some Detail	Clear; some details missing
Includes Geography:				✓
Includes Health Info:			✓	✓
Includes Research:				
Exceptions made for:				
Geography:	✓			✓
Health Information:			✓	
Research:			✓	✓

**Navigation...**

To get more information specific to the contents of the Acts on geography, health information or research, click on the name of the Act at the top and the information will appear in this box.

To open up a copy of the Act in a new window, click on its [Reference](#) above

Scope: Who the Act applies to  
Exemptions: Groups exempt from the Act  
PID: Personally Identifiable Data

Use the map to the left to navigate to province-specific legislation

world
©Philip AbdelMalik

The right-hand side of the screen displays the legislative summary in a grid. The first row of the grid contains the name of the Act or Statute; clicking on the name will display additional information in the box below the grid. The items in the grid are as follows:

- |             |   |
|-------------|---|
| Reference:  | Clicking this opens up a copy of the Act or Statute |
| Scope:      | The entities to whom the Act or Statute applies     |
| Exemptions: | Exemptions to those covered under the scope         |



PID Definition:	Whether or not "personally identifiable data" is defined.
Includes Geography:	Whether or not the PID definition includes geography
Includes Health Info:	Whether or not the PID definition includes health information
Includes Research:	Whether or not the PID definition includes research
Exceptions made for:	
Geography:	Whether or not the Act or Statute provides any exceptions or circumstances that allow the release or use of geography
Health Info:	Whether or not the Act or Statute provides any exceptions or circumstances that allow the release or use of health information
Research:	Whether or not the Act or Statute provides any exceptions for research purposes

The design of the tool provides a quick and intuitive visual of what is and is not covered in the legislation, as well as the relevant sections and links for further reference.

### **11.3. APPLYING THE MULTIDIMENSIONAL POINT TRANSFORM**

In Part II, we explored some novel methods for working with data. Among these was the novel Multidimensional Point Transform, and a static image of what an application might look like was designed and referenced in the published manuscript [153]. Since publication, however, a slightly modified design has been created as found in Figure 23 and through the tools page of the Website.

**Figure 23: Example of a user interface for a tool to implement the Multidimensional Point Transform**

Multidimensional Point Transform

Data to transform:  ...

Successfully read 400 records preview

**Location Information**

Areal Unit:

Latitude Field:  Longitude Field:

**Base Population(s) Used**

---

Minimum Anonymity  $k$ :  The minimum anonymity level required. For example,  $k=5$  means any given transformed records is indistinguishable on defined dimensions from at least 4 others.

Perturbation Distance: Minimum:  The minimum distance a record can be displaced, **in metres**      Maximum:  The maximum distance a record can be displaced, **in metres**

Add random distance: Minimum:  The minimum random distance to add to the perturbation distance **in metres**      Maximum:  The maximum random distance to add to the perturbation distance **in metres**

Allow Original Point Selection?     YES     NO    Whether or not to allow random selection of the original point's location as the transformed point's location.

**Dimensions to Match & Anonymise On:**  
This is where you define the dimensions that the transform uses to achieve the desired minimum  $k$ -anonymity level you defined above to appropriately anonymise each record.

Dimension	Priority	Perturb	Increments	Interval	
Location	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<b>Priority:</b> A higher priority attempts to minimise that dimension's perturbation. Default is 1 (low)
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<b>Perturb:</b> Whether or not the dimension's original values can be perturbed. If set to YES, then <b>Increments &amp; Intervals</b> will be required.
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<b>Increments:</b> The maximum number of increments allowed for a dimension
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<b>Interval:</b> The number of units within each increment.
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	If <b>Perturb</b> is YES and <b>Increments</b> is 0, the dimension will be re-coded from its minimum value to its maximum value.
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	If <b>Perturb</b> is YES, <b>Increments</b> is NOT 0 but <b>Interval</b> is 0, the dimension will be incremented using equal intervals, calculated as the range divided by the number of <b>Increments</b> .

The proposed design splits the screen into two parts. The top part is where the user identifies the location of the case data (e.g. C:\My Data\NY\_patients.csv), the base population to use (e.g. New York County) and whether to use an areal or coordinate-based location. If an aerial location is chosen, the user must identify the appropriate field in the case data (e.g. postcode or zip code). In the current example, a coordinate-based location is chosen and the case data fields containing the latitude (Y) and longitude (X) for each record have been specified. Once the data are successfully imported, the tool then indicates the number of successfully read records (e.g. 400) and allows the user to preview the data (e.g. to check for import errors).

The next part of the screen is where the user defines all the parameter settings. The minimum required  $k$ -anonymity, the minimum and maximum perturbation distances, the

minimum and maximum added random distance, and whether or not to allow selection of the original point. The user can then set the dimensions on which the transform will operate, specifying the field name for each, the priority, whether or not the dimension is allowed to be perturbed, and if so, the degree to which it can be perturbed. Since the transform is designed to integrate location, which in this example is specified by latitude and longitude, the "LOCATION" dimension will not appear in the user's case data - instead, it is a function that would be built into the application and would use the location information specified above (in this example, the latitude and longitude as defined by the user to be represented by Y and X respectively).

In the example appearing in the published manuscript ([153]; Figure 24), the data are transformed on the dimensions of location, age and sex. All three are given equal priority, which will cause the algorithm to proceed as described in Chapter 8. Location is allowed to be perturbed, but as indicated in the image (and described in Chapter 8) the transform can also be allowed to perturb other dimensions only if so desired, thereby disregarding the location dimension altogether. The age dimension is allowed to be perturbed, as described in Chapter 8 - that is, in 1-unit intervals to a maximum of 4-unit increments (in the case of age, the units would typically be years). Note that, as described in the algorithm, this does not simply increment the age as provided in the case data, but rather slots the age into the calculated categories defined by the increments and intervals, starting at age 0. In other words, for an individual case who is 3 years of age, the age perturbation does not categorise to 3-4 years, 3-5 years, 3-6 years and 3-7 years, but rather 2-3 years, 3-5 years, 0-3 years and 0-4 years, thereby maintaining consistent categorisation across the dataset.

**Figure 24: Multidimensional Point Transform tool interface design as originally published showing example settings**

### Multidimensional Point Transform

Data to transform:  ...

Successfully read 400 records PREVIEW

Latitude Field:  Longitude Field:

**Base Population(s) Used**  
- New York County

Minimum Anonymity  $k$ :  The minimum anonymity level required. For example,  $k = 5$  means any given transformed record is indistinguishable on defined dimensions from at least 4 others.

Maximum Perturbation Distance:  The maximum distance a record can be displaced, in metres.

Allow Original Point Selection?  YES  NO Whether or not to allow random selection of the original point's location as the transformed point's location.

Dimensions to Match & Anonymise On:  
This is where you define the dimensions that the transform uses to achieve the desired minimum  $k$ -anonymity level you defined above to appropriately anonymise each record.

Dimension	Priority	Perturb	Increments	Interval
LOCATION	1	YES	0	0
Age	1	YES	4	1
Sex	1	NO		

**Priority:** A higher priority attempts to minimise that dimension's perturbation. Default is 1 (low).

**Perturb:** Whether or not the dimension's original values can be perturbed. If set to YES, then **Increments & Interval** will be requested.

**Increments:** The maximum number of increments allowed for a dimension

**Interval:** The number of units within each increment

If **Perturb** is YES and **Increments** is 0, the dimension will be re-coded from its minimum value to its maximum value. If **Perturb** is YES, **Increments** is NOT 0 but **Interval** is 0, the dimension will be incremented using equal intervals, calculated as the range divided by the number of **Increments**.

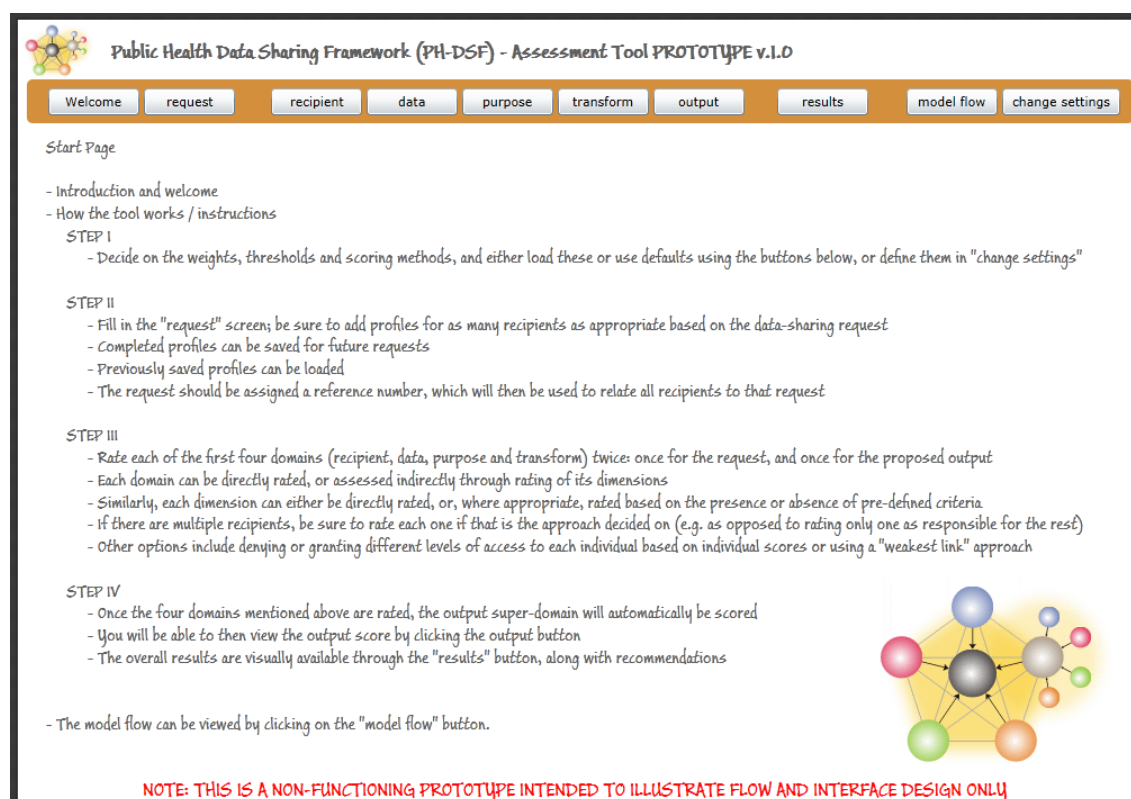
Example

← CLEAR
TRANSFORM →

## 11.4. GUIDING DATA-SHARING

Finally, in Part III, we explored a novel framework for assessing data-sharing initiatives and providing guidance on domains and dimensions to consider, and illustrated its use through its application to a fictitious scenario. To facilitate this use, and also allow individuals to "play" with the results, the scenario was implemented through a functioning Microsoft Excel workbook, allowing users to set weights and thresholds and experiment with scoring of the dimensions. In addition, a design of a prototype Web-based application (Figure 25) has been developed using Microsoft Expression Blend Sketchflow. The prototype requires an internet connection, a Web browser capable of using Microsoft Silverlight, and Microsoft Silverlight itself. Silverlight can be downloaded at no cost at <http://www.microsoft.com>. Both the Excel file and the prototype can be found in the tools section of the research Web site as specified previously.

**Figure 25: Prototype tool available through the study Website showing an example of how a practical implementation of the framework might look**



The Microsoft Excel Workbook allows users to implement the framework by defining weights and thresholds, and scoring the individual domains and dimensions. The results are displayed graphically through colour-coded circles (only visible with MS Office Excel 2007 or higher) and a colour-coded recommendation to allow quick visual assessment and identification of strengths and weaknesses. This provides the quantitative scoring aspect of the framework which should be complemented, as described previously, with qualitative justification.

The workbook consists of eight tabs:

**ASSESSMENT:** This is where the user scores the domains and dimensions (Figure 26). For each of the first four domains - **Recipient, Data, Purpose and Transform** - the user provides a score for the initial request, as well as for the **Output** super-domain. These scores inform the final **Output** score, and the final results are also

automatically computed. Based on the final weighted score, a colour-coded recommendation is given at the top and bottom of the worksheet: red for high risk (data-sharing not recommended), yellow for medium risk (review required) and green for low risk (data-sharing okay) based on the user-defined thresholds.

**Figure 26: Assessment worksheet of the Microsoft Excel workbook developed as an example implementation of the data-sharing framework.**

PRIVACY FRAMEWORK WORKSHEET			PLEASE REVIEW LOW SCORING DOMAINS																																									
<p><b>RECIPIENT</b> <span style="float:right">REQUEST <span style="color:red">●</span> OUTPUT <span style="color:red">●</span></span></p> <p>PLEASE RATE THE RECIPIENT ON THE FOLLOWING:</p> <table border="1"> <tr><td>LEVEL OF TRUST:</td><td>MEDIUM</td><td>LOW</td></tr> <tr><td>RECIPIENT'S SECURITY:</td><td>HIGH</td><td>LOW</td></tr> <tr><td>RECIPIENT'S RELEVANT TRAINING:</td><td>LOW</td><td>LOW</td></tr> <tr><td>IS THERE A WRITTEN CONTRACT?</td><td>YES</td><td>NO</td></tr> <tr><td><b>USER SCORE:</b></td><td><b>0.857142857</b></td><td><b>0.3</b></td></tr> <tr><td><b>USER RATING:</b></td><td><b>HIGH</b></td><td><b>LOW</b></td></tr> </table>			LEVEL OF TRUST:	MEDIUM	LOW	RECIPIENT'S SECURITY:	HIGH	LOW	RECIPIENT'S RELEVANT TRAINING:	LOW	LOW	IS THERE A WRITTEN CONTRACT?	YES	NO	<b>USER SCORE:</b>	<b>0.857142857</b>	<b>0.3</b>	<b>USER RATING:</b>	<b>HIGH</b>	<b>LOW</b>	<p><b>DATA</b> <span style="float:right">REQUEST <span style="color:red">●</span> OUTPUT <span style="color:green">●</span></span></p> <p>PLEASE RATE THE DATA ON THE FOLLOWING:</p> <table border="1"> <tr><td>GRANULARITY</td><td>HIGH</td><td>LOW</td></tr> <tr><td>MULTIPLICITY</td><td>MEDIUM</td><td>LOW</td></tr> <tr><td>SENSITIVITY</td><td>HIGH</td><td>LOW</td></tr> <tr><td>SIZE</td><td>HIGH</td><td>LOW</td></tr> <tr><td><b>DATA SCORE:</b></td><td><b>0.38888889</b></td><td><b>1</b></td></tr> <tr><td><b>DATA RATING:</b></td><td><b>LOW</b></td><td><b>HIGH</b></td></tr> </table>			GRANULARITY	HIGH	LOW	MULTIPLICITY	MEDIUM	LOW	SENSITIVITY	HIGH	LOW	SIZE	HIGH	LOW	<b>DATA SCORE:</b>	<b>0.38888889</b>	<b>1</b>	<b>DATA RATING:</b>	<b>LOW</b>	<b>HIGH</b>			
LEVEL OF TRUST:	MEDIUM	LOW																																										
RECIPIENT'S SECURITY:	HIGH	LOW																																										
RECIPIENT'S RELEVANT TRAINING:	LOW	LOW																																										
IS THERE A WRITTEN CONTRACT?	YES	NO																																										
<b>USER SCORE:</b>	<b>0.857142857</b>	<b>0.3</b>																																										
<b>USER RATING:</b>	<b>HIGH</b>	<b>LOW</b>																																										
GRANULARITY	HIGH	LOW																																										
MULTIPLICITY	MEDIUM	LOW																																										
SENSITIVITY	HIGH	LOW																																										
SIZE	HIGH	LOW																																										
<b>DATA SCORE:</b>	<b>0.38888889</b>	<b>1</b>																																										
<b>DATA RATING:</b>	<b>LOW</b>	<b>HIGH</b>																																										
<p><b>PURPOSE</b> <span style="float:right">REQUEST <span style="color:green">●</span> OUTPUT <span style="color:green">●</span></span></p> <p>PLEASE RATE THE PURPOSE ON THE FOLLOWING:</p> <table border="1"> <tr><td>CONTRIBUTION</td><td>HIGH</td><td>HIGH</td></tr> <tr><td>NECESSITY</td><td>HIGH</td><td>HIGH</td></tr> <tr><td>EFFORT</td><td>LOW</td><td>LOW</td></tr> <tr><td>IMPACT: BENEFITS OF SHARING</td><td>HIGH</td><td>HIGH</td></tr> <tr><td>IMPACT: RISKS OF SHARING</td><td>HIGH</td><td>LOW</td></tr> <tr><td><b>PURPOSE SCORE:</b></td><td><b>1</b></td><td><b>1</b></td></tr> <tr><td><b>PURPOSE RATING:</b></td><td><b>HIGH</b></td><td><b>HIGH</b></td></tr> </table>			CONTRIBUTION	HIGH	HIGH	NECESSITY	HIGH	HIGH	EFFORT	LOW	LOW	IMPACT: BENEFITS OF SHARING	HIGH	HIGH	IMPACT: RISKS OF SHARING	HIGH	LOW	<b>PURPOSE SCORE:</b>	<b>1</b>	<b>1</b>	<b>PURPOSE RATING:</b>	<b>HIGH</b>	<b>HIGH</b>	<p><b>TRANSFORM</b> <span style="float:right">REQUEST <span style="color:red">●</span> OUTPUT <span style="color:green">●</span></span></p> <p>PLEASE RATE THE TRANSFORM ON THE FOLLOWING:</p> <table border="1"> <tr><td>UNIQUENESS</td><td>HIGH</td><td>LOW</td></tr> <tr><td>IDENTIFICATION RISK</td><td>HIGH</td><td>LOW</td></tr> <tr><td>EROSION</td><td>LOW</td><td>HIGH</td></tr> <tr><td>ANALYTICAL EFFECT</td><td>LOW</td><td>HIGH</td></tr> <tr><td><b>TRANSFORM SCORE:</b></td><td><b>0.58333333</b></td><td><b>0.88888889</b></td></tr> <tr><td><b>TRANSFORM RATING:</b></td><td><b>LOW</b></td><td><b>HIGH</b></td></tr> </table>			UNIQUENESS	HIGH	LOW	IDENTIFICATION RISK	HIGH	LOW	EROSION	LOW	HIGH	ANALYTICAL EFFECT	LOW	HIGH	<b>TRANSFORM SCORE:</b>	<b>0.58333333</b>	<b>0.88888889</b>	<b>TRANSFORM RATING:</b>	<b>LOW</b>	<b>HIGH</b>
CONTRIBUTION	HIGH	HIGH																																										
NECESSITY	HIGH	HIGH																																										
EFFORT	LOW	LOW																																										
IMPACT: BENEFITS OF SHARING	HIGH	HIGH																																										
IMPACT: RISKS OF SHARING	HIGH	LOW																																										
<b>PURPOSE SCORE:</b>	<b>1</b>	<b>1</b>																																										
<b>PURPOSE RATING:</b>	<b>HIGH</b>	<b>HIGH</b>																																										
UNIQUENESS	HIGH	LOW																																										
IDENTIFICATION RISK	HIGH	LOW																																										
EROSION	LOW	HIGH																																										
ANALYTICAL EFFECT	LOW	HIGH																																										
<b>TRANSFORM SCORE:</b>	<b>0.58333333</b>	<b>0.88888889</b>																																										
<b>TRANSFORM RATING:</b>	<b>LOW</b>	<b>HIGH</b>																																										
<p><b>OUTPUT</b> <span style="float:right">REQUEST <span style="color:green">●</span> OUTPUT <span style="color:green">●</span></span></p> <p>AUTO-SCORES BASED ON ABOVE ASSESSMENTS</p> <table border="1"> <tr><td>RECIPIENT</td><td>LOW</td><td><span style="color:red">●</span></td></tr> <tr><td>DATA</td><td>HIGH</td><td><span style="color:green">●</span></td></tr> <tr><td>PURPOSE</td><td>HIGH</td><td><span style="color:green">●</span></td></tr> <tr><td>TRANSFORM</td><td>HIGH</td><td><span style="color:green">●</span></td></tr> <tr><td><b>OVERALL OUTPUT SCORE:</b></td><td><b>0.90462963</b></td><td></td></tr> <tr><td><b>OVERALL OUTPUT RATING:</b></td><td><b>HIGH</b></td><td></td></tr> </table>			RECIPIENT	LOW	<span style="color:red">●</span>	DATA	HIGH	<span style="color:green">●</span>	PURPOSE	HIGH	<span style="color:green">●</span>	TRANSFORM	HIGH	<span style="color:green">●</span>	<b>OVERALL OUTPUT SCORE:</b>	<b>0.90462963</b>		<b>OVERALL OUTPUT RATING:</b>	<b>HIGH</b>		<p><b>RESULTS</b></p> <table border="1"> <tr><td>RECIPIENT</td><td>4.28571429</td><td><span style="color:green">●</span></td></tr> <tr><td>DATA</td><td>1.16666667</td><td><span style="color:red">●</span></td></tr> <tr><td>PURPOSE</td><td>5</td><td><span style="color:green">●</span></td></tr> <tr><td>TRANSFORM</td><td>1.75</td><td><span style="color:red">●</span></td></tr> <tr><td>OUTPUT</td><td>4.52314815</td><td><span style="color:green">●</span></td></tr> <tr><td><b>RECOMMENDATION</b></td><td><b>0.79645377</b></td><td><span style="color:orange">●</span></td></tr> </table> <p><b>PLEASE REVIEW LOW SCORING DOMAINS</b></p>			RECIPIENT	4.28571429	<span style="color:green">●</span>	DATA	1.16666667	<span style="color:red">●</span>	PURPOSE	5	<span style="color:green">●</span>	TRANSFORM	1.75	<span style="color:red">●</span>	OUTPUT	4.52314815	<span style="color:green">●</span>	<b>RECOMMENDATION</b>	<b>0.79645377</b>	<span style="color:orange">●</span>			
RECIPIENT	LOW	<span style="color:red">●</span>																																										
DATA	HIGH	<span style="color:green">●</span>																																										
PURPOSE	HIGH	<span style="color:green">●</span>																																										
TRANSFORM	HIGH	<span style="color:green">●</span>																																										
<b>OVERALL OUTPUT SCORE:</b>	<b>0.90462963</b>																																											
<b>OVERALL OUTPUT RATING:</b>	<b>HIGH</b>																																											
RECIPIENT	4.28571429	<span style="color:green">●</span>																																										
DATA	1.16666667	<span style="color:red">●</span>																																										
PURPOSE	5	<span style="color:green">●</span>																																										
TRANSFORM	1.75	<span style="color:red">●</span>																																										
OUTPUT	4.52314815	<span style="color:green">●</span>																																										
<b>RECOMMENDATION</b>	<b>0.79645377</b>	<span style="color:orange">●</span>																																										

The figure illustrates the scoring of each dimension and domain, the visual assessments and recommendation

**ADMINISTRATION:** This worksheet is where the user sets the weights for each domain and dimension, and the thresholds for each domain and the final result ( Figure 27). Note that the weights for the domains under the output heading (appearing in black) should not be modified, as they are set in the **Output** super-domain box (in beige) and automatically reflected.

**Figure 27: The Administration worksheet of the Microsoft Excel workbook developed as an example implementation of the data-sharing framework allowing users to set domain and dimension weights as well as assessment scoring thresholds**

	INITIAL REQUEST					OUTPUT			
	WEIGHT	thresholds (bound by 0 and 1)				WEIGHT	thresholds (bound by 0 and 1)		
		HIGH	MEDIUM	LOW		HIGH	MEDIUM	LOW	
<b>RECIPIENT</b>	5	0.85	0.7	0.25	<b>1</b>	0.85	0.7	0.25	
Trust	2				1				
Security	5				1				
Training	1				1				
Contract	4				1				
<b>DATA</b>	3	0.85	0.7	0.25	<b>4</b>	0.85	0.7	0.25	
Granularity	4				4				
Multiplicity	2				2				
Sensitivity	4				4				
Size	2				2				
<b>PURPOSE</b>	5	0.85	0.7	0.25	<b>4</b>	0.85	0.7	0.25	
Contribution	5				5				
Necessity	4				4				
Effort	2				2				
Impact	3				5				
<b>TRANSFORM</b>	3	0.85	0.7	0.25	<b>4</b>	0.85	0.7	0.25	
Uniqueness	5				5				
Identification Risk	5				5				
Erosion	3				1				
Analytical Effect	3				1				
<b>OUTPUT</b>	5	0.85	0.7	0.25	<b>RESULTS</b>	0.85	0.7	0.25	
Recipient	1								
Data	4								
Purpose	3								
Transform	4								

The remaining tabs are where the calculations are performed for each domain, and are provided to allow the user to view the logic and formulae and encourage exploration of the effects of various changes.

## 12. Future Directions

### 12.1. LEGISLATION

As has been indicated throughout this study, the clash between the right to privacy and the right to health is not a mere perception, but rather a reality. However, there is a lack of clarity around the patchwork of legislation surrounding the two rights. Expansion of the legislative tool to multiple countries and jurisdictions would be beneficial, particularly if reviewed and endorsed by a legal entity (though care must be taken for it to remain in plain language). Tools such as this are a good start to helping the public health practice community determine the legal circumstances and implications of the sharing of personally identifiable information, but as described in Part I, we must move beyond the prescriptive legal approach to one of ethical and professional responsibility. Furthermore, the lack of unified legislation both across and within countries does not benefit individuals when it comes to primary and public health.

As stated in this study, the legislation is, in many cases, out-dated, and must therefore be modernised. This provides a good opportunity to harmonise the legislation not only with the requirements of public health practice, but also across national and international jurisdictions. The European Union has recently initiated renewal efforts of its Data Protection directive, and therefore has an opportunity to pioneer such an approach. However, as stated by the United Kingdom's Right Honourable Kenneth Clarke, Lord Chancellor and Secretary of State for Justice, this must be done within the context of a robust and appropriate paradigm that facilitates data-sharing, not a restrictive and prescriptive one:

*...we must also guard against regulations or reactions...that become obsessed with privacy or data protection without recognising the harm that also results to citizens from failure to share information, as well as from careless stewardship of data. Detailed prescription will not in itself make our citizens safer, or more free, in this complex, modern world. [198]*



Legislation must therefore be grounded in principles that are inclusive of individual and societal rights and benefits, lucid enough to provide clarity and guidance yet flexible enough to allow changes consistent with evolving technologies. Public health must gain the appreciation and understanding of society of its significance to individual and societal health, and its practitioners must adopt ethical standards that foster the public's trust in the what, why and how of the practice. Such changes will take time to implement and will no doubt have to overcome political, legal and philosophical obstacles. However, in the absence of such a concerted approach and with the increasing adoption of electronic health records, both medical and personal, the chasm between privacy and public health practice will only increase.

## **12.2. TRANSFORMATIONS**

Even in the presence of appropriate regulatory frameworks, governance structures and enabling legislation, there may still be scenarios and circumstances in which the right to privacy supersedes a public health activity. In such circumstances, data transformations are a reasonable consideration, provided the concerns raised throughout this study (and presented in the transform domain of the framework) are taken into consideration. This provides fertile ground for the exploration and development of novel transformations.

### **12.2.1. SMART AGGREGATION**

One example of novel transformations is the improvement of traditional aggregation algorithms. As described in Part II, traditional aggregation methods are subject to the modifiable areal unit problem in which areal units are arbitrarily, politically or administratively defined and therefore irrelevant or not well suited to the study at hand. Such aggregations can also blur demographic differences which in turn significantly impacts public health practice. Aggregation has also been shown to have a negative analytical impact, as demonstrated through studies on disease cluster detection.

Two novel approaches were described in this study that can be combined and enhanced to provide what I refer to as "smart aggregation" - or *SMAGGRO* for short. In the contribution to the work on the management of re-identification risk, I created a Canada Grid in which each cell was 1Km by 1Km. Such a grid can be re-created using smaller cell sizes, with each cell containing the summarised attributes of the population it captures. An adjacency matrix as described under "Rethinking Spatial Aggregation" can then be built for the grid cells to include spatial proximity, along with other relevant attributes such as population and socio-demographic factors. Cells can then be aggregated based on pre-defined rules for attribute similarities. This allows the aggregation of areal units small enough (based on the defined cell-size) and similar enough (based on pre-defined attribute requirements) to minimise aggregation effects on analysis. Areal units are then also dynamically aggregated, not based on pre-defined boundaries, thereby overcoming the issues created by the modifiable areal unit problem.

Assigning the cells population counts and demographics based on a granular underlying population would further enhance the aggregation effects, as well as allow for control of uniqueness and identification risk. In the absence of a real granular population, it would be useful to use a synthesised population instead, as was discussed in the multidimensional point transform. This is therefore another area for additional work.

### **12.2.2. SYNTHESISED POPULATIONS**

As described in Part II, a synthesised population for the United States has already been developed and proven useful for agent-based modelling of infectious diseases. The multidimensional point transform demonstrated a novel use for such a population, but such a population's usefulness to public health – particularly in light of privacy-related issues – are tremendous. A synthesised population can be used in the *SMAGGRO* method described above, or to model impacted populations for

environmental health risks, analyse the impact of chemical, biological, radiological and nuclear events, assess changes in demographic factors, synthesise social networks, etc. Provided the synthesised population appropriately mirrors the real population, the possibilities for its use are limited only by one's initiative and creativity.

After much searching and consultation across various academic institutions and public health organisations in Canada, it became apparent that while all those consulted agreed that such a synthetic population would be immensely useful, none had created one. I therefore took the initiative to begin synthesising one, starting with the city of Ottawa as a prototype using the 2006 census of Canada at the smallest released census profile geography - the Dissemination Area (DA). This work is yet to be completed, however the SAS code pursued to date is attached in Appendix F (Volume II). Future work would continue the development of models to create, validate and use synthesised populations, updated with every census release and incorporating projections between census years.

### **12.2.3. THE MULTIDIMENSIONAL POINT TRANSFORM**

As outlined in Chapter 8, and through the suggested interface design, the MPT algorithm can be further refined to allow the user more control over various settings and the flow of the transform. For example, the user can be allowed to set priority levels to the transformable dimensions, thereby allowing control over what dimensions get transformed when and by how much. Minimum thresholds can also be set, in parallel with maximum ones, and further control over anonymisation within the dataset as well as within the underlying population can be integrated. Other future work in this area would also allow the transform to operate across multiple spatial datasets to allow for inclusion of additional geographic and demographic attributes.

The MPT algorithm also requires further testing on its data erosion and analytical effects. The impact on cluster detection, for example, should be assessed and the code

to do so has been created as an addendum to the current code (please see Appendix E in Volume II). Further testing on different-sized cities, as previously mentioned, different population densities and additional dimensions is also required.

### **12.3. GOVERNANCE STRUCTURES & DATA-SHARING**

Complementing the proposed legislative-paradigm shift is the implementation of governance structures within public health practice to facilitate and regulate the sharing of information. One step in this direction is through the development, adoption and implementation of decision-making frameworks such as the one developed and presented in this study. Continued testing and development of the framework across multiple scenarios and contexts will not only be informed by requirements but will also help define them, identify gaps and establish standards.

Standards around thresholds for anonymity and re-identification need to be further explored and implemented. Again, there is no one-size fits all solution, and thresholds will therefore have to be context-specific. What frameworks and governance structures must allow is the inclusion of multiple domains and dimensions in the assessment of context-specificity. It is not sufficient to apply blanket rules to a given health event or scenario, but consideration must also be given to the person or organisation with whom the information is being shared, the purpose for which it is being shared, the contents of the data, any modifications or transforms performed on them and their implications, and the end result and its delivery or dissemination. Future work must therefore test and refine the proposed framework, moulding it to the requirements of the user and providing an intuitive application to facilitate its implementation.

Public health should also take advantage of the tremendous potential of electronic medical and personal health records. As they become the de-facto method of recording and sharing health information between the individual and clinicians, they can also be used to facilitate public health. Where appropriate, patients can be given the ability to

indicate consent for the use of their information in public health practice, making such an approach more practicable and feasible than current consent requirements. Such an electronic medium can then also be used as a feedback loop, providing findings and back to the patient, thereby making these records a means of transparency, accountability, education and promotion. As indicated, surveys suggest that the public would be agreeable to such uses of their personal information. Such control, however, must be tempered by its necessity and assessed within the guidance offered through implemented frameworks - as mentioned in this study, consent is not always beneficial or appropriate for public health practice (e.g. in an outbreak situation, or where it will introduce sufficient bias so as to render the practice uninformative).

#### **12.4. TOOLS**

As demonstrated in the previous chapter, there is much opportunity to develop tools to implement the various novel aspects of this study. The continued development of Web-based technologies provides an ideal medium through which to develop and deliver such tools, and continued work on the secure delivery of applications over the Web must be considered where data are involved (such as with the implementation of a Web-based application of the MPT). Continued refinement of security protocols must therefore also be a part of tool development. Most important, though, is the development of multi-disciplinary tools that do not simply address public health issues in isolation of other issues - both within public health and across other disciplines. In the current study, the proposed solutions span multiple disciplines, including epidemiology, law, mathematics, geomatics and informatics. These disciplines are not to be treated as silos but rather as synergistic fields that inform and influence one another. Just as privacy should not be an afterthought in public health policies, neither should location be an afterthought in anonymisation algorithms. Future developments and applications must therefore be multidisciplinary.

## 13. Review and Concluding Remarks

### 13.1. OBJECTIVES REVISITED & ACCOMPLISHMENTS REVIEWED

In Chapter 1, the objectives of this research were framed within the context of a contribution to the resolution of the public health-privacy debate. Four objectives were identified and addressed throughout this study:

#### 13.1.1. OBJECTIVE 1

The first objective was to review privacy legislation as it pertains to *place* and public health in Canada, the UK, and various other countries around the world

A review of the privacy legislation was completed and an original contribution to a manuscript was published in the *International Journal of Health Geographics*. The review formed the backdrop against which the current study was set. In addition, a prototype tool was developed and presented as a means to summarise the relevant legislation in plain language to public health practitioners and provide the relevant legislative sections for public health events and research.

#### 13.1.2. OBJECTIVE 2

The second objective was to formally collect and synthesise the perspectives and requirements of public health professionals in Canada and the UK on the current issue, with a focus on the role of *place*.

A survey was completed with public health practitioners in Canada and the United Kingdom, and the study was published in *BMC Public Health*. Up until that point, no other study was found directly assessing the perceptions of public health professionals on the implications of privacy to public health, particularly from a location-privacy perspective. The findings of the survey further informed and justified the remainder of the study.

### 13.1.3. OBJECTIVE 3

The third objective was to develop a novel technique to allow spatial public health analysis at a granular level without compromising privacy

Several approaches were explored in the course of the study, and contributions were made to manuscripts on managing re-identification risk and rethinking spatial aggregation. These involved the development of a Canadian spatial grid and adjacency matrices, both of which were original contributions to the studies. The culmination of this work, however, was the development of the novel and flexible multidimensional algorithm named the "MPT" - Multidimensional Point Transform. As demonstrated with a New York County synthesised population, it is a dynamic, adaptive algorithm that addresses many of the recognised deficiencies in already existing techniques. The algorithm has been submitted and accepted for publication in *Methods of Information in Medicine* and is currently available on the journal Website as an electronic publication ahead of print. In addition, a user-interface for the implementation of the MPT was designed and further developments and areas for refinement and improvement proposed.

### 13.1.4. OBJECTIVE 4

The final objective was to develop a conceptual framework to guide public health practice in the appropriate evaluation of the privacy implications of data-sharing with a particular emphasis on location-privacy

This was accomplished through the proposed conceptual framework presented in Chapter 10. A run-through implementation of the framework using a fictitious scenario was described to demonstrate its use using a customisable and therefore fully flexible and context-specific scoring approach. A workbook was developed in Microsoft Excel to demonstrate the implementation, and allow users to experiment with the effects of

domain and dimension scoring and weighting, and threshold settings. A prototype application was also developed and made available.

To facilitate the sharing of findings and the tools as appropriate, a research Website was created and maintained at <http://www.personplacetime.org>. All four objectives have been successfully accomplished, though there is yet much work to be done to facilitate data-sharing – particularly identifiable location-data – in public health practice.

## 13.2. CONCLUDING THOUGHTS

*You cannot make men good by law; and without good men, you cannot have a good society*

**C.S. Lewis  
Mere Christianity**

Because privacy is a broad “catch-word” of sorts capturing subjective and potentially emotionally-charged perspectives, simply uttering the word tends to draw attention and immediate reaction. It is rather interesting – and sometimes amusing – to sit in on meetings, collaborations and conferences within the public health community and simply watch the commotion and discussion caused by simply asking if everyone has considered the privacy implications, or simply, “..and what about privacy issues?” Yet the flurry of opinions quickly offered in response to such questions sometimes lack rationality, prompted instead by a flawed emotional conviction that if a little privacy is good, more must be better; if one does not think so then one must be against privacy [199]. However, as part of the paradigm shift we must undergo, we need to acknowledge the reality that, whether we like it or not, privacy is circumstantially at odds with other rights and privileges. Suggesting the temporary suspension of privacy in order to save lives does not imply that one is against privacy. Unfortunately, however, the prioritisation gradient along which we are willing to suspend our privacy is itself subjective and muddy at best. In the face of immediate, life threatening danger, we are more likely to quickly acquiesce, often with an expectation of imminent results, leading to a mindset of “you can have my data, but only in an emergency”. But how about not-so-obvious threats with not-so-immediate results? What if the use of



identifiable information could lead to policies that would significantly reduce the incidence of a cancer, thereby not only saving thousands of lives but also prolonging others? What if such data could lead to healthier citizens in a healthier world? Is this not the very purpose of public health practice? The foremost epidemiologist of the twentieth century, Sir William Richard Doll, did exactly that. The first to establish a link between lung cancer and tobacco smoking, he was quoted by the Academy of Medical Sciences to have said "much of my research on the effects ionising radiation and the use of oral contraceptives, leave alone smoking, would have been impossible without the facility of obtaining unbiased access to medical records" [200]. His work would not have been possible had he not had access to granular, identifiable data. Although the human condition has not changed much since Sir Doll's publication of his findings in 1950, privacy legislation has, making such work considerably more difficult to conduct today.

The issue lies not with privacy itself, but rather in the fact that as technology has progressed at an incredible pace, privacy issues have come to the forefront and policies and legislation have become more reactive than proactive, trying to keep up and leaving public health trailing far behind. What is required is a sensible and proportionate approach, and an attempt has been made to capture this within this study. Use of personally identifiable information in public health practice must be addressed on a case-by-case basis; privacy for the sake of privacy is grossly inappropriate, and has the grave potential of forsaking the most basic right to life and health in favour of the right to control one's own information. In doing so, it has the potential to not only harm society as a whole, but also those individuals whose very privacy is being protected. Ideally, informative and protective mechanisms and frameworks such as the novel approaches presented in this study would be in place prior to data collection; privacy requirements must be built into the design phase of studies and projects, as opposed to being an after-thought or requirement forced by public, ethical, operational, technological or other demands.

In advocating for such change, one must be diligent in carefully weighing out potential harm with potential benefit; the aim of instituted policies is generally to prevent harmful use, but in doing so, their focus tends to overlook the encouragement and facilitation of beneficial use. Society is so busy implementing restrictive policies to prevent harmful use when in contrast the focus should be on implementing governance structures that not only permit but also facilitate beneficial use of data in public health. One of the issues is that it seems to be much easier to focus on what *cannot* be done as opposed to what *can* in the context of “blanket control” residing with the individual. As put by Fred H. Cate, “...focus on control ignores the extent to which many uses of personal information pose no risk of harm to individuals, while creating significant benefits for data subjects and society more broadly. Laws that facilitate that control, therefore, often create significant costs, without yielding net benefits.” [199] Obvious examples of this are the legal documents (e.g. licence agreements, terms of use, etc.) often signed by customers or clients in exchange for a service or product, but which are rarely actually read [199]. What is needed is a shift in perspective and emphasis, though it should be noted that just as the right to privacy should not be used to justify or legitimise a public health injustice, the right to health should likewise not be abused.

It is interesting to note that we are very much predominantly social beings who crave not only contact with others, but also the sharing of our lives and details with others - to a large degree, we behave in a communitarian fashion, sharing our information rather freely. A simple reference to the explosion of social media is sufficient to demonstrate this. Our lives have become intricately intertwined with one another in a complex social web, and the explosive growth spurt experienced by rapidly-devoured and integrated social media sites such as Facebook, Twitter and Foursquare in their relatively short life spans is evidence of the human hunger for social contact, networking and sharing of personal events and details. Facebook – whose mission is to “give people the power to share and make the world more open and connected” – currently claims to have

over 500 million active users since its February 2004 launch, with over 250 million logging

into the site on any given day. According to the site, the average user has 130 friends, creates 90 pieces of content every month and spends, on average, somewhere in the neighbourhood of 1.5 hours per day on the site [201]. Twitter, launched in July 2006, reported a mind-boggling 14 million new accounts created in the space of just one month spanning February to March, 2011 and about one billion tweets per week [202]. Foursquare, which allows users to share their location, was launched in March 2009 and has managed to amass over 10 million users in just over two years [203]. As much as we might like to feel like we have control over our personal information, we still give it away quite liberally.

It has been said that the future of medicine is not about drugs and procedures, but rather about data [204]. For public health, this future is now. Although implications are extremely complex, requiring us to sift through seemingly conflicting human rights to health, to information, and to privacy, we must nonetheless rise to the occasion, look beyond the politics and controls, and focus instead on the enabling of trusted and ethical professionals whose goal is, indeed, healthy individuals in healthy communities in a healthy world.



**PART V**  
**REFERENCES**



## References

1. Spasoff RA, Harris SS, Thuriaux MC: **A dictionary of epidemiology**. New York: Oxford University Press, Inc., 2001
2. Hoffman RE, Lopez W, Matthews GW, Rothstein MA, Foster KL: **Law in Public Health Practice**. New York: Oxford University Press, 2007
3. Warren SD, Brandeis LD: **The right to privacy**. *Harvard Law Review* (1890), **4(5)**:193-220.
4. Annas GJ: **Book Review: Is privacy the enemy of public health**. *Health Affairs* (1999), **18(4)**:197-198.
5. Bayer R, Colgrove J: **Public health vs. civil liberties**. *Science* (2002), **297**:1811.  
[\[www.sciencemag.org\]](http://www.sciencemag.org)
6. Lohr, Steve. **Google and Microsoft look to change health care**. *The New York Times*. 14-8-2007  
[http://www.nytimes.com/2007/08/14/technology/14healthnet.html?\\_r=2&th=&adxnnl=0&oref=slogin&emc=th&adxnnlx=1187098251-66i/USF/vUDfnFhnsyfbQ&pagewanted=print](http://www.nytimes.com/2007/08/14/technology/14healthnet.html?_r=2&th=&adxnnl=0&oref=slogin&emc=th&adxnnlx=1187098251-66i/USF/vUDfnFhnsyfbQ&pagewanted=print)
7. **Microsoft HealthVault**. Last Updated: 2011  
[\[http://www.microsoft.com/en-us/healthvault/\]](http://www.microsoft.com/en-us/healthvault/)
8. Sweeney L: **k-Anonymity: a model for protecting privacy**. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* (2002), **10(5)**:557-570.
9. Jeffery C, Ozonoff A, Forsberg L, Nuño M, Pagano M: **The cost of obfuscation when reporting locations of cases in syndromic surveillance systems**. *Advances in Disease Surveillance* (2006), **1**:36.
10. Olson KL, Grannis SJ, Mandl KD: **Privacy protection versus cluster detection in spatial epidemiology**. *American Journal of Public Health* (2006), **96(11)**:2002-2008.
11. Curtis AJ, Mills JW, Leitner M: **Spatial Confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina**. *International Journal of Health Geographics* (10-10-2006), **5(44)**.
12. Brownstein JS, Cassa CA, Kohane IS, Mandl KD: **An unsupervised classification method for inferring original case locations from low-resolution disease maps**. *International Journal of Health Geographics* (8-12-2006), **5(56)**.
13. Brownstein JS, Cassa CA, Mandl KD: **No Place to Hide — Reverse Identification of Patients from Published Maps**. *New England Journal of Medicine* (19-10-2006), **355(16)**:1741-1742.  
[\[http://content.nejm.org/cgi/content/extract/355/16/1741\]](http://content.nejm.org/cgi/content/extract/355/16/1741)
14. Barrett G, Cassell JA, Peacock JL, Coleman MP: **National survey of British public's views on use of identifiable medical data by the National Cancer Registry**. *British Medical Journal* (2006), **332**:1068-1072.

15. Boulos MNK, Cai Q, Padgett JA, Rushton G: **Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses.** *Journal of Biomedical Informatics* (2006), **39(2)**:160-170.  
[<http://www.sciencedirect.com/science/article/B6WHD-4GR32TM-1/2/be0cb959aa15839693f49f582633e59b>]
16. Armstrong MP, Rushton G, Zimmerman DL: **Geographically masking health data to preserve confidentiality.** *Statistics in Medicine* (1999), **18**:497-525.  
[<http://www3.interscience.wiley.com/cgi-bin/fulltext/45002090/PDFSTART>]
17. Cassa CA, Grannis SJ, Overhage JM, Mandl KD: **A novel, context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection.** *Advances in Disease Surveillance* (2006), **1**:10.
18. Kwan M-P, Schuurman N: **Introduction: issues of privacy protection and analysis of public health data.** *Cartographica* (2004), **39(2)**.
19. Marchildon GP: **Health Systems in Transition - Canada.** *Report (Vol. 7 No. 3)*: World Health Organization, 2005  
[<http://www.euro.who.int/document/e87954.pdf>]
20. Schoen C, Osborn R, Huynh PT, Doty M, Davis K, Zapert K, Peugh J: **Primary care and health system performance: adults' experiences in five countries.** *Health Affairs* (28-10-2004), **W4**:487.  
[<http://content.healthaffairs.org/cgi/reprint/hlthaff.w4.487v1>]
21. Schoen C, Osborn R, Doty MM, Bishop M, Peugh J, Murukutla N: **Toward higher-performance health systems: adults' health care experiences in seven countries, 2007.** *Health Affairs* (31-10-2007), **26(6)**:w717.  
[<http://content.healthaffairs.org/cgi/reprint/26/6/w717>]
22. Davis K, Schoen C, Schoenbaum SC, Doty MM, Holmgren AL, Kriss JL, Shea KK: **Mirror, mirror on the wall: an international update on the comparative performance of American health care.** *Report (Pub. no. 1027)*: The Commonwealth Fund, May 2007  
[[http://www.commonwealthfund.org/publications/publications\\_show.htm?doc\\_id=482678](http://www.commonwealthfund.org/publications/publications_show.htm?doc_id=482678)]
23. Naylor D, Basrur S, Bergeron MG, Brunham RC, Butler-Jones D, Dafoe G, Ferguson-Paré M, Lussing F, McGeer A, Neufeld KR, Plummer F: **Learning from SARS: renewal of public health in Canada. A report of the National Advisory Committee on SARS and Public Health.** *Report (Publication Number: 1210)*, Ottawa, Ontario, Canada: Health Canada, 2003
24. **What is public health? (Website).** Last Updated: 2011  
[<http://www.whatispublichealth.org/index.html>]
25. Hodge JG, Gostin LO: **Public health practice vs. research. A report for public health practitioners including cases and guidance for making distinctions.** *Report* May 2004
26. **What is public health practice? As defined by the Public Health Practice Council.** Last Updated: 31-1-2007  
[[www.cdc.gov/od/ocphp/PHPCouncil/Docs/Meetings/03\\_15\\_2007/Handouts/Definition%20of%20PH%20Practice.doc](http://www.cdc.gov/od/ocphp/PHPCouncil/Docs/Meetings/03_15_2007/Handouts/Definition%20of%20PH%20Practice.doc)]



27. **A framework for core functions in public health.** Province of British Columbia: Population Health and Wellness, Ministry of Health Services, 2005 [[http://www.health.gov.bc.ca/library/publications/year/2005/core\\_functions.pdf](http://www.health.gov.bc.ca/library/publications/year/2005/core_functions.pdf)]
28. Sherman G, Campione-Piccardo J: **Distinguishing surveillance from research.** *Critical Public Health* (2007), **17(4)**:279-287.
29. Pan American Health Organisation: **Essential Public Health Functions.** *Report (CD42/15)* July 2000 [[http://www.paho.org/english/gov/cd/cd42\\_15-e.pdf](http://www.paho.org/english/gov/cd/cd42_15-e.pdf)]
30. **Public Health Agency of Canada website.** Last Updated: 2011 [<http://www.phac-aspc.gc.ca>]
31. National Public Health Partnership Group: **National Delphi study on public health functions in Australia.** *Report*, Melbourne, Victoria, Australia January 2000 [<http://www.nphp.gov.au/publications/phpractice/delphi-body.pdf>]
32. World Health Organisation: Regional office for the Western Pacific: **Essential public health functions: the role of ministries of health.** *Report (WPR/RC53/10)* July 2002 [<http://www.wpro.who.int/internet/resources.ashx/RCM/RC53-10.pdf>]
33. **10 Essential public health services.** Last Updated: 9-12-2010 [Centers for Disease Control and Prevention website at <http://www.cdc.gov/nphpsp/essentialServices.html>]
34. **The three core public health functions and the essential public health services.** Last Updated: 2011 [Iowa Department of Public Health website at <http://www.idph.state.ia.us/>]
35. The Secretariat: **Public health practice in Australia today.** *Report*: Melbourne (Vic): National Public Health Partnership, 2000 [<http://www.nphp.gov.au/publications/phpractice/phprac.pdf>]
36. Department of Health: **Shifting the balance of power within the NHS: Securing delivery.** *Report (United Kingdom)* July 2001 [[http://www.dh.gov.uk/prod\\_consum\\_dh/groups/dh\\_digitalassets/@dh/@en/documents/digitalasset/dh\\_4076522.pdf](http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4076522.pdf)]
37. **The impact of HIPAA privacy regulations on public health data functions.** Testimony to the National Committee on Vital and Health Statistics, Subcommittee on Privacy and Confidentiality: November 12, 2003
38. Jacquez GM: **Current practices in the spatial analysis of cancer: flies in the ointment.** *International Journal of Health Geographics* (2004), **3**:22-31.
39. Oppong JR: **Data problems in GIS and health.** In *Proceedings of Health and Environment Workshop 4: Health Research Methods and Data*. Turku, Finland (1999)
40. Hodge JG Jr, Brown EF, O'Connell JP: **The HIPAA privacy rule and bioterrorism planning, prevention and response.** *Biosecurity and Bioterrorism: Biodefense, Strategy, Practice, and Science* (2004), **2(2)**:73-80.

41. Pickle LW, Waller LA, Lawson AB: **Current practices in cancer spatial data analysis: a call for guidance.** *International Journal of Health Geographics* (2005), 4:3.
42. Meyerson A, Williams R: **General k-anonymization is hard.** *Report (CMU-CS-03-113)*, Pittsburgh (PA): School of Computer Science, Carnegie Mellon University, 2003
43. Hulton L, Brandon G, McAlister S, Sage J: [www.citystats.org](http://www.citystats.org) - **Better local information. From breastfeeding to badgers - Brighton & Hove's one stop interactive statistics and mapping service.** *Report*, United Kingdom: South East England Public Health Observatory, NHS, October 2004
44. Boulos MNK: **Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom.** *International Journal of Health Geographics* (2004), 3:1.  
[<http://www.ij-healthgeographics.com/content/3/1/1>]
45. Crawford CAG, Young LJ: **A spatial view of the ecological inference problem.** *Chapter 10 In Ecological Inference: New Methodological Strategies* (King G, Rosen O, Tanner MA (Eds)). Cambridge (UK): Cambridge University Press, 2004  
[<http://www.iq.harvard.edu/NewsEvents/Past/EIC/may/gotwayyoung062103.pdf>]
46. Cotton C, Crippen DW, Kapadia F, Morgan A, Murray HN, Ross G: **Ethics roundtable debate: is a physician-patient confidentiality relationship subservient to a greater good?** *Critical Care* (2005), 9:233-237.  
[<http://ccforum.com/content/9/3/233>]
47. **Gordon v Canada (Health).** (February 2008)  
[<http://decisions.fct-cf.gc.ca/en/2008/2008fc258/2008fc258.html>]
48. **Your Guide to Canadian Law: Answers to the most frequently asked questions.** Markham, ON: Fitzhenry and Whiteside Limited, 2006
49. World Health Organisation: **Declaration of Alma Ata.** Alma Ata, USSR (1978)
50. Lengwiler M: **Privacy, justice and equality. The history of privacy legislation and its significance for civil society.** *Report (Discussion Paper Nr. SP IV 2004-503)*: Wissenschaftszentrum Berlin für Sozialforschung gGmbH, 2004
51. Saxena N, MacKinnon MP, Watling J, Willison D, Swinton M: **Understanding Canadians' attitudes and expectations: Citizens' dialogue on privacy and the use of personal information for health research in Canada.** *Report (Research Report P|09)*: Canadian Policy Research Networks Inc., March 2006
52. Cate FH: **Privacy in perspective.** Washington, D.C.: The AEI Press, 2001
53. Westin AF: **Intrusions: Privacy tradeoffs in a free society.** *Public Perspective* (2000), 11(6):8-11.  
[<http://webapps.ropercenter.uconn.edu/ppscan/116%5C116008.pdf>]

54. Cate FH: **The failure of fair information practice principles.** *Chapter 13 In Consumer protection in the age of the "information economy"* (Winn JK (Ed)).: Ashgate Publishing Limited, 2006  
[\[http://books.google.ca/books?id=EVe40UcXGkUC&printsec=frontcover&dq=consumer+protection+in+the+age+of+the+information+economy+2006&source=bl&ots=AXd0vFRzgK&sig=u9aXBbNWf-UkR6uVBiGzuOZn4t0&hl=en&ei=z6usTb68MqmR0QG5rcn5CA&sa=X&oi=book\\_result&ct=result&resnum=1&ved=0CBgQ6AEwAA#v=onepage&q&f=false\]](http://books.google.ca/books?id=EVe40UcXGkUC&printsec=frontcover&dq=consumer+protection+in+the+age+of+the+information+economy+2006&source=bl&ots=AXd0vFRzgK&sig=u9aXBbNWf-UkR6uVBiGzuOZn4t0&hl=en&ei=z6usTb68MqmR0QG5rcn5CA&sa=X&oi=book_result&ct=result&resnum=1&ved=0CBgQ6AEwAA#v=onepage&q&f=false)
55. OECD: **OECD guidelines on the protection of privacy and transborder flows of personal data.**  
[http://www.oecd.org/document/18/0,3343,en\\_2649\\_34255\\_1815186\\_1\\_1\\_1\\_1\\_00.html](http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_1_00.html) (September 1980).
56. **Fair Information Practice Principles.** Last Updated: 25-6-2007  
[\[http://www.ftc.gov/reports/privacy3/fairinfo.shtm\]](http://www.ftc.gov/reports/privacy3/fairinfo.shtm)
57. Asia-Pacific Economic Cooperation Secretariat: **APEC Privacy Framework. Report:** [http://www.apec.org/Groups/Committee-on-Trade-and-Investment/~media/Files/Groups/ECSG/05\\_ecsq\\_privacyframewk.ashx](http://www.apec.org/Groups/Committee-on-Trade-and-Investment/~media/Files/Groups/ECSG/05_ecsq_privacyframewk.ashx), 2005  
[\[http://www.apec.org/Groups/Committee-on-Trade-and-Investment/~media/Files/Groups/ECSG/05\\_ecsq\\_privacyframewk.ashx\]](http://www.apec.org/Groups/Committee-on-Trade-and-Investment/~media/Files/Groups/ECSG/05_ecsq_privacyframewk.ashx)
58. El Emam K, Brown A, AbdelMalik P: **Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk.** *Journal of the American Medical Informatics Association* (10-12-2008), **16(2)**:256-266.  
[\[http://www.jamia.org/cgi/content/full/16/2/256\]](http://www.jamia.org/cgi/content/full/16/2/256)
59. **Generally accepted privacy principles: CPA and CA practitioner version.** (August 2009)  
[\[http://www.cica.ca/service-and-products/privacy/gen-accepted-privacy-principles/item10511.pdf\]](http://www.cica.ca/service-and-products/privacy/gen-accepted-privacy-principles/item10511.pdf)
60. **Personal Information Protection and Electronic Documents Act, S.C. 2000, c. 5 P-8.6.** Canada. (2000).  
[\[http://laws.justice.gc.ca/en/P-8.6/section-\[section-no\].html\]](http://laws.justice.gc.ca/en/P-8.6/section-[section-no].html)
61. ISTPA: **Analysis of privacy principles: making privacy operational.** *Report (Version 2.0):* International Security, Trust and Privacy Alliance, May 2007  
[\[http://www.istpa.org\]](http://www.istpa.org)
62. O'Neill O: **Informed consent and public health.** *Philosophical Transactions of the Royal Society B: Biological Sciences* (29-6-2004), **359(1447)**:1133-1136.  
[\[http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1693386/pdf/15306401.pdf\]](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1693386/pdf/15306401.pdf)
63. Neuffield Council on Bioethics: **Public health: ethical issues.** *Report,* Cambridge, UK: Cambridge Publishers Ltd, November 2007
64. Lowrance WW: **Learning from Experience: Privacy and the Secondary Use of Data in Health Research.** *Report:* The Nuffield Trust, November 2002  
[\[http://www.nuffieldtrust.org.uk/ecomm/files/161202learning.pdf\]](http://www.nuffieldtrust.org.uk/ecomm/files/161202learning.pdf)
65. Wartenberg D, Thompson WD: **Privacy versus public health: the impact of current confidentiality rules.** *American Journal of Public Health* (2010), **100(3)**:407-412.

66. Rubinstein HG: **If I am only for myself, what am I? A communitarian look at the privacy stalemate.** *American Journal of Law and Medicine* (1999), **25**:203-231.
67. Office of the Privacy Commissioner of Canada: **Privacy: annual report to parliament.** *Report (2007-2008)*: Minister of Public Works and Government Services Canada, 2008  
[<http://www.privcom.gc.ca>]
68. Nass SJ, Levit LA, Gostin LO: **Beyond the HIPAA privacy rule: enhancing privacy, improving health through research.**: National Academies Press, 2009  
[<http://www.ncbi.nlm.nih.gov/books/NBK9578/pdf/TOC.pdf>]
69. Curtis AJ, Mills JW, Agustin L, Cockburn M: **Confidentiality risks in fine scale aggregations of health data.** *Computers, Environment and Urban Systems* (2010), **In press**.  
[doi:10.1016/j.compenvurbsys.2010.08.002]
70. AbdelMalik P, Boulos MNK, Jones R: **The perceived impact of location privacy: a web-based survey of public health perspectives and requirements in the UK and Canada.** *BMC Public Health* (2008), **8**:156.  
[<http://www.biomedcentral.com/1471-2458/8/156>]
71. **Canadian Charter of Rights and Freedoms**29-3-1982).  
[<http://laws.justice.gc.ca/en/charter/1.html>]
72. **R. v. Morgentaler**, [1988] 1 S.C.R. 30. Supreme Court of Canada: January 28, 1988  
[<http://scc.lexum.umontreal.ca/en/1988/1988rcs1-30/1988rcs1-30.html>]
73. **Hunter et al. v. Southam Inc.**, [1984] 2 S.C.R. 145. Supreme Court of Canada: September 17, 1984  
[<http://csc.lexum.umontreal.ca/en/1984/1984rcs2-145/1984rcs2-145.html>]
74. **The Privacy Act, R.S.C. 1985, c. P-21.** Canada. (1985).  
[[http://laws.justice.gc.ca/en/P-21/section-\[section-no\].html](http://laws.justice.gc.ca/en/P-21/section-[section-no].html)]
75. **European Communities Act 1972.** Chapter 68. (1972).  
[[http://www.opsi.gov.uk/acts/acts1972/pdf/ukpga\\_19720068\\_en.pdf](http://www.opsi.gov.uk/acts/acts1972/pdf/ukpga_19720068_en.pdf)]
76. **European Judicial Network in civil and commercial matters: Legal Order - England and Wales.** Last Updated: 19-8-2004  
[[http://ec.europa.eu/civiljustice/legal\\_order/legal\\_order\\_eng\\_en.htm](http://ec.europa.eu/civiljustice/legal_order/legal_order_eng_en.htm)]
77. **Universal Declaration of Human Rights.** Res. 217 A (III). (10-12-1948).  
[<http://www.un.org/Overview/rights.html>]
78. **Convention for the Protection of Human Rights and Fundamental Freedoms as amended by Protocol No. 11**(2003).  
[<http://www.echr.coe.int/NR/rdonlyres/D5CC24A7-DC13-4318-B457-5C9014916D7A/0/EnglishAnglais.pdf>]
79. **Charter of Fundamental Rights of the European Union (2000/C 364/01)**2000).  
[<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2000:364:0001:0022:EN:PDF>]

80. **Directive 95/46/EC of the European Parliament and of the council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.** OJL 281. (23-11-1995).  
[\[http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML\]](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML)
81. **Directive 97/66/EC of the European Parliament and of the Council of 15 December 1997 concerning the processing of personal data and the protection of privacy in the telecommunications sector**10-1-1998).  
[\[http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:1998:024:0001:0008:EN:PDF\]](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:1998:024:0001:0008:EN:PDF)
82. **Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)**31-7-2002).  
[\[http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:NOT\]](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:NOT)
83. **Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC**3-5-2006).  
[\[http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006L0024:EN:NOT\]](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006L0024:EN:NOT)
84. **Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data**12-1-2001).  
[\[http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:008:0001:0022:EN:PDF\]](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:008:0001:0022:EN:PDF)
85. **Naomi Campbell and MGN Limited, [2002] EWCA Civ 1373.** Supreme Court of Judicature, Court of Appeal (Civil Division): October 14, 2002  
[\[http://www.bailii.org/ew/cases/EWCA/Civ/2002/1373.html\]](http://www.bailii.org/ew/cases/EWCA/Civ/2002/1373.html)
86. **Health and Social Care Act 2008 (c. 14).** United Kingdom. (2008).  
[\[http://www.opsi.gov.uk/acts/acts2008/ukpga\\_20080014\\_en\\_1\]](http://www.opsi.gov.uk/acts/acts2008/ukpga_20080014_en_1)
87. **Office of Public Sector Information.** Last Updated: 16-6-2009  
[\[http://www.opsi.gov.uk/\]](http://www.opsi.gov.uk/)
88. **Public Health etc. (Scotland) Act 2008 (asp 5).** United Kingdom. (2008).  
[\[http://www.opsi.gov.uk/legislation/scotland/acts2008/asp\\_20080005\\_en\\_1\]](http://www.opsi.gov.uk/legislation/scotland/acts2008/asp_20080005_en_1)
89. **Health and Social Care (Reform) Act (Northern Ireland) 2009 (c. 1).** United Kingdom. (2009).  
[\[http://www.opsi.gov.uk/legislation/northernireland/acts/acts2009/nia\\_20090001\\_en\\_1\]](http://www.opsi.gov.uk/legislation/northernireland/acts/acts2009/nia_20090001_en_1)

90. **Campbell v. MGN Limited, [2004] UKHL 22.** House of Lords Appellate Committee: May 6, 2004  
[\[http://www.bailii.org/uk/cases/UKHL/2004/22.html\]](http://www.bailii.org/uk/cases/UKHL/2004/22.html)
91. **Personal Information Protection Act, S.A. 2003, c. P-6.5.** Alberta, Canada. (2003).  
[\[http://www.canlii.org/en/ab/laws/stat/sa-2003-c-p-6.5/latest\]](http://www.canlii.org/en/ab/laws/stat/sa-2003-c-p-6.5/latest)
92. **Freedom of Information and Protection of Privacy Act, C.C.S.M. c. F-175.** Manitoba, Canada. (1997).  
[\[http://www.canlii.org/mb/laws/sta/f-175/20090324/whole.html\]](http://www.canlii.org/mb/laws/sta/f-175/20090324/whole.html)
93. **R. v. Patrick, [2009] S.C.C. 17.** Supreme Court of Canada: April 9, 2009  
[\[http://scc.lexum.umontreal.ca/en/2009/2009scc17/2009scc17.html\]](http://scc.lexum.umontreal.ca/en/2009/2009scc17/2009scc17.html)
94. **Office of the Privacy Commissioner of Canada.** Last Updated: 2011  
[\[http://www.priv.gc.ca/\]](http://www.priv.gc.ca/)
95. **European Data Protection Supervisor.** Last Updated: 2009  
[\[http://www.edps.europa.eu/EDPSWEB/\]](http://www.edps.europa.eu/EDPSWEB/)
96. **Information Commissioner's Office.** Last Updated: 2009  
[\[http://www.ico.gov.uk/\]](http://www.ico.gov.uk/)
97. Levin A, Nicholson MJ: **Privacy law in the United States, the EU and Canada: the allure of the middle ground.** *University of Ottawa Law and Technology Journal* (2005), **2(2)**:357-395.
98. Esmail, Nadeem. **The black hole that is Canada's medicare.** *Times Colonist.* 14-2-2005  
[http://www.fraserinstitute.org/Commerce.web/article\\_details.aspx?pubID=3526](http://www.fraserinstitute.org/Commerce.web/article_details.aspx?pubID=3526)
99. Irvine B, Ferguson S, Cackett B: **Background briefing: the Canadian health care system.** *Report:* Centre for the New Europe, 2005  
[\[http://www.cne.org/pub\\_pdf/2002\\_08\\_health\\_care\\_in\\_canada.pdf\]](http://www.cne.org/pub_pdf/2002_08_health_care_in_canada.pdf)
100. Turner C, Bishay H, Peng B, Merifield A: **The ALPHA project: an architecture for leveraging public health applications.** *International Journal of Medical Informatics* (2006), **75(10-11)**:741-754.
101. Last JM: **A dictionary of public health.** New York: Oxford University Press, Inc., 2007
102. **National Library of Medicine - Medical Subject Headings.** Last Updated: 2007  
[\[http://www.nlm.nih.gov/cgi/mesh/2007/MB\\\_cgi?mode=&index=21704&field=all&HM=&II=&PA=&form=&input=\]\(http://www.nlm.nih.gov/cgi/mesh/2007/MB\_cgi?mode=&index=21704&field=all&HM=&II=&PA=&form=&input=\)](http://www.nlm.nih.gov/cgi/mesh/2007/MB_cgi?mode=&index=21704&field=all&HM=&II=&PA=&form=&input=)
103. **Privacy, funding doubts shutter Calif. RHIO.** Last Updated: 8-3-2007  
[\[http://www.govhealthit.com/online/news/97855-1.html\]](http://www.govhealthit.com/online/news/97855-1.html)
104. **E-Health Insider: Researchers underline need for access to records.** Last Updated: 11-6-2007  
[\[http://www.e-health-insider.com/news/2766/resarchers\\_underline\\_need\\_for\\_access\\_to\\_records\]](http://www.e-health-insider.com/news/2766/resarchers_underline_need_for_access_to_records)  
 (Retrieved on June 13, 2007)]

105. Munro, Margaret. **Our privacy rules 'block health research': Important studies held back, scientist says.** *The Vancouver Sun*. 20-10-2004
106. **NewScientist.com: Strict data protection may stifle health research.** Last Updated: 17-1-2006  
[<http://www.newscientist.com/article/dn8595.html>]
107. Gutmann MP, Stern PC: **Putting people on the map: protecting confidentiality with linked social-spatial data.** Washington, D.C.: The National Academies Press, 2007
108. **Health and Social Care Act 2001.** c. 15, United Kingdom. (2001).  
[[http://www.opsi.gov.uk/acts/acts2001/ukpga\\_20010015\\_en\\_1](http://www.opsi.gov.uk/acts/acts2001/ukpga_20010015_en_1)]
109. **Data Protection Act 1998.** c. 29, United Kingdom. (1998).  
[<http://www.opsi.gov.uk/ACTS/acts1998/19980029.htm>]
110. **Data protection & medical research.** (2005)
111. Pickle LW, Szczur M, Lewis DR, Stinchcomb DG: **The crossroads of GIS and health information: a workshop on developing a research agenda to improve cancer control.** *International Journal of Health Geographics* (21-11-2006), **5(51)**.  
[<http://www.ij-healthgeographics.com/content/5/1/51>]
112. Kwan Mei-Po, Casas I, Schmitz BC: **Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks?** *Cartographica* (2004), **39(2)**:15-28.
113. **Office of the Privacy Commissioner of Canada. Fact Sheet: The Privacy Act: Not an excuse to promote secrecy.** Last Updated: 2006  
[[http://www.privcom.gc.ca/fs-fi/02\\_05\\_d\\_29\\_e.asp](http://www.privcom.gc.ca/fs-fi/02_05_d_29_e.asp)]
114. **PersonPlaceTime.org.** Last Updated: 2007  
[<http://www.personplacetime.org>]
115. **NHS Health Informatics Community.** Last Updated: 2007  
[<http://www.espace.connectingforhealth.nhs.uk/>]
116. **The Map & Data Exchange (Public Health Agency of Canada).** Last Updated: 2007  
[<https://php-psp.phac-aspc.gc.ca>]
117. Eysenbach G: **Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES).** *Journal of Medical Internet Research* (2004), **6(3)**:e34. doi:10.2196/jmir.6.3.e34.  
[<http://www.jmir.org/2004/3/e34>]
118. Souhami R: **Governance of research that uses identifiable personal data.** *British Medical Journal* (2006), **333**:315-316.
119. Singleton P, Wadsworth M: **Consent for the use of personal medical data in research.** *British Medical Journal* (2006), **333**:255-258.
120. Hewison J, Haines A: **Overcoming barriers to recruitment in health research.** *British Medical Journal* (2006), **333**:300-302.

121. **Canadian Institutes of Health Research.** Last Updated: 2007  
[<http://www.cihr-irsc.gc.ca/e/23019.html>]
122. **Office of the Privacy Commissioner of Canada - Canadians and the privacy landscape.** Last Updated: 2007  
[[http://www.privcom.gc.ca/information/survey/2007/ekos\\_2007\\_02\\_e.asp](http://www.privcom.gc.ca/information/survey/2007/ekos_2007_02_e.asp)]
123. **Convention for the Protection of Human Rights and Fundamental Freedoms.** C.E.T.S No. 005. (4-11-1950).  
[<http://conventions.coe.int/Treaty/en/Treaties/Html/005.htm>]
124. Sims H: **Public confidence in government, and government service delivery.** *Report (Report # P105B):* Canadian Centre for Management Development, 2001  
[[http://www.cspc-efpc.gc.ca/Research/publications/pdfs/HarveySimms\\_e.pdf](http://www.cspc-efpc.gc.ca/Research/publications/pdfs/HarveySimms_e.pdf)]
125. **In government we don't trust.** (1997)  
[<http://www.foreignpolicy.com/Ning/archive/archive/108/ingowwedontrust.pdf>]
126. Green JM, Draper AK, Dowler EA, Fele G, Hagenhoff V, Rusanen M, Rusanen T: **Public understanding of food risks in four European countries: a qualitative study.** *European Journal of Public Health* (2005), **15(5):**523-527.
127. Newton K: **Political support: social capital, civil society and political and economic performance.** *Political Studies* (2006), **54:**846-864.
128. Dalton RJ: **The social transformation of trust in government.** *International Review of Sociology* (2005), **15(1):**133-154.
129. Hudson J: **Institutional trust and subjective well-being across the EU.** *Kyklos* (2006), **59(1):**43-62.
130. Matthews S.A.: **GIS and privacy.** *Report (GIS\_RD\_03-51)*, University Park (PA): Pennsylvania State University, February 2003  
[[http://www.pop.psu.edu/gia-core/pdfs/gis\\_rd\\_03-51.pdf](http://www.pop.psu.edu/gia-core/pdfs/gis_rd_03-51.pdf)]
131. **Every Canadian's Guide To The Law.** Toronto, ON: HarperCollins Publishers Ltd, 2005
132. El Emam K, Buckeridge D, Tamblyn R, Neisa A, Jonker E, Verma A: **The re-identification risk of Canadians from longitudinal demographics.** *BMC Medical Informatics and Decision Making* (2011), **11(46).**  
[<http://dx.doi.org/10.1186/1472-6947-11-46>]
133. **Jerry Ratcliffe's Home Page - Web site. The Modifiable Areal Unit Problem.** Last Updated: 2005  
[<http://www.jratcliffe.net/research/maup.htm>]
134. King G, Rosen O, Tanner MA: **Ecological Inference: New Methodological Strategies.** Cambridge (UK): Cambridge University Press, 2004
135. Ozonoff A, Jeffery C, Manjourides J, White LF, Pagano M: **Effect of spatial resolution on cluster detection: a simulation study.** *International Journal of Health Geographics* (27-11-2007), **6(52).**  
[<http://www.ij-healthgeographics.com/content/6/1/52>]



136. Wiggins L: **Using geographic information systems technology in the collection, analysis, and presentation of cancer registry data: a handbook of basic practices.** Springfield (IL): North American Association of Central Cancer Registries, 2002
137. **University of British Columbia Web Site. Shifting boundaries, shifting results: The modifiable areal unit problem.** Last Updated: 2001  
[\[http://www.geog.ubc.ca/courses/geog516/talks\\_2001/scale\\_maup.html\]](http://www.geog.ubc.ca/courses/geog516/talks_2001/scale_maup.html)
138. Yang T-C: **Modifiable areal unit problem.** *Report (GIS\_RD\_05-65)*, University Park (PA): Pennsylvania State University, February 2005  
[\[http://www.pop.psu.edu/gia-core/pdfs/gis\\_rd\\_05-65.pdf\]](http://www.pop.psu.edu/gia-core/pdfs/gis_rd_05-65.pdf)
139. Marceau DJ: **The scale issue in social and natural sciences.** *Canadian Journal of Remote Sensing* (1999), **25(4)**:347-356.
140. Bivand R: **A review of spatial statistical techniques for location studies.** (August 1998)  
[\[http://www.nhh.no/geo/qib/qib1998/qib98-3/lund.pdf\]](http://www.nhh.no/geo/qib/qib1998/qib98-3/lund.pdf)
141. **Jerry Ratcliffe's Home Page - Web site. The ecological fallacy.** Last Updated: 2005  
[\[http://www.jratcliffe.net/research/ecolfallacy.htm\]](http://www.jratcliffe.net/research/ecolfallacy.htm)
142. Rushton G: **GIS and health information: moving ahead to improve cancer control - Confidentiality restrictions and methods to allow analysis and presentation of local data.** Bethesda (MD): National Cancer Institute
143. Cassa CA, Grannis SJ, Overhage JM, Mandl KD: **A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection.** *Journal of the American Medical Informatics Association* (2006), **13(2)**:160-165.
144. Wieland SS, Cassa CA, Mandl KD, Berger B: **Revealing the spatial distribution of a disease while preserving privacy.** *Proceedings of the National Academy of Sciences of the United States of America* (2008), **105(46)**:17608-17613.
145. Hampton KH, Fitch MK, Allshouse WB, Doherty IA, Gesink DC, Leone PA, Serre ML, Miller WC: **Mapping health data: improved privacy protection with donut method geomasking.** *American Journal of Epidemiology* (2010), **172(9)**:1062-1069.
146. El Emam K, Brown A, AbdelMalik P, Neisa A, Walker M, Bottomley J, Roffey T: **A method for managing re-identification risk from small geographic areas in Canada.** *BMC Medical Informatics and Decision Making* (2010), **10**:18.  
[\[http://www.biomedcentral.com/1472-6947/10/18\]](http://www.biomedcentral.com/1472-6947/10/18)
147. El Emam K, Dankar FK: **Protecting Privacy Using k-Anonymity.** *Journal of the American Medical Informatics Association* (25-6-2008), **15**:627-637.  
[\[http://www.jamia.org/cgi/content/abstract/15/5/627\]](http://www.jamia.org/cgi/content/abstract/15/5/627)
148. Dankar FK, El Emam K: **A method for evaluating marketer re-identification risk.** *Proceedings of the 2010 EDBT/ICDT Workshops* (2010), **28**.

149. Aggarwal CC, Yu PS: **A general survey of privacy-preserving data mining models and algorithms**. Chapter 2 In *Privacy-preserving data mining: models and algorithms* (Aggarwal CC, Yu PS (Eds)). New York: Springer Science+Business Media, 2008
150. Cassa CA, Wieland SC, Mandl KD: **Re-identification of home addresses from spatial locations anonymized by Gaussian skew**. *International Journal of Health Geographics* (2008), **7(45)**.  
[<http://www.ij-healthgeographics.com/content/7/1/45>]
151. Aggarwal CC, Yu PS: **Privacy-preserving data mining: models and algorithms**. New York: Springer Science+Business Media, 2008
152. Ratcliffe JH, McCullagh MJ: **Hotbeds of crime and the search for spatial accuracy**. *Journal of Geographical Systems* (1999), **1**:385-398.
153. AbdelMalik P, Kamel Boulos MN: **Multidimensional point transform for public health practice**. *Methods of Information in Medicine* (2011)(In press).  
[<http://dx.doi.org/10.3414/ME11-01-0001>]
154. Boulos MNK, Curtis AJ, AbdelMalik P: **Musings on privacy issues in health research involving disaggregate geographic data about individuals**. *International Journal of Health Geographics* (20-7-2009), **8**:46.  
[<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2716332/pdf/1476-072X-8-46.pdf>]
155. De Moor GJE, Claerhout B, De Meyer F: **Privacy enhancing techniques**. *Methods of Information in Medicine* (2003), **42**:148-153.
156. Robling MR, Hood K, Houston H, Pill R, Fay J, Evans HM: **Public attitudes towards the use of primary care patient record data in medical research without consent: a qualitative study**. *Journal of Medical Ethics* (2004), **30**:104-109.  
[<http://jme.bmj.com/cgi/content/abstract/30/1/104>]
157. Jones C: **The utilitarian argument for medical confidentiality: a pilot study of patients' views**. *Journal of Medical Ethics* (2003), **29(6)**:348-352.  
[<http://jme.bmj.com/cgi/content/abstract/29/6/348>]
158. Onsrud HJ, Johnson JP, Lopez XR: **Protecting personal privacy in using geographic information systems**. *Photogrammetric Engineering & Remote Sensing* (1994), **60(9)**:1083-1095.
159. Domingo-Ferrer J, Torra V: **A critique of k-anonymity and some of its enhancements**. In *IEEE*. (2008)
160. Dalenius T: **Finding a needle in a haystack or identifying anonymous census records**. *Journal of Official Statistics* (1986), **2(3)**:329-336.
161. Claerhout B, De Moor GJE: **Privacy protection for HealthGrid applications**. *Methods of Information in Medicine* (2005), **44**:140-143.
162. **Models of Infectious Disease Agent Study: Synthesized data**. Last Updated: 2009  
[<https://www.epimodels.org/midas/pubsyntdata1.do>]

163. **Statistics Canada: 2006 Census release topics - Canada.** Last Updated: 17-12-2009  
[\[http://www12.statcan.gc.ca/census-recensement/2006/rt-td/index-eng.cfm\]](http://www12.statcan.gc.ca/census-recensement/2006/rt-td/index-eng.cfm)
164. **Office for National Statistics: Census 2001 - United Kingdom.** Last Updated: 1-11-2005  
[\[http://www.statistics.gov.uk/census2001/census2001.asp\]](http://www.statistics.gov.uk/census2001/census2001.asp)
165. **2005 American Community Survey: Age and sex population.** Last Updated: 2008  
[\[http://factfinder.census.gov/servlet/STTable?\\_bm=y&-geo\\_id=01000US&-qr\\_name=ACS\\_2005\\_EST\\_G00\\_S0101&-ds\\_name=ACS\\_2005\\_EST\\_G00\\_&-lang=en&-redoLog=false\]](http://factfinder.census.gov/servlet/STTable?_bm=y&-geo_id=01000US&-qr_name=ACS_2005_EST_G00_S0101&-ds_name=ACS_2005_EST_G00_&-lang=en&-redoLog=false)
166. **RAMDisk software.** Last Updated: 2010  
[\[http://memory.dataram.com/products-and-services/software/ramdisk\]](http://memory.dataram.com/products-and-services/software/ramdisk)
167. Zimmerman DL, Pavlik C: **Quantifying the Effects of Mask Metadata Disclosure and Multiple Releases on the Confidentiality of Geographically Masked Health Data.** *Geographical Analysis* (2008), **40(1)**:52-76.  
[\[http://www3.interscience.wiley.com/journal/119390400/abstract?CRETRY=1&SRETRY=0\]](http://www3.interscience.wiley.com/journal/119390400/abstract?CRETRY=1&SRETRY=0)
168. El Emam K, Kamal Dankar F, Issa R, Jonker E, Amyot D, Cogo E, Corriveau J-P, Walker M, Chowdhury S, Vaillancourt R, Roffey T, Bottomley J: **A globally optimal k-anonymity method for the de-identification of health data.** *Journal of the American Medical Informatics Association* (2009), **16(5)**:670-682.  
[\[http://dx.doi.org/10.1197/jamia.M3144\]](http://dx.doi.org/10.1197/jamia.M3144)
169. Wheaton WD, Cajka JC, Chasteen BM, Wagener DK, Cooley PC, Ganapathi L, Roberts DJ, Allpress JL: **Synthesized population databases: a US geospatial database for agent-based models.** *Report (RTI Press publication No. MR-0010-0905)*, Research Triangle Park, NC: RTI International, May 2009
170. Black N: **Secondary use of personal data for health and health services research: why identifiable data are essential.** *Journal of Health Services Research and Policy* (2003), **8(Supplement 1)**:S1:36-S1:40.  
[\[http://www.ncbi.nlm.nih.gov/pubmed/12869337\]](http://www.ncbi.nlm.nih.gov/pubmed/12869337)
171. Rotter JB: **A new scale for the measurement of interpersonal trust.** *Journal of Personality* (1967), **35(4)**:651-665.
172. Liu C, Marchewka JT, Lu J, Yu C-S: **Beyond concren - a privacy-trust-behavioral intention model of electronic commerce.** *Information & Management* (2005), **42**:289-304.
173. Reilly P, Cullen R: **Information privacy and trust in government: a citizen-based perspective.** *Report* January 2006
174. **The Direct Project.** Last Updated: 2011  
[\[http://directproject.org/\]](http://directproject.org/)
175. **The Direct Project Overview.** Last Updated: 11-10-2010  
[\[http://wiki.directproject.org/file/view/DirectProjectOverview.pdf\]](http://wiki.directproject.org/file/view/DirectProjectOverview.pdf)
176. ISTPA: **Privacy management reference model.** *Report (Version 2.0)*: International Security, Trust & Privacy Alliance, 2009  
[\[http://www.istpa.org\]](http://www.istpa.org)

177. **International Organization for Standardization.** Last Updated: 2011  
[<http://www.iso.org/iso/home.htm>]
178. **ISO 27799:2008: Health informatics -- information security management in health using ISO/IEC 27002.** Last Updated: 2011  
[[http://www.iso.org/iso/catalogue\\_detail?csnumber=41298](http://www.iso.org/iso/catalogue_detail?csnumber=41298)]
179. DHS Privacy Office: **Guide to implementing privacy. Report (Version 1.0)**  
June 2010  
[<http://www.dhs.gov/privacy>]
180. **Organisation for Economic Co-operation and Development.** Last Updated: 2011  
[<http://www.oecd.org>]
181. Department of Health: **Information security management: NHS code of practice.** *Report (No. 280361):* United Kingdom, April 2007
182. Canadian Institutes of Health Research: **CIHR best practices for protecting privacy in health research.**: Public Works and Government Services Canada, 2005
183. **NHS - Information Governance.** Last Updated: 2011  
[<http://www.connectingforhealth.nhs.uk/systemsandservices/infogov>]
184. **NHS Information Governance Toolkit.** Last Updated: 2011  
[<https://www.igt.connectingforhealth.nhs.uk/Home.aspx?tk=407642394438910&cb=e9763189-dcf9-43f6-a8b8-205c25fd38c5&Inv=7&clnav=YES>]
185. **Privacy Act 1988. Act No. 119 of 1988 as amended up to Act No. 170 of 2006. Government of Australia.** Act No. 199 of 1988, as amended up to Act No. 170 of 2006. (1988).[
186. **Personal Data Act (523/1999).** Finland. (1999).  
[<http://www.tietosuoja.fi/uploads/hopxtvf.HTM>]
187. **Personal Data Act (1998:204).** Sweden. (29-4-1998).  
[<http://www.datainspektionen.se/pdf/ovrigt/pul-eng.pdf>]
188. **Guidance for the Classification Marking of NHS Information.** Last Updated: 2009  
[[https://www.igt.connectingforhealth.nhs.uk/KnowledgeBaseNew/DH\\_NHS%20G%20-%20Info%20Classifications.pdf](https://www.igt.connectingforhealth.nhs.uk/KnowledgeBaseNew/DH_NHS%20G%20-%20Info%20Classifications.pdf)]
189. North East London NHS: **Caldicott and safe havens policy.** (July 2009),  
**Policy No: IT002.**
190. Mc Cullagh K: **Data sensitivity: resolving the conundrum.** In *British & Irish Law, Education and Tehcnology Association - 2007 Annual Conference.* Hertfordshire (2007)  
[<http://www.bileta.ac.uk/Document%20Library/1/Data%20Sensitivity%20-%20resolving%20the%20conundrum.pdf>]
191. **Should we care what we share?** Last Updated: 23-4-2011  
[[http://news.bbc.co.uk/2/hi/programmes/click\\_online/9465244.stm](http://news.bbc.co.uk/2/hi/programmes/click_online/9465244.stm)]
192. Kickbusch I: **Global health diplomacy: how foreign policy can influence health.** *British Medical Journal* (2011), **342**:d3154.

193. Fayerman, Pamela. **Should smokers be precluded from jobs in the health sector?** *The Vancouver Sun*. 28-2-2011  
<http://communities.canada.com/vancouver/blogs/medicinematters/archive/2011/02/28/looming-on-the-horizon-smokers-need-not-apply-for-these-jobs.aspx>
194. CIHI: **'Best Practice' guidelines for managing the disclosure of de-identified health information.** *Report*: Canadian Institute for Health Information, October 2010
195. **Association of Schools of Public Health website.** Last Updated: 2010  
<http://www.asph.org/>
196. El Emam K, Mercer J, Moreau K, Grava-Gubins I, Buckeridge D, Jonker E: **Physician privacy concerns when disclosing patient data for public health purposes during a pandemic influenza outbreak.** *BMC Public Health* (2011), **11(454)**.  
[doi:10.1186/1471-2458-11-454]
197. **pipes from Yahoo!** Last Updated: 2011  
<http://pipes.yahoo.com/pipes/>
198. **Lord Chancellor and Secretary of State for Justice Rt.Hon Kenneth Clarke MP: Data Protection (transcript).** (May 2011)  
<http://amberhawk.typepad.com/files/kenneth-clarke-final-speech-as-delivered-1.pdf>
199. Cate FH: **Principles for protecting privacy.** *Carto Journal* (2002), **22(1):33-57**.
200. The Academy of Medical Sciences: **Personal data for public good: using health information in medical research.** *Report* January 2006
201. **Facebook Statistics.** Last Updated: 2011  
<http://www.facebook.com/press/info.php?statistics>
202. **Twitter Blog #numbers.** Last Updated: 2011  
<http://blog.twitter.com/2011/03/numbers.html>
203. **Foursquare hits 10 million user milestone.** Last Updated: 20-6-2011  
<http://aboutfoursquare.com/>
204. **The coming revolution: data-driven, patient-centered health care (TED2010 Pioneer Portfolio).** (2010)  
<http://vimeo.com/9426281>



**MULTIDIMENSIONAL EPIDEMIOLOGICAL TRANSFORMATIONS:  
ADDRESSING LOCATION-PRIVACY IN PUBLIC HEALTH PRACTICE**

**VOLUME II**

**PHILIP SAMI ONSY ABDELMALIK**  
MHS<sub>c</sub> (Community Health & Epidemiology)

**Thesis Submission  
in partial fulfillment for the Degree of**

**DOCTOR OF PHILOSOPHY (PHD)**

**Submitted to the University of Plymouth  
Faculty of Health and Education**

**November 2011**

**© Philip AbdelMalik. All rights reserved.**





## TABLE OF CONTENTS

### VOLUME II

<b>PART VI APPENDICES .....</b>	<b>1</b>
<b>A. Web Survey Logistics &amp; Business Specifications .....</b>	<b>3</b>
A.1. Plain Language Statement .....	3
A.2. Functional Requirements for Web Implementation .....	5
A.3. Target Population .....	6
A.4. Sample Email Invitation .....	8
A.5. Consent to Participate .....	10
A.6. Data Storage and Handling Policies .....	12
A.7. Glossary .....	13
A.8. Application Screenshots .....	15
<b>B. Survey Questionnaires .....</b>	<b>17</b>
B.1. Public Health Professional Questionnaire CANADA .....	17
B.2. L'impact de la protection des renseignements personnels sur la pratique en santé publique CANADA .....	32
B.3. Public Health Professional Questionnaire UNITED KINGDOM .....	54
<b>C. Full survey findings .....</b>	<b>69</b>
<b>D. Multidimensional Point Transform algorithm: SAS code .....</b>	<b>113</b>
<b>E. Example Code Modification to Allow for Cluster Insertion .....</b>	<b>135</b>
<b>F. (Partial) Synthetic Population Generation: SAS code .....</b>	<b>140</b>
<b>G. Publications &amp; Reports - Attached as published .....</b>	<b>172</b>

### LIST OF FIGURES

Figure A.8.1: Login screen for the Web-based survey.....	15
Figure A.8.2: Screenshot of the consent screen.....	15
Figure A.8.3: Example screen from Web-based survey showing the first question.....	16
Figure A.8.4: Administrator screen.....	16

## **DECLARATION**

The author declares that at no time during registration for this research degree of Doctor of Philosophy was he registered for any other University award.

This study was partially financed and supported by the Public Health Agency of Canada.

During the course of this research, the author was invited to participate in collaborative work with other groups and authors. Research arising from such collaborations was informed by the author's current degree work and is clearly identified as such throughout the thesis. Similarly, original work and publications arising directly from this work are identified and were conceived and completed in their entirety by the author.

Relevant publications, presentations, reports, awards, grants and activities are identified in Volume I. Please note that workshops, seminars and meetings attended by the author in which he was a participant and not a presenter are not listed.

In addition, a Website was created for the study and can be found at <http://www.personplacetime.org>. Prototype tools developed through the course of this work can be found through this site and are referenced and described in Volume I where relevant.

The University of Plymouth is hereby granted permission to allow the thesis to be copied and or distributed, in whole or in part, for academic purposes, subject to the acknowledgement of the author.

**PART VI**  
**APPENDICES**



## **A. Web Survey Logistics & Business Specifications**

### **A.1. PLAIN LANGUAGE STATEMENT**

In public health research, the location of any specific health event is of paramount importance to investigating the event and trying to identify intervention or prevention strategies. However, once such a location is provided, it becomes quite easy to identify the specific individual to which it refers. This, in turn, violates privacy laws, which, of course, is unacceptable. Therefore, in public health research, events are either grouped (aggregated) and investigated within larger areas that prevent the identification of individuals, or access to the information and its analysis at the finer level of detail is simply restricted or prohibited. In either case, this hinders the research, and provides less-than-optimal information for improving public health systems.

The aim of this study is to improve the capacity of public health research around the globe by providing an innovative and accurate method for the analysis of real health data at the individual level - particularly in space and time (i.e. spatial-temporal) - while simultaneously respecting the privacy and confidentiality of the individual. To assess the importance of this study, though, a survey will first be conducted to ask public health professionals whether or not they do consider this to be an issue, and therefore whether or not such a study would be useful and valuable to them.

The survey will be divided into eight short sections to better categorize and organize the type of information being captured, and should take no more than 20 minutes to complete. Respondents may skip any question(s) they would rather not answer, though they are reminded that this compromises the quality of the research. The survey will utilise skip logic – that is, logic that allows questions posed to be dependent on response received – wherever possible to streamline the survey and shorten the respondents' participation time as much as possible. Question types will vary, but include single-choice, multiple-choice, scale and free-form response questions.

Initial contact with potential participants will be through direct email invitation to specific public health professionals. Recipients of this email will also be asked to forward the invite to any other public health professionals they may deem appropriate within the larger target population. The quality of the research will be maximised by ensuring that each participant only completes the survey once, and that only public health professionals participate. Participation is completely voluntary, and each participant will have to give their consent to participate on the first page of the survey.

The responses received from this survey will help inform the direction of the research. Participants will not otherwise have any direct role in the design and development of the research itself.

## **A.2. FUNCTIONAL REQUIREMENTS FOR WEB IMPLEMENTATION**

The following requirements were specifically identified to the ALPHA group prior to the commencement of the Web implementation:

- Will be administered externally to the public health communities in Canada and the UK
- Must incorporate skip logic: ability to ask different questions based on responses received (see process flow)
- Must require the participant to enter a unique code in order to complete the survey. This ensures that only qualified respondents complete the survey. Participants can acquire a unique code by directly contacting Philip AbdelMalik
- Must allow participants to complete the survey over multiple sessions
- All questions are optional in terms of response requirements
- Must allow for “pop-up” boxes with definitions when italicised words are clicked (a glossary of sorts)
- Data collected from the response must be downloadable to a local computer as a comma delimited file for analysis
- Question types:
  - o Multiple choice, single response
  - o Multiple choice, multiple responses
  - o Yes / No
  - o Rating scale
  - o Matrix / cross tabular
  - o Free text

### A.3. TARGET POPULATION

Staff of the following institutions and organisations were invited to participate in the Web-based survey:

#### Canada

- Public Health Agency of Canada
  - o Office of Public Health Practice
  - o GIS Infrastructure
  - o Field Epidemiology Program
  - o Skills Enhancement
  - o Information Sharing Practices
  - o Public Health Law
  - o Centre for Emergency Preparedness and Response
  - o Emergency Operations Centre
  - o Foodborne, Waterborne and Zoonotic Infectious Diseases (Guelph)
  - o Laboratory for Foodborne Zoonoses (Québec)
  - o Pandemic Preparedness Secretariat
  - o Tuberculosis Prevention and Control
  - o FluWatch
  
- Health Canada
  - o First Nations & Inuit Health Branch
  - o Communicable Disease Division
  - o Healthy Environments and Consumer Safety Branch
  - o Access to Information and Privacy
  
- GIS Infrastructure clients on the Map and Data Exchange (n~300)
- Public Health Sciences staff (University of Toronto; University of Ottawa)
- Grey Bruce Health Unit, Ontario



- Nova Scotia Organized Breast Cancer Screening Program
- New Brunswick Lung Association

### The United Kingdom

- City University GIS Masters Program  
<http://www.soi.city.ac.uk/pgcourses/gis/index.html>
- University of Sheffield Public Health GIS unit (Centre of Excellence)  
<http://www.shef.ac.uk/scharr/sections/ph/research/gis>
- Imperial College Environmental Epidemiology and Small Area Health Statistics Unit  
<http://www1.imperial.ac.uk/medicine/about/divisions/ephpc/eph/projects/eresh/>
- NLH/NHS Informatics UK Health GIS Special Interest Group (SIG)  
<http://www.informatics.nhs.uk/groups/group3/index.html>
- UK AGI Health SIG  
<http://www.agi.org.uk/bfora/user/systems/sig/view.asp?sig=278&arg=1>

#### A.4. SAMPLE EMAIL INVITATION

##### Canadian E-Mail Invitation

Dear NAME,

**Re: The Impact of Privacy on Public Health Practice – Survey**

You have been identified as a Public Health professional in Canada, either through the Web, mailing lists, personal contacts, referrals, or word of mouth. As a public health professional, your input and opinions on the impact of privacy on public health practice are critical to shaping the current research project, being conducted simultaneously in Canada and the United Kingdom. The target audience for the survey is all public health professionals (including directors, consultants, analysts, researchers, strategic staff, managers, epidemiologists, etc.) in the two countries, and your contribution may help significantly improve the ways in which public health is researched and applied.

**Click [here](http://www.personplacetime.org) to complete the survey, or visit <http://www.personplacetime.org> for more information. Your Unique Access Code is XXXXXXXXX.**

Privacy and confidentiality issues have repeatedly been identified as potential obstacles to public health research in many parts of the world. However, there is limited literature on the extent of their impact, sensitivity around further research, and a resultant lack of a satisfactory solution. This survey is part of a PhD research project aimed at investigating, developing, and evaluating novel transformations on the fundamental building blocks of the epidemiological triad – person, place and time – in order to promote public health analysis at a granular level whilst protecting individual confidentiality.

The objectives of the survey, which will take about 20 minutes to complete, are as follows:

1. To assess the impact of privacy and confidentiality on public health research in Canada and the UK
2. To assess the usefulness of the proposed project to public health in Canada and the UK
3. To allow public health professionals in Canada and the UK to volunteer to pilot and evaluate the end result

The results of this survey will be critical to guiding further research and the development of valid transformations. A document containing study details and a glossary of terms used throughout the survey is also attached as a separate file.

All information collected will be completely anonymous (i.e. your access code will be disassociated from the results, unless you specify otherwise. Please note that as a survey sponsored by the Office of Public Health Practice, Public Health Agency of Canada, your personal information will be protected according to the Access to Information Act and the Privacy Act. For more details on survey privacy and confidentiality policies, please visit <http://www.personplacetime.org/SurveyPrivacy>

**Please pass this email on to any other public health personnel you feel appropriate for participation. Please be sure to have them contact me for their unique access code.**

Thank you again for your assistance with this important research.

## **A.5. CONSENT TO PARTICIPATE**

Invited participants who agreed to complete the survey were presented with a consent screen prior to beginning. Consent was required in order to proceed; invitees who did not consent were thanked for their consideration and logged out of the survey. The text of the consent screen appears below:

### **CONSENT TO PARTICIPATE**

By clicking below and completing this survey, you are giving consent to participate in this PhD research project, entitled “Multidimensional Epidemiological Transformations”. All questions in the survey are voluntary, and you may choose to skip any that you do not wish to answer. Please keep in mind, however, that this compromises the quality of the research. You may terminate the survey at any time, and have the option of completing it over multiple sessions, at your convenience. To minimise error, please ensure that you respond to the survey once, and once only. All responses will remain anonymous, and can in no way be linked back to you (your login information will be disassociated from your responses), unless you specifically indicate otherwise in the last section of the survey. Similarly, your provision of your contact information at the end of the survey is completely voluntary, and will only be used for the items you specify (e.g. follow-up, networking and future participation). Participating in this survey poses no risks or benefits to you, other than the time spent completing the survey (about 30 minutes); rather, gathered responses will benefit the scientific community as a whole, and have the potential to improve public health practice wherever issues of privacy are of concern.

Please note that any intellectual property rights in your responses will vest in, and remain the property of Her Majesty the Queen in Right of Canada, and as such, may also be made available to and used by the Public Health Agency of Canada to improve their operations and service provision to the public health community. As an Office of Public Health Practice, Public Health Agency of Canada survey, all data collected, including your personal information, will be protected according to the Access to Information Act (R.S., 1985, c. A-1) and the Privacy Act in Canada (R.S., 1985, c. P-21), as well as the confidentiality, privacy and data protection laws of the UK.

For more details on data storage and handling policies, as well as contact information, please click *here*. More details on the research and the survey itself can be found on the research website at *www.peopleplacetime.org*.

If you do not consent, thank you for your time.

This research has been reviewed and approved by:

- The Research Ethics Board, Office of the Chief Scientist, Health Canada, Government of Canada; Approval #2006-0033
- The Southwest Multi-centre Research Ethics Committee, National Health Service, United Kingdom; Approval #06/MRE06/67
- The Access to Information & Privacy Division, Corporate Services Branch, Health Canada
- The Public Opinion Research & Evaluation Division, Communications, Marketing & Consultations Directorate, Health Canada

I consent to participating in this research

**NOTE: Review by the above does *NOT* indicate their endorsement of the research**

## **A.6. DATA STORAGE AND HANDLING POLICIES**

Survey participants were informed of the policies around the storage and handling of their information and responses through a Web-page integrated in the survey. The text on this page appeared as follows:

### **Data Storage and Handling Policies**

All responses received from the survey will be stored on a secure, password-protected drive within the Public Health Agency of Canada. This drive sits on a secure Agency server, and can only be accessed through the Agency's secure intranet by the principal investigator, and authorised IT personnel of the Public Health Agency of Canada, should the need arise. Since the data will be stored on a Public Health Agency of Canada computer, all personal information will be protected according to the Access to Information Act, as well as the Privacy Act of Canada.

Identifiable data will also be shared with the PhD supervisors in the United Kingdom. This will be done via data postings on the Public Health Agency of Canada's secure *Public Health Portal*. This site can only be accessed via a user name and password, and the data will be shared through a secure folder accessible only to the supervisors, the principal investigator, and administrators of the portal.

The survey will be administered for a total of three to four months, ending on January 31, 2007 (the duration of administration depends on when it can be commenced, and is dependant on ethics approval timelines). All data will be retained by the Government of Canada, and archived upon completion of the research (expected April 1, 2011 or sooner).

## A.7. GLOSSARY

The following terms were defined and included as part of the Web survey's business specifications.

### **Aggregate Data**

These are data that have been “combined and summarised” such that no individual cases can be identified; usually this is the result of manipulated and/or analysed data.

This is the opposite of “granular data”.

### **Epidemiological Triad**

This triad consists of person, time and place, and forms the fundamental elements of an epidemiological study

### **Granular Data**

These are data on an individual, case-by-case level; usually the original data in their raw form, and contain personally identifiable data (defined below). This is the opposite of “aggregate data”.

### **Personally Identifiable Data (PID)**

**CANADA:** This refers to any information that can be used to identify an individual, and is the focus of this research. Canada's *Personal Information Protection and Electronic Documents Act* (PIPEDA) defines “personal information” as **any** information about an identifiable individual – or, more simply stated, any information that can be used to identify an individual.

**UK:** This refers to any information that can be used to identify an individual, and is the focus of this research. In the UK, Chapter 20, Part I, Section I(I) of *The Data Protection*

*Act, 1998*, defines "personal data" as data which relate to a living individual who can be identified either from those data alone, or in combination with other information that is or may come into the possession of the data controller. The *Manual for Caldicott Guardians* was used to identify such data in this survey.

### **Public Health Practice**

The term "Public Health Practice" is used throughout this research to capture the range of activities required for effective provision of health services and the resulting improvement of the population's health. These activities include surveillance, research and analysis, strategic decision-making, program implementation, etc.

### **Relational Integrity**

This refers to the relationships between the components of the epidemiological triad: person (within the health data; e.g. where one case is relative to another); place (between the health data and their location; e.g. where cases are relative to a school or environmental feature); and time (between the health data and time; e.g. what season or time of year a case occurred in).

### **Software Agent**

This is an application (software) with built-in artificial intelligence, which would perform analyses on personally identifiable data and return only the aggregated, and therefore anonymised, results. In this scenario, you would never see the actual data, but would be able to perform analyses on them through these automated applications.

### **Transformation**

In the context of this research, the term "transformation" is used in its mathematical sense to refer to a geometric or spatial alteration of the original data – in this case, to anonymise patient or case identity. Standard geometric transformations include rotation, translation, reflection, etc.



## A.8. APPLICATION SCREENSHOTS

Figure A.8.1: Login screen for the Web-based survey

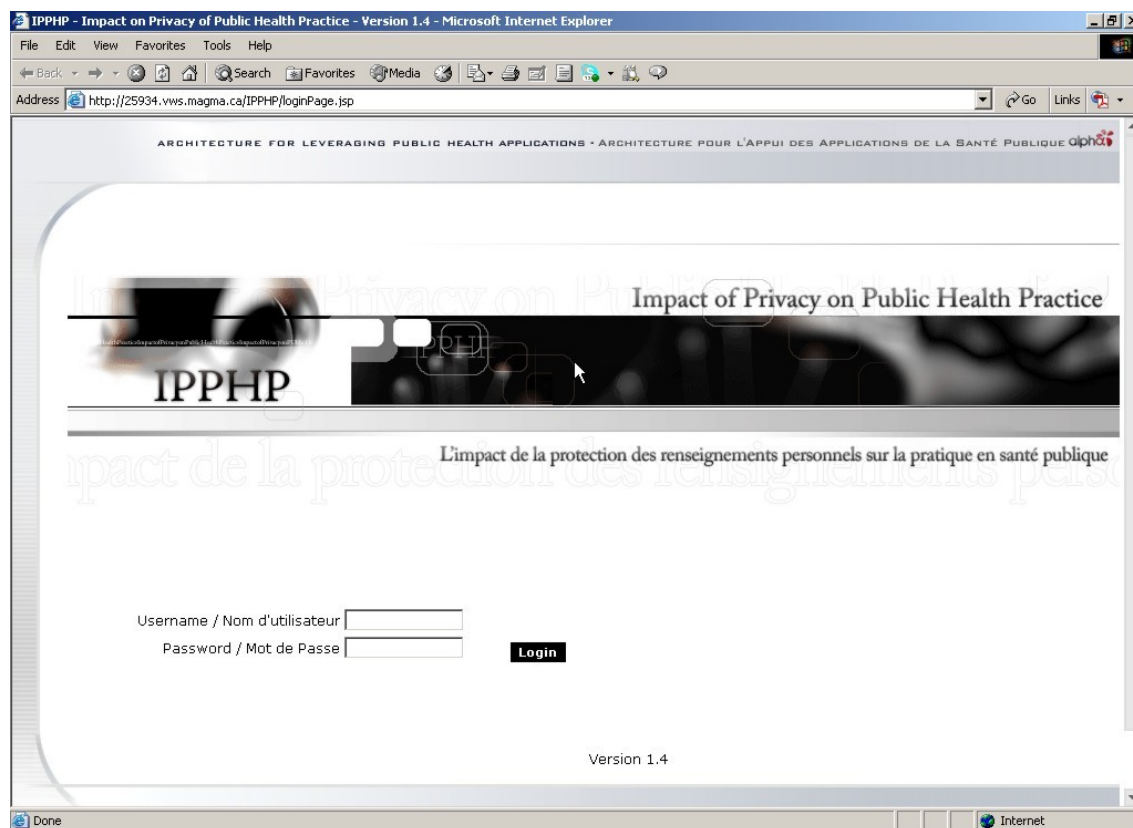
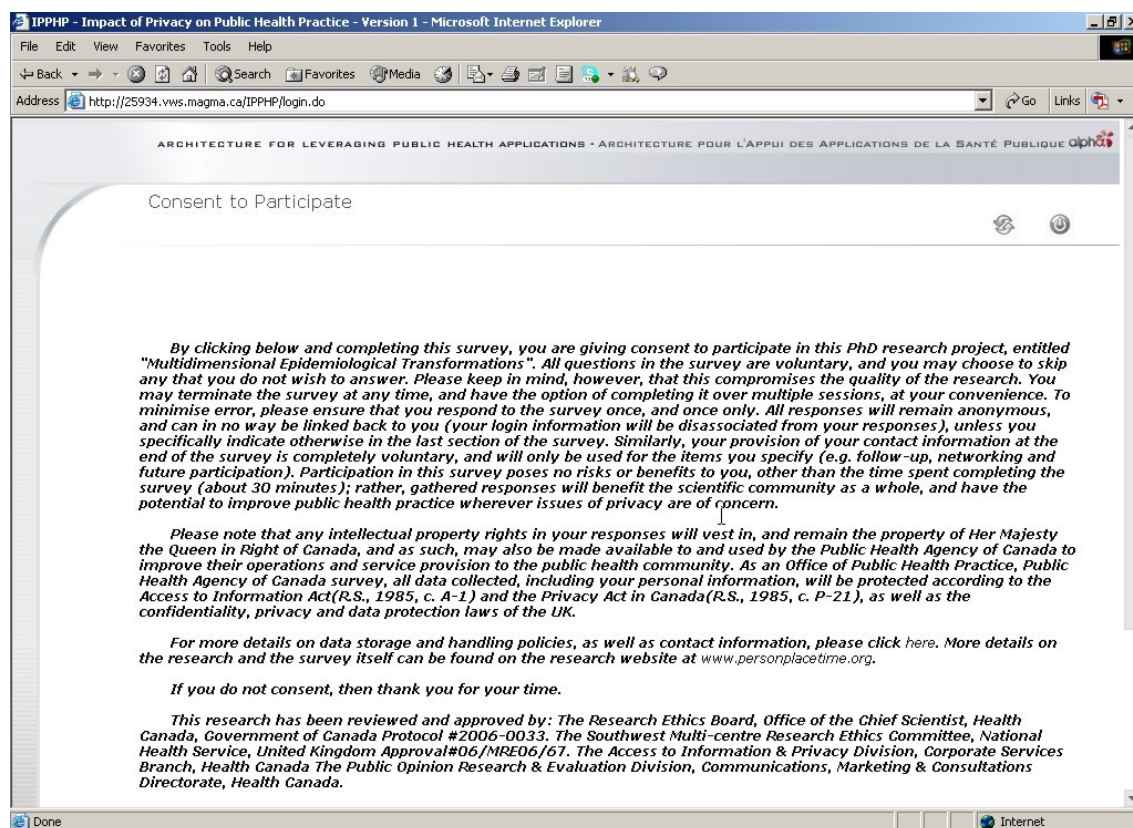
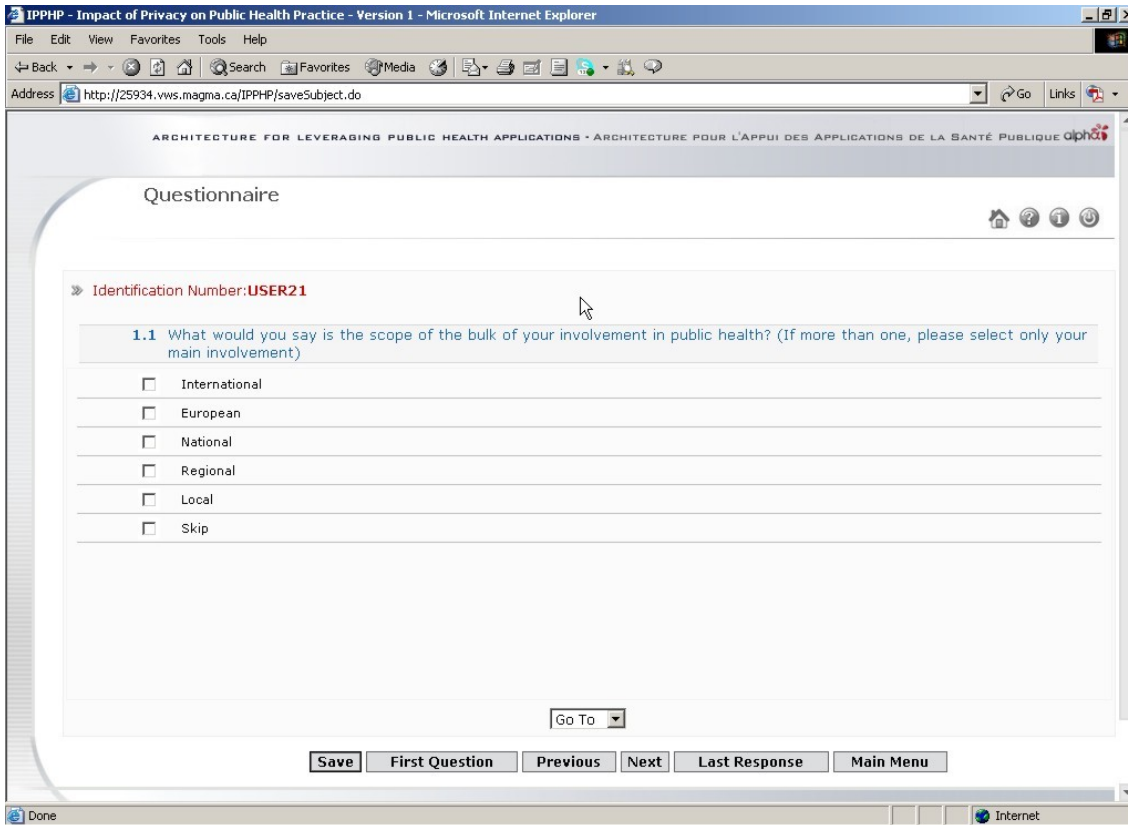


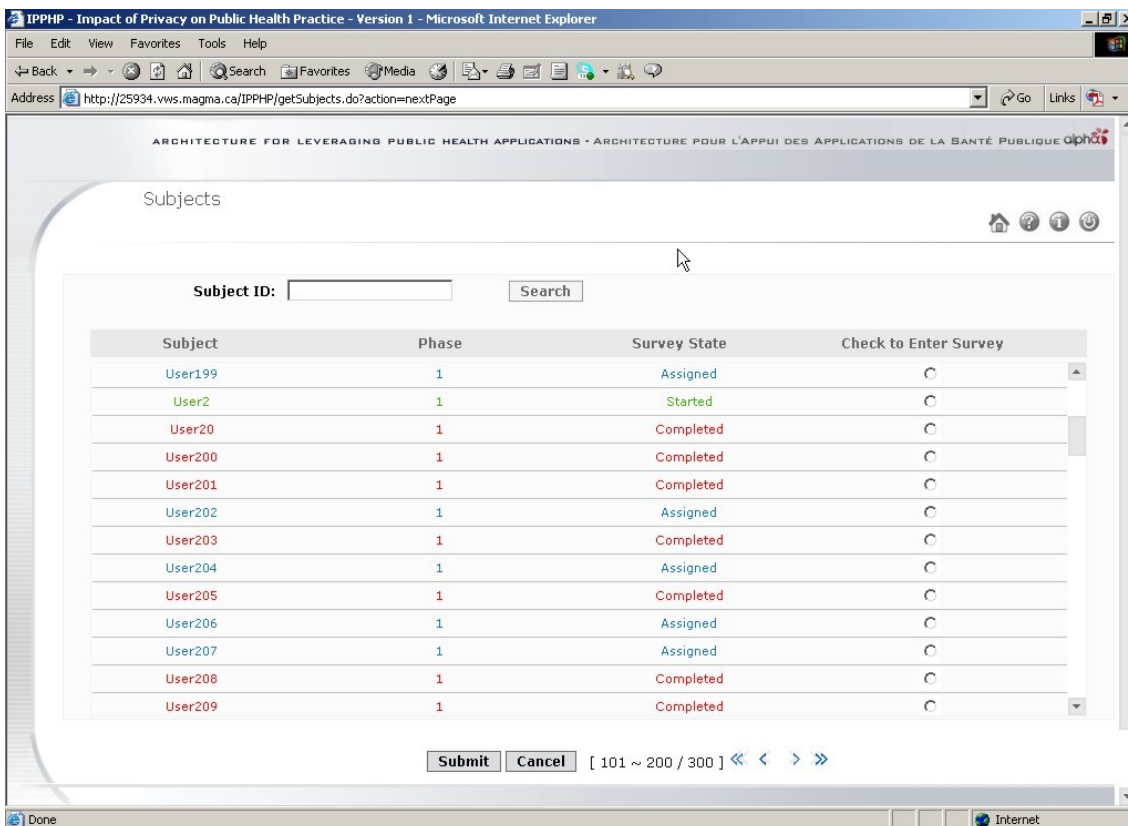
Figure A.8.2: Screenshot of the consent screen



**Figure A.8.3: Example screen from Web-based survey showing the first question**



**Figure A.8.4: Administrator screen**



## B. Survey Questionnaires

### B.1. PUBLIC HEALTH PROFESSIONAL QUESTIONNAIRE CANADA

**ESTIMATED TIME TO COMPLETE THIS QUESTIONNAIRE IS 20 TO 30 MINUTES**

Thank you for taking the time to complete this questionnaire – your contribution is highly valued and critical to the research.

This questionnaire is divided into eight short sections to better categorize and organize the type of information being captured. You may skip any question(s) you would rather not answer, however please keep in mind that this compromises the quality of the research. Please also note that, unless you choose to complete section VIII of the questionnaire, your responses will remain anonymous and cannot be linked back to you. Therefore, please ensure that all responses are clearly marked.

Please note that your responses will vest in, and remain the property of Her Majesty the Queen in Right of Canada, and as such, may also be made available to and used by the Public Health Agency of Canada to improve their operations and service provision to the public health community. As an Office of Public Health Practice, Public Health Agency of Canada survey, all data collected, including your personal information, will be protected according to the Access to Information Act and the Privacy Act in Canada.

SECTION I – A little about you...

(~ 5 minutes)

I-1 What would you say is the scope of the **bulk** of your involvement in public health?  
(If more than one, please select only your **main** involvement)

- International    N. American    National    Provincial/ Territorial    Regional/ Local

I-2 With which public health organization are you currently employed / affiliated?  
(If more than one, please select only your **main** organization)

- Public Health Agency of Canada  
 Health Canada  
 Other Federal Government Agency  
 Provincial Government (e.g. Ministry of Health)  
 Regional / Local Health Authority or Unit  
 Canadian Public Health Association  
 Other Non-Government Association (e.g. provincial epidemiological, diabetes, cancer, etc.)  
 University / Academia  
 Other

Please specify: \_\_\_\_\_

I-3. Please indicate your current specific area(s) of expertise: (Check as many as apply)

- |  |  |
|--|--|
| <input type="checkbox"/> Aboriginal Health   | <input type="checkbox"/> Food & Nutrition  |
| <input type="checkbox"/> Chronic Diseases (cancer, diabetes, etc.)                   | <input type="checkbox"/> Genetics  |
| <input type="checkbox"/> Child / Paediatric Public Health                            | <input type="checkbox"/> Health Services (needs, delivery, etc.)   |
| <input type="checkbox"/> Communicable / Infectious Diseases                          | <input type="checkbox"/> Injuries / Disability   |
| <input type="checkbox"/> Dental Public Health  | <input type="checkbox"/> Mental Health & Substance Misuse  |
| <input type="checkbox"/> Emergency Preparedness & Response                           | <input type="checkbox"/> Occupational Health   |
| <input type="checkbox"/> Environment (pollution, climate, water & food safety, etc.) | <input type="checkbox"/> Social Determinants of Health (e.g. poverty, education, social exclusion, etc.) |
| <input type="checkbox"/> Ethics, Public Health Law, Privacy, etc.                    | <input type="checkbox"/> Surveillance  |
| <input type="checkbox"/> Other   |  |

Please specify: \_\_\_\_\_

I-4. Which of the following **best** describe your roles or functions as a public health professional? (If more than one, please select only your **main** roles)

- |  |   |
|--|---|
| <input type="checkbox"/> Strategic decision / policy maker | <input type="checkbox"/> Research and Analysis                          |
| <input type="checkbox"/> Manager or Coordinator            | <input type="checkbox"/> Front-line responder / patient care / clinical |
| <input type="checkbox"/> Consultant                        |   |
| <input type="checkbox"/> Other                             |   |

Please specify: \_\_\_\_\_

I-5. Thinking of your regular activities, how much of your time (roughly, as a percentage) would you typically spend doing each of the following?

Strategic decision / policy making	_____ %
Management / Coordination	_____ %
Consultation	_____ %
Research / Analysis	_____ %
Front-line response / patient care / clinical	_____ %
Other (as specified in I-4)	_____ %

I-6. In which of the roles you identified above are you **most likely** to use or require the use of personally identifiable data?

- |  |   |
|--|---|
| <input type="checkbox"/> Strategic decision / policy maker | <input type="checkbox"/> Research and Analysis                          |
| <input type="checkbox"/> Manager or Coordinator            | <input type="checkbox"/> Front-line responder / patient care / clinical |
| <input type="checkbox"/> Consultant                        |   |
| <input type="checkbox"/> Other                             |   |

Please specify: \_\_\_\_\_

I-7. Do you have or foresee a need for including geographic location of health data in your roles or organization?

- YES  NO

I-8. Geographic Information Systems (GIS) are tools that allow you to visualise and analyse your data spatially – that is, using their geographical location on earth. In which of the roles you identified above would GIS be useful?

- |  |   |
|--|---|
| <input type="checkbox"/> Strategic decision / policy maker | <input type="checkbox"/> Research and Analysis                          |
| <input type="checkbox"/> Manager or Coordinator            | <input type="checkbox"/> Front-line responder / patient care / clinical |
| <input type="checkbox"/> Consultant                        |   |
| <input type="checkbox"/> Other                             |   |

Please specify: \_\_\_\_\_

I-9. What GIS application(s) do you currently use, or have you used in the past?

- Public Health Map Generator (PHMG)  
 Other Web-based: specify \_\_\_\_\_

Desktop GIS products:

- ESRI ArcGIS products  
 MapInfo  
 AutoDesk products  
 PCI Geomatics products  
 Intergraph products  
 Other

Please specify: \_\_\_\_\_

- I have never used any GIS applications, and have no use for them  
 I have never used any GIS applications, but am interested in learning more

I-10 At what level(s) of geography do you visualise your data and/or conduct spatial analyses for each product you use?

	PHMG	Other Web-Based	Desktop GIS
Latitude and Longitude	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Street address	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Dissemination area	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Postal Code	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Census Subdivision	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Census Division	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Forward Sortation Area	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Urban – Rural	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Provincial	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please specify: \_\_\_\_\_

I-11 Are you or have you been restricted in your use of GIS for any public health activity because of privacy concerns (i.e. map or data might identify an individual or community)?

YES → Please explain

No → Please explain

I-12 Setting privacy issues aside and in light of your response to the previous question, at what level(s) of geography would you **ideally** like to visualise your data and/or conduct spatial analyses for each product you use?

	PHMG	Other Web-Based	Desktop GIS
Latitude and Longitude	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Street address	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Dissemination area	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Postal Code	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Urban – Rural	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please specify: \_\_\_\_\_

SECTION II – Current access to data

(~ 4 minutes)

The questions in this section all pertain to the role you identified in question I-6 in Section I. If you do **not** have access to any of the *Personally Identifiable Data (PID)* listed in question II-1, please mark the last option in question II-1 and skip to Section III – No current access to data. Otherwise, please complete this section, and then skip to Section IV – Privacy Issues.

**NOTE:** The term “access” as used in this survey implies the ability to actually acquire individual level data so you can work with it directly.

II-1. What PID do you currently have access to? (Check as many as apply)

- |  |   |
|--|---|
| <input type="checkbox"/> First Name  | <input type="checkbox"/> Street Address           |
| <input type="checkbox"/> Last Name   | <input type="checkbox"/> Postal Code              |
| <input type="checkbox"/> Initials  | <input type="checkbox"/> Community Name           |
| <input type="checkbox"/> Sex   | <input type="checkbox"/> City / Town / Village    |
| <input type="checkbox"/> Date of Birth / Age   | <input type="checkbox"/> Region / Geographic Area |
| <input type="checkbox"/> Date of Death   | <input type="checkbox"/> Latitude & Longitude     |
| <input type="checkbox"/> Provincial Health Insurance Plan Number   |   |
| <input type="checkbox"/> Hospital ID   |   |
| <input type="checkbox"/> Registered GP / Family Physician  |   |
| <br>   |   |
| <input type="checkbox"/> Other   |   |
| Please specify: _____  |   |
| <br>   |   |
| <input type="checkbox"/> I do NOT currently have access to any of the above (please skip to Section III) |   |

The following questions all pertain to the *PID* you have access to, as identified in the previous question.

II-2 From a privacy and organisational bureaucracy perspective, how easy would you say it is for you to access this *PID* when you need it?  
Please circle the appropriate number, 1 being “Extremely difficult”, and 10 being “Very easy”

	Don't Know	Extremely difficult <span style="float: right;">→</span>								Very easy	
		1	2	3	4	5	6	7	8	9	10
First Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Last Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Provincial Health Insurance Plan Number	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hospital ID	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Postal Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

II-3. On average, how often do you access the *PID* you identified above?

		Rarely <span style="float: right;">→</span> All the time									
	Not Applicable	1	2	3	4	5	6	7	8	9	10
First Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Last Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Provincial Health Insurance Plan Number	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hospital ID	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Postal Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

II-4. How useful / important is this *PID* to you and your roles and responsibilities?  
 Please circle the appropriate number, with 1 being “Not at all useful”, and 10 being “Critical to my roles and responsibilities”

		Not at all useful <span style="float: right;">→</span> Critical									
	Not Applicable	1	2	3	4	5	6	7	8	9	10
First Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Last Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Provincial Health Insurance Plan Number	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hospital ID	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Postal Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



- II-5. What impact would removal of your access to this *PID* have on the quality of your work and resulting public health decisions?  
 Please circle the appropriate number, with 1 being “No impact – quality would not suffer”, and 10 being “Severe Impact - results and decisions would be severely compromised”

		No Impact <span style="float: right;">→</span> Severe Impact									
	Not Applicable	1	2	3	4	5	6	7	8	9	10
First Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Last Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Provincial Health Insurance Plan Number	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hospital ID	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Postal Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

- II-6. What *PID* do you currently **NOT** have access to, but believe would be beneficial to you to further enhance your work and resulting public health decisions?  
 (Check as many as apply)

- |  |   |
|--|---|
| <input type="checkbox"/> First Name                              | <input type="checkbox"/> Street Address           |
| <input type="checkbox"/> Last Name                               | <input type="checkbox"/> Postal Code              |
| <input type="checkbox"/> Initials                                | <input type="checkbox"/> Community Name           |
| <input type="checkbox"/> Sex                                     | <input type="checkbox"/> City / Town / Village    |
| <input type="checkbox"/> Date of Birth / Age                     | <input type="checkbox"/> Region / Geographic Area |
| <input type="checkbox"/> Date of Death                           | <input type="checkbox"/> Latitude & Longitude     |
| <input type="checkbox"/> Provincial Health Insurance Plan Number |   |
| <input type="checkbox"/> Hospital ID                             |   |
| <input type="checkbox"/> Registered GP / Family Physician        |   |

Other  
 Please specify: \_\_\_\_\_

None

Please skip to Section IV – Privacy Issues

SECTION III – No current access to data

(~ 2 minutes)

If you have access to *Personally Identifiable Data (PID)* and completed Section II above, then please skip to Section IV – Privacy Issues.

**NOTE:** The term “access” as used in this survey implies the ability to actually acquire individual level data so you can work with it directly.

III-1. Having access to which of the following *PID* would facilitate your roles and responsibilities, or enhance your work and improve resulting public health decisions? (Check as many as apply)

- |  |   |
|--|---|
| <input type="checkbox"/> First Name                              | <input type="checkbox"/> Street Address           |
| <input type="checkbox"/> Last Name                               | <input type="checkbox"/> Postal Code              |
| <input type="checkbox"/> Initials                                | <input type="checkbox"/> Community Name           |
| <input type="checkbox"/> Sex                                     | <input type="checkbox"/> City / Town / Village    |
| <input type="checkbox"/> Date of Birth / Age                     | <input type="checkbox"/> Region / Geographic Area |
| <input type="checkbox"/> Date of Death                           | <input type="checkbox"/> Latitude & Longitude     |
| <input type="checkbox"/> Provincial Health Insurance Plan Number |   |
| <input type="checkbox"/> Hospital ID                             |   |
| <input type="checkbox"/> Registered GP / Family Physician        |   |
| <br>   |   |
| <input type="checkbox"/> Other                                   |   |
| Please specify: _____  |   |
| <br>   |   |
| <input type="checkbox"/> None                                    |   |

III-2. How useful to you and your roles and responsibilities (as identified in Section I) would access to the *PID* you identified above be?

Please circle the appropriate number, with 1 being “Not at all useful”, and 10 being “Very useful – would greatly enhance by roles and responsibilities”

		Not at all useful <span style="font-size: 2em;">→</span> Very useful									
	Don't Know	1	2	3	4	5	6	7	8	9	10
First Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Last Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Provincial Health Insurance Plan Number	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hospital ID	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Postal Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

III-3 How easy would it be for you to access the *PID* you identified above, if you were to need it?

Please circle the appropriate number, with 1 being “Impossible”, and 10 being “Very easy”

		Impossible <span style="float: right;">→</span> Very easy									
	Don't Know	1	2	3	4	5	6	7	8	9	10
First Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Last Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Provincial Health Insurance Plan Number	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hospital ID	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Postal Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

III-4. What impact has your lack of access to this *PID* had on the quality of your work and resulting public health decisions?

Please circle the appropriate number, with 1 being “No impact – quality has not suffered”, and 10 being “Severe Impact - results and decisions have been severely compromised”

		No impact <span style="float: right;">→</span> Severe Impact									
	Don't Know	1	2	3	4	5	6	7	8	9	10
First Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Last Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Provincial Health Insurance Plan Number	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hospital ID	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Postal Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

SECTION IV – Privacy Issues

(~ 3 minutes)

This section pertains to the field of public health in general, and uses the term “*public health practice*” to refer to its various activities, including research, surveillance, health service delivery, strategic policy and decision making, etc. The goal is to get your opinion, as a public health professional, on the overall impact of restricted access to *PID* on *public health practice* in Canada. These questions ask for your opinion; if you’re not sure how to answer a question in this section, please just hazard a guess!

**NOTE:** The term “access” as used in this survey implies the ability to actually acquire individual level data so you can work with it directly.

IV-1. In your opinion, do current restrictions on access to *PID* pose an obstacle to any aspects of *public health practice* (e.g. research, surveillance, etc.)?

Please circle the appropriate number, with 1 being “Not an obstacle at all”, and 10 being “Yes, they pose a serious threat to accurate public health practice”

1	2	3	4	5	6	7	8	9	10	
Not an issue								→	Serious Threat	

IV-2. How amenable would **you** be to other professionals in the public health field having access to **your** *PID* for public health research and analyses (e.g. **your** address, date of birth, etc.) to improve public health delivery, service, etc.?

Sure, go ahead                       No Way                       Not Sure → Please Explain

IV-3a. In your opinion, what proportion of the public is aware of the impact of restricted access to *PID* on public health practice? (Please just guess!)

Please circle the approximate proportion

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
No one								→	Everyone	

IV-3b. How do you think we could increase this proportion?

IV-4a. In your opinion, what proportion of the public would allow the use of *PID* for public health practice if they were asked and educated on the usefulness of such data to public health practice? (Please just guess!) - Please circle the approximate proportion

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
No one								→	Everyone	

IV-4b. How do you think we could increase this proportion?

SECTION V – Current data holdings and provision to others...

(~ 2 minutes)

This section gathers information on the sharing of PID within and between organizations.

V-1. What would you say is the one **most** critical obstacle in the sharing or acquisition of PID linked to health data? (Please select only one; give your opinion!)

- National legislation     Public disapproval     Organizational bureaucracy  
 Lack of knowledge     Public Paranoia     Practitioner Paranoia  
 Other (please specify): \_\_\_\_\_

V-2. Do you or your organization currently collect individual-level health data for any purpose (e.g. research, surveillance, service delivery, etc.), or act as the custodian of such data?

- YES → Continue                       NO → Go to Section VI

V-3. For what specific purpose(s) is this data collected? (check as many as apply)

- Research                       Surveillance                       Service Delivery  
 Other  
 Please specify: \_\_\_\_\_

V-4. What data is collected?

- First Name                       Street Address  
 Last Name                       Postal Code  
 Initials                       Community Name  
 Sex                       City / Town / Village  
 Date of Birth / Age                       Region / Geographic Area  
 Date of Death                       Latitude & Longitude  
 Provincial Health Insurance Plan Number  
 Hospital ID  
 Registered GP / Family Physician  
 Other  
 Please specify: \_\_\_\_\_

V-5. How difficult is it for other public health professionals such as yourself to acquire access to your PID and linked health data holdings if they are outside your immediate working team, but within...

	D/K	Impossible <span style="float: right;">→</span> Very Easy									
		1	2	3	4	5	6	7	8	9	10
Your organisation?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The Federal Government?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Your Provincial Government?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A Provincial Government other than your own?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A regional or public health authority?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A University or Research Facility	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Another national government (e.g. CDC in the US, NHS in the UK, etc.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The World Health Organization	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

The proposed research will seek to apply a method (called a *transformation*) to public health data such that important relationships within and between the data are preserved, but the actual identity of the individual is anonymized. So, for example, if you were looking at an infectious outbreak in children, you might be interested in where the infected children are relative to one another, as well as where the schools are, arenas, community centres, etc. You would then preserve the *relationship* between these points of interest, and change everything else, so that the original points can no longer be identified back to their original owners. In this way, you have *transformed* the data so that you're still looking at individual-level data, but can't determine who it belongs to (i.e., it has become anonymous). Assuming the data custodians allow the data derived from such a *transformation* to be made available to the public health professional community:

For all scales, circle '0' if you "Don't know"

VI-1. How useful would such a transformation be to you in your current role?

Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

0	1	2	3	4	5	6	7	8	9	10	
Not useful		→									Very useful

VI-2. How useful do you think such a transformation would be to the field of public health in general?

Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

0	1	2	3	4	5	6	7	8	9	10	
Not useful		→									Very useful

VI-3. Imagine you are a data custodian, and that a method has been developed to take your individual level data and mask it or change it somehow, while still keeping it at an individual-by-individual level. Would you allow such a method to be conducted on your data so that it can be shared with other public health professionals for public health research and practice?

YES       NO → Please explain why not       MAYBE → Please explain

A specific disease or health condition will be used to test and evaluate the developed method(s). This condition must have a known aetiology, with well-known patterns and relationships, to serve as a starting point for the research. It must also be a disease of interest to the public health community.

VI-4. What diseases, health conditions, or databases most immediately come to mind as potential subjects for this research?

VI-5. Based on your knowledge of the condition you identified in the previous question, what relationships to the physical environment would a *transformation* as defined in the opening paragraph of this section have to retain in order for the data to be meaningful and useful to you (e.g. where patients or cases are relative to each other, to schools, to restaurants, etc; where schools are relative to a type of industry; etc...)?

Another proposed solution to the issue at hand is to build what are called automated *software agents*. You can think of these as applications that would go into a dataset wherever it is housed (i.e. at the custodian's location), perform the analyses for you (on the *personally identifiable data*) and return only the aggregated, and therefore anonymised, results. In other words, you would never see the actual data, but would have this "agent" do the analyses for you, directly on the *PID*; you simply get the results of the analyses, as long as, of course, they don't compromise privacy. As a simple analogy, it would be like you giving me an equation or function to perform on my data, and I giving you back the result of that function without you ever needing to see my actual data. Assuming the data custodians allow such a *software agent* to analyse their data and make the results available to the public health professional community:

VI-6. How useful do you think such a *software agent* would be to you in your current role?  
Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

0	1	2	3	4	5	6	7	8	9	10	
Not useful		→								Very useful	

VI-7. How useful do you think such a *software agent* would be to the field of public health in general?  
Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

0	1	2	3	4	5	6	7	8	9	10	
Not useful		→								Very useful	

VI-8. If you were (or are) a data custodian, would you allow such a *software agent* to access your data, conduct the analyses, and return the results to the public health professional community for research and analysis?

YES                       NO → Please explain why not                       MAYBE → Please explain

VI-9. To summarise, if a solution is found such that privacy is no longer an issue, which of the following would you prefer? (Please select only one)

- I would prefer to be able to work directly with the raw data, so I can access information on a case-by-case basis.
- I have no need to see the raw data, and would prefer to access information and results on an aggregate basis.

SECTION VII – Qualitative Component

(~ 5 minutes)

VII-1. How knowledgeable do you consider yourself on privacy and confidentiality issues / legislation?

Please circle the appropriate number, 1 being “Not at all knowledgeable”, and 10 being “Expert”

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

Not Knowledgeable → Expert

VII-2. How do you feel about the impact of privacy and confidentiality legislation – in particular the restrictions on access to personally identifiable data (e.g. The Privacy Act, The Personal Information Protection and Electronic Documents Act, etc.) – on public health?

VII-3. What do you think of the proposed research (development of a *transformation*)?

VII-4. What do you think of the “*software agent*” idea?

VII-5. Do you have any other thoughts or comments regarding this issue, the proposed research, or this questionnaire that you would like to share?



## SECTION VIII – Further Participation and Contact

Please indicate your desired level of anonymity and interest in further participation; you may check multiple boxes as applicable. Please note that leaving this section empty will default your response to “absolute anonymity”, rendering you answers on this questionnaire personally unidentifiable and removing yourself from any further contact or participation.

- You may link my identity to my responses on this questionnaire for clarifications and follow-up
- Please send me a summary of the findings of this questionnaire, once completed
- Please send me periodic updates on the progress of this research by email to the address given below
- I am interested in piloting the results of this research

If you checked any of the above boxes, please complete your details below. Your personal information will be stored in a password-protected directory within the Public Health Agency of Canada, and will be protected according to the Access to Information Act and the Privacy Act.

**Name:** \_\_\_\_\_

**Title:** \_\_\_\_\_

**Organization:** \_\_\_\_\_

**Address:** \_\_\_\_\_  
\_\_\_\_\_

**E-Mail:** \_\_\_\_\_

**Phone:** \_\_\_\_\_

**Preferred Method of Contact:**  Phone  Fax  E-Mail  Mail

Are you in possession of any personally identifiable data that you can use for testing and evaluation of the developed transformation?

YES  NO

Dear Public Health Professional,

Thank you so much for taking the time to complete this questionnaire; your responses will help assess the impact of privacy and confidentiality legislation on public health research, and will be used to investigate and develop a disease-specific solution, which will, in turn, enhance strategic decisions and research in public health.

Once again, many thanks for your time, and I look forward to enhancing public health practice by exploring this issue, and its solution, further with you. Should you have any comments, questions or concerns, please feel free to send me an email or give me a ring.

Best wishes,

Philip AbdelMalik

## **B.2. L'IMPACT DE LA PROTECTION DES RENSEIGNEMENTS PERSONNELS SUR LA PRATIQUE EN SANTÉ PUBLIQUE CANADA**

**TEMPS REQUIS ESTIMÉ POUR RÉPONDRE À CE QUESTIONNAIRE : 20 À 30 MINUTES**

*Merci de prendre le temps de remplir ce questionnaire - votre contribution est bien appréciée et critique à la recherche.*

Ce questionnaire est divisé en huit courtes sections pour mieux classer par catégorie et organiser les types d'information. Vous pouvez sauter pardessus n'importe quel question. Notez toutefois que chaque saut compromet la qualité de la recherche. Veuillez noter également que, à moins que vous choisissiez d'accomplir la section VIII du questionnaire, vos réponses demeureront anonymes. Assurez-vous svp que toutes vos réponses sont inscrites clairement.

Veuillez noter que vos réponses deviendront la propriété du gouvernement Canada, et comme tels, peuvent également être rendues disponibles pour être employées par l'agence de santé publique du Canada pour améliorer ses opérations et pour assurer des services à la communauté de santé publique. Ce questionnaire est mené par le bureau de la pratique en santé publique de l'agence de la santé publique du Canada. Donc, toutes les données rassemblées, y compris votre information personnelle, seront protégées selon la loi d'accès à l'information et la loi de protection des renseignements personnels du Canada. Notez également que, à moins que vous choisissiez d'accomplir la section VIII du questionnaire, vos réponses demeureront anonymes. Par conséquent, assurez-vous svp que toutes les réponses sont clairement marquées.

## SECTION I – Un aperçu à votre sujet

(~ 5 minutes)

- I-1 A quel niveau se situe la portée de l'**essentiel** de vos activités en santé publique ?  
(Si plus d'un choix s'applique, veuillez ne cocher que la case correspondant au niveau **principal** de vos activités)

<input type="checkbox"/> International	<input type="checkbox"/> Amérique du Nord	<input type="checkbox"/> National	<input type="checkbox"/> Provincial/ Terr itori al	<input type="checkbox"/> Régional/ Local
--	---	-----------------------------------	---	---

- I-2 Après de quel type d'organisation de santé publique travaillez-vous ou êtes-vous affilié ?  
(Si plus d'un choix s'applique, veuillez ne cocher que la case correspondant au type d'organisation **principal** de vos activités)

Agence de santé publique du Canada  
 Santé Canada  
 Autre agence gouvernementale fédérale  
 Gouvernement provincial (ministère de la santé, etc.)  
 Autorité ou régie régionale ou locale  
 Association Canadienne de santé publique  
 Autre organisme non gouvernemental (provincial, épidémiologie, diabète, cancer, etc.)  
 Université ou établissement d'enseignement  
 Autre  
 (préciser svp) \_\_\_\_\_

- I-3. Veuillez préciser le ou les domaine(s) d'expertise dans lesquels vous exercez  
présentement : (Cocher toutes les mentions pertinentes)

<input type="checkbox"/> Santé des Autochtones	<input type="checkbox"/> Aliments et nutrition
<input type="checkbox"/> Maladies chroniques (cancer, diabète, etc.)	<input type="checkbox"/> Génétique
<input type="checkbox"/> Santé publique des enfants / pédiatrique	<input type="checkbox"/> Services et soins de santé (besoins, prestation des services, etc.)
<input type="checkbox"/> Maladies transmissibles / infectieuses	<input type="checkbox"/> Préjudices corporels / Invalidité
<input type="checkbox"/> Santé publique dentaire	<input type="checkbox"/> Santé mentale et toxicomanies
<input type="checkbox"/> Protection civile / mesures d'urgence	<input type="checkbox"/> Santé en milieu de travail
<input type="checkbox"/> Environnement (pollution, climat, sécurité de l'eau et des aliments, etc.)	<input type="checkbox"/> Déterminants sociaux de la santé (pauvreté, éducation, exclusion sociale, etc.)
<input type="checkbox"/> Éthique, droit de la santé publique, protection des renseignements personnels, etc.	<input type="checkbox"/> Surveillance
<input type="checkbox"/> Autre (préciser svp) _____	

- I-4. Quelle attribution, parmi celles mentionnées ci-après, décrit le mieux votre rôle ou vos fonctions à titre de professionnel de la santé publique ?  
(Si plus d'un choix s'applique, veuillez ne cocher que la case correspondant à votre rôle ou vos fonctions **principales**)

<input type="checkbox"/> Décideur / concepteur de politiques stratégiques	<input type="checkbox"/> Recherche et analyse
<input type="checkbox"/> Gestionnaire ou coordonnateur	<input type="checkbox"/> Intervenant de première ligne / soins aux patients / services cliniques
<input type="checkbox"/> Consultant	
<input type="checkbox"/> Autre (préciser svp) _____	

I-5. Dans le cadre de vos activités régulières, combien de temps consacrez-vous (approximativement, en pourcentage) ordinairement à l'accomplissement de chacune des tâches suivantes ?

Décideur / concepteur de politiques stratégiques	_____ %
Gestionnaire ou coordonnateur	_____ %
Consultant	_____ %
Recherche et analyse	_____ %
Intervenant de première ligne / soins aux patients / services cliniques	_____ %
Autre (comme indiqué dedans I-4)	_____ %

I-6. Dans l'accomplissement de quelle(s) tâche(s) mentionnées ci-dessus est-il **le plus probable** que vous utilisiez ou auriez besoin d'utiliser des renseignements nominatifs (permettant d'identifier un individu) ?

- |   |  |
|---|--|
| <input type="checkbox"/> Décideur / concepteur de politiques stratégiques | <input type="checkbox"/> Recherche et analyse  |
| <input type="checkbox"/> Gestionnaire ou coordonnateur                    | <input type="checkbox"/> Intervenant de première ligne / soins aux patients / services cliniques |
| <input type="checkbox"/> Consultant                                       |  |
| <input type="checkbox"/> Autre  |  |
- (préciser svp) \_\_\_\_\_

I-7. Avez-vous besoin ou prévoyez-vous le besoin d'inclure des données sur la zone géographique des données recueillies en matière de santé, dans le cadre de vos activités ou de celles de votre organisation ?

OUI  NON

I-8. Un système d'information géographique (SIG) est un outil vous permettant de visualiser et d'analyser vos données en y intégrant une composante spatiale, c'est-à-dire en tenant compte de leur emplacement géographique sur la planète. Dans l'accomplissement de quelle(s) tâche(s) le recours au SIG pourrait-il vous être utile ?

- |   |  |
|---|--|
| <input type="checkbox"/> Décideur / concepteur de politiques stratégiques | <input type="checkbox"/> Recherche et analyse  |
| <input type="checkbox"/> Gestionnaire ou coordonnateur                    | <input type="checkbox"/> Intervenant de première ligne / soins aux patients / services cliniques |
| <input type="checkbox"/> Consultant                                       |  |
| <input type="checkbox"/> Autre  |  |
- (préciser svp) \_\_\_\_\_

I-9. Quelle(s) application(s) de SIG utilisez-vous présentement ou avez-vous utilisées par le passé ?

- Générateur de cartes en santé publique (GCSP)  
 Autre (disponibles sur le web; svp indiquez) \_\_\_\_\_

Produits SIG utilisés en bureautique :

- ESRI (ArcGIS)  
 MapInfo  
 AutoDesk  
 PCI Geomatics  
 Intergraph  
 Autre

(préciser svp) \_\_\_\_\_

- Je n'ai jamais utilisé des applications de SIG, et je n'y vois aucune utilité dans mes activités  
 Je n'ai jamais utilisé des applications de SIG, mais je souhaite en apprendre davantage à ce sujet

I-10 À quel niveau de géographie est-ce que vous visualisez vos données ou en faites une analyse spatiale pour chacun des produits que vous utilisez, le cas échéant ?

	GCSP	Autres niveaux disponibles sur le web	Produits SIG utilisés en bureautique
Latitude et Longitude	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Adresse civique	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Aire de diffusion	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Code postal	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Subdivision de recensement	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Division de recensement	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Région de tri d'acheminement	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Urbain – Rural	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Provincial	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Autre			

(préciser svp) \_\_\_\_\_

I-11 Êtes-vous restreint dans l'utilisation de SIG ou devez-vous restreindre votre utilisation de SIG pour accomplir quelque activité de santé publique en raison de préoccupations au niveau de la protection des renseignements personnels (par exemple, de crainte que la carte ou les données soient susceptibles de permettre l'identification d'un individu ou d'une communauté) ?

OUI → svp expliquez

NON → svp expliquez

I-12 En faisant abstraction pour l'instant de vos préoccupations au niveau de la protection des renseignements personnels et selon votre réponse à la question précédente, à quel niveau de géographie est-ce que vous souhaiteriez **idéalement** pouvoir visualiser vos données ou en faire une analyse spatiale pour chacun des produits que vous utilisez, le cas échéant ?

	GCSP	Autres niveaux disponibles sur le web	Produits SIG utilisés en bureautique
Latitude et Longitude	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Adresse civique	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Aire de diffusion	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Code postal	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Urbain – Rural	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Autre			

(préciser svp) \_\_\_\_\_

Toutes les questions dans cette section concernent le rôle que vous avez identifié à la question I-6 dans la section I. Si vous n'avez pas accès à des données personnellement identifiables (PID) énumérées à la question II-1, cochez svp la dernière option à la question II-1 et sautez à la section III - Aucun accès courant aux données. Sinon, remplissez svp cette section, et passez ensuite à la section IV sur la protection des renseignements personnels.

**REMARQUE :** Dans ce questionnaire, le terme accès se rapporte à votre capacité d'acquérir effectivement des renseignements nominatifs de manière à pouvoir travailler directement avec ces données.

II-1. En tenant compte des tâches dans lesquelles vous êtes le plus susceptible d'avoir besoin d'utiliser des données nominatives, à quels renseignements nominatifs avez-vous présentement accès ? (Cocher toutes les mentions pertinentes)

- |  |   |
|--|---|
| <input type="checkbox"/> Prénom  | <input type="checkbox"/> Adresse civique                |
| <input type="checkbox"/> Nom   | <input type="checkbox"/> Code postal                    |
| <input type="checkbox"/> Initiales   | <input type="checkbox"/> Nom de la communauté           |
| <input type="checkbox"/> Sexe  | <input type="checkbox"/> Ville / Municipalité / Village |
| <input type="checkbox"/> Date de naissance/ Age  | <input type="checkbox"/> Région / Zone géographique     |
| <input type="checkbox"/> Date de décès   | <input type="checkbox"/> Latitude et Longitude          |
| <input type="checkbox"/> Numéro d'assurance-maladie  |   |
| <input type="checkbox"/> Numéro d'identification d'une carte d'hôpital   |   |
| <input type="checkbox"/> Médecin généraliste ou en médecine familiale  |   |
| <input type="checkbox"/> Autre<br>(préciser svp) _____   |   |
| <input type="checkbox"/> <b>Je n'ai PAS accès présentement à l'un ou l'autre de ces éléments (svp sautez à la Section III)</b> |   |

Les questions qui suivent se rapportent toutes aux renseignements nominatifs auxquels vous avez accès, tel que vous l'avez indiqué à la question précédente à ce sujet.

- II-2 Dans une perspective de protection des renseignements personnels et organisationnels, comment jugez-vous le degré de facilité avec lequel vous avez accès à des renseignements nominatifs lorsque vous en avez besoin ?  
**Encercler le chiffre correspondant à votre réponse, 1 = « Extrêmement difficile » et 10 = « Très facile »**

		Extrêmement difficile  Très facile									
	Ne sais pas	1	2	3	4	5	6	7	8	9	10
Prénom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initiales	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sexe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de naissance/ Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de décès	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'assurance- maladie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'identification d'une carte d'hôpital	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Médecin généraliste ou en médecine familiale	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Adresse civique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Code postal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom de la communauté	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ville / Municipalité / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Région / Zone géographique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

II-3. En général, combien de fois accédez-vous aux renseignements nominatifs que vous avez mentionnés ci-dessus ?

Encercler le chiffre correspondant à votre réponse, 1 = « rarement » et 10 = « Tout le temps »

		Rarement							Tout le temps		
	Non applicable	1	2	3	4	5	6	7	8	9	10
Prénom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initiales	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sexe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de naissance/ Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de décès	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'assurance- maladie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'identification d'une carte d'hôpital	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Médecin généraliste ou en médecine familiale	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Adresse civique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Code postal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom de la communauté	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ville / Municipalité / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Région / Zone géographique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



- II-4. Dans quelle mesure ces renseignements nominatifs vous sont-ils utiles ou utiles dans l'accomplissement des rôles et responsabilités qui vous incombent ?  
**Encercler le chiffre correspondant à votre réponse, 1 = « Pas du tout utiles » et 10 = « Essentiels à l'accomplissement des rôles et responsabilités qui m'incombent »**

		Pas du tout utiles								Essentiels	
	Ne sais pas	1	2	3	4	5	6	7	8	9	10
Prénom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initiales	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sexe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de naissance/ Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de décès	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'assurance- maladie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'identification d'une carte d'hôpital	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Médecin généraliste ou en médecine familiale	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Adresse civique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Code postal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom de la communauté	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ville / Municipalité / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Région / Zone géographique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

II-5. Quel impact aurait le fait de perdre votre accès à ces renseignements nominatifs sur la qualité de votre travail et les décisions en matière de santé publique qui en résultent ?  
**Encercler le chiffre correspondant à votre réponse, 1 = « Aucun impact – aucune incidence sur la qualité » et 10 = « Impact sérieux – les résultats et les décisions en seraient gravement compromis »**

	Ne sais pas	Aucun impact						Impact sérieux			
		1	2	3	4	5	6	7	8	9	10
Prénom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initiales	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sexe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de naissance/ Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de décès	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'assurance- maladie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'identification d'une carte d'hôpital	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Médecin généraliste ou en médecine familiale	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Adresse civique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Code postal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom de la communauté	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ville / Municipalité / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Région / Zone géographique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

II-6. À quels renseignements nominatifs n'avez-vous **PAS** présentement accès, mais dont l'accès vous serait bénéfique à votre avis à fin d'améliorer votre travail et les décisions en matière de santé publique qui en résultent ? **(Cocher toutes les mentions pertinentes)**

- |  |   |
|--|---|
| <input type="checkbox"/> Prénom  | <input type="checkbox"/> Adresse civique                |
| <input type="checkbox"/> Nom   | <input type="checkbox"/> Code postal                    |
| <input type="checkbox"/> Initiales                                     | <input type="checkbox"/> Nom de la communauté           |
| <input type="checkbox"/> Sexe  | <input type="checkbox"/> Ville / Municipalité / Village |
| <input type="checkbox"/> Date de naissance/ Age                        | <input type="checkbox"/> Région / Zone géographique     |
| <input type="checkbox"/> Date de décès                                 | <input type="checkbox"/> Latitude et Longitude          |
| <input type="checkbox"/> Numéro d'assurance-maladie                    |   |
| <input type="checkbox"/> Numéro d'identification d'une carte d'hôpital |   |
| <input type="checkbox"/> Médecin généraliste ou en médecine familiale  |   |
| <input type="checkbox"/> Autre<br>(préciser svp) _____                 |   |
| <input type="checkbox"/> Aucun   |   |

**PERSON SHOULD NOW BE TAKEN TO SECTION IV – PRIVACY ISSUES**

**If the person completed Section II above, then skip to Section IV – Privacy Issues.**

**REMARQUE :** Dans ce questionnaire, le terme accès se rapporte à votre capacité d'acquérir effectivement des renseignements nominatifs de manière à pouvoir travailler directement avec ces données.

III-1. L'accès à quels renseignements nominatifs, parmi la liste suivante, vous serait bénéfique à votre avis à fin d'améliorer votre travail et les décisions en matière de santé publique qui en résultent ?

(Cocher toutes les mentions pertinentes)

- |                          |   |                          |                                |
|--------------------------|---|--------------------------|--------------------------------|
| <input type="checkbox"/> | Prénom  | <input type="checkbox"/> | Adresse civique                |
| <input type="checkbox"/> | Nom   | <input type="checkbox"/> | Code postal                    |
| <input type="checkbox"/> | Initiales                                     | <input type="checkbox"/> | Nom de la communauté           |
| <input type="checkbox"/> | Sexe  | <input type="checkbox"/> | Ville / Municipalité / Village |
| <input type="checkbox"/> | Date de naissance/ Age                        | <input type="checkbox"/> | Région / Zone géographique     |
| <input type="checkbox"/> | Date de décès                                 | <input type="checkbox"/> | Latitude et Longitude          |
| <input type="checkbox"/> | Numéro d'assurance-maladie                    |                          |                                |
| <input type="checkbox"/> | Numéro d'identification d'une carte d'hôpital |                          |                                |
| <input type="checkbox"/> | Médecin généraliste ou en médecine familiale  |                          |                                |
| <input type="checkbox"/> | Autre<br>(préciser svp) _____                 |                          |                                |
| <input type="checkbox"/> | Aucun   |                          |                                |

III-2. Dans quelle mesure l'accès aux renseignements nominatifs que vous avez indiqués ci-dessus serait-il utile à l'accomplissement des rôles et responsabilités qui vous incombent (que vous avez indiqués en réponse à la Section I) ?

Encercler le chiffre correspondant à votre réponse, 1 = « Pas du tout utile » et 10 = « Très utile – aideraient grandement les rôles et responsabilités qui m'incombent »

	Pas du tout utile <span style="font-size: 2em;">→</span> Très utile										
	Ne sais pas	1	2	3	4	5	6	7	8	9	10
Prénom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initiales	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sexe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de naissance/ Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de décès	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'assurance- maladie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'identification d'une carte d'hôpital	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Médecin généraliste ou en médecine familiale	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Adresse civique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Code postal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom de la communauté	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ville / Municipalité / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Région / Zone géographique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

III-3 Dans quelle mesure cela serait-il facile pour vous d'accéder aux renseignements nominatifs que vous avez identifiés ci-dessus, si vous deviez en avoir besoin ?

Encercler le chiffre correspondant à votre réponse, 1 = « Impossible » et 10 = « Très facile »

		Impossible <span style="float: right;">→</span> Très facile									
	Ne sais pas	1	2	3	4	5	6	7	8	9	10
Prénom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initiales	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sexe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de naissance/ Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de décès	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'assurance- maladie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'identification d'une carte d'hôpital	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Médecin généraliste ou en médecine familiale	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Adresse civique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Code postal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom de la communauté	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ville / Municipalité / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Région / Zone géographique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

III-4. Quel impact le fait de ne pas avoir accès à ces renseignements nominatifs a-t-il sur la qualité de votre travail et les décisions en matière de santé publique qui en résultent ?  
**Encercler le chiffre correspondant à votre réponse, 1 = « Aucun impact – aucune incidence sur la qualité » et 10 = « Impact sérieux – les résultats et les décisions ont été gravement compromis »**

	Aucun impact <span style="float: right;">→</span> Impact sérieux										
	Ne sais pas	1	2	3	4	5	6	7	8	9	10
Prénom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initiales	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sexe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de naissance/ Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date de décès	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'assurance- maladie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Numéro d'identification d'une carte d'hôpital	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Médecin généraliste ou en médecine familiale	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Adresse civique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Code postal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nom de la communauté	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ville / Municipalité / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Région / Zone géographique	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

## SECTION IV – Protection des renseignements personnels

(~ 3 minutes)

Cette section a trait au domaine de la santé publique dans son ensemble; l'emploi de l'expression « pratique en santé publique » vise les diverses activités réalisées dans ce domaine : recherche, surveillance de la santé, prestation des services de santé, élaboration des politiques et prise de décisions stratégiques, etc. L'objectif est d'obtenir votre opinion, en tant que professionnel de la santé publique, au sujet de l'impact global de la restriction de l'accès à des renseignements nominatifs *sur la pratique en santé publique* au Canada. Ces questions ont pour objet de recueillir votre opinion à ce sujet; si vous n'êtes pas certain de la réponse à donner à une question donnée dans cette section, veuillez quand même donner une réponse selon votre bon jugement.

**REMARQUE :** Dans ce questionnaire, le terme accès se rapporte à votre capacité d'acquérir effectivement des renseignements nominatifs de manière à pouvoir travailler directement avec ces données.

- IV-1. À votre avis, les restrictions actuelles sur l'accès aux renseignements nominatifs constituent-elles un obstacle à quelque volet de la *pratique en santé publique* (recherche, surveillance, etc.) ?  
 Encercler le chiffre correspondant à votre réponse, 1=«Pas du tout un obstacle» et 10=«Elles constituent une grave menace à l'exercice efficace et convenable de la pratique en santé publique»

1	2	3	4	5	6	7	8	9	10	
Pas du tout un obstacle								→	Une grave menace	

- IV-2. Dans quelle mesure seriez-vous disposé à ce que d'autres professionnels du domaine de la santé publique aient accès à des renseignements nominatifs **vous concernant** aux fins de recherche et d'analyse en santé publique (**votre** adresse, date de naissance, etc.) à fin d'améliorer la prestation des services en matière de santé publique, etc. ?

D'accord, pas de problème     Aucunement     Pas certain → Préciser pourquoi :

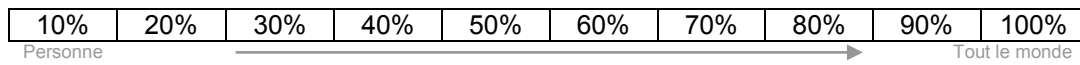
- IV-3a. À votre avis, dans quelle proportion le grand public est-il au courant de l'impact de l'accès restreint aux renseignements nominatifs sur la pratique en santé publique ?  
 Svp encercler la proportion approximative

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
Personne							→	Tout le monde		

- IV-3b. De quelle manière croyez-vous que l'on pourrait augmenter cette proportion ?

IV-4a. À votre avis, quelle proportion du grand public permettrait que l'on utilise les renseignements nominatifs des individus aux fins de la pratique en santé publique si on interrogeait les gens et les sensibilisait au sujet de l'utilité de tels renseignements aux fins de la pratique en santé publique ?

Svp encercler la proportion approximative



IV-4b. Comment pourrait-on augmenter cette proportion ?



SECTION V – Données actuelles et dispositions pour l'obtention d'autres renseignements...

(~ 2 minutes)

Cette section vise à recueillir des informations au sujet du partage des renseignements nominatifs au sein de votre organisation et entre des organisations.

Dans l'échelle des choix de réponse, encrer « 0 » si votre réponse est « Ne sais pas ».

V-1. À votre avis, quel serait le **principal** obstacle au partage ou à l'acquisition de renseignements nominatifs liés à des informations sur la santé des individus ?  
(Veuillez ne cocher qu'un seul choix – il s'agit de faire valoir votre opinion !)

- |   |   |   |
|---|---|---|
| <input type="checkbox"/> Les lois nationales        | <input type="checkbox"/> La réprobation du public | <input type="checkbox"/> La bureaucratie            |
| <input type="checkbox"/> L'absence de connaissances | <input type="checkbox"/> La paranoïa du public    | <input type="checkbox"/> La paranoïa des praticiens |
| <input type="checkbox"/> Autre (préciser svp) _____ |   |   |

V-2. Est-ce que vous ou votre organisation recueillez à l'heure actuelle des renseignements nominatifs en matière de santé à quelque fin que ce soit (recherche, surveillance, prestation des services, etc.), ou agissez à titre de dépositaire de tels renseignements ?

- OUI                       NON **(Go to Section VI)**

V-3. À quelles fins spécifiques ces renseignements sont-ils recueillis ?  
(Cocher toutes les mentions pertinentes)

- |   |                                       |  |
|---|---------------------------------------|--|
| <input type="checkbox"/> Recherche                  | <input type="checkbox"/> Surveillance | <input type="checkbox"/> Prestation des services |
| <input type="checkbox"/> Autre (préciser svp) _____ |                                       |  |

V-4. Quels renseignements sont recueillis ?

- |  |   |
|--|---|
| <input type="checkbox"/> Prénom  | <input type="checkbox"/> Adresse civique                |
| <input type="checkbox"/> Nom   | <input type="checkbox"/> Code postal                    |
| <input type="checkbox"/> Initiales                                     | <input type="checkbox"/> Nom de la communauté           |
| <input type="checkbox"/> Sexe  | <input type="checkbox"/> Ville / Municipalité / Village |
| <input type="checkbox"/> Date de naissance/ Age                        | <input type="checkbox"/> Région / Zone géographique     |
| <input type="checkbox"/> Date de décès                                 | <input type="checkbox"/> Latitude et Longitude          |
| <input type="checkbox"/> Numéro d'assurance-maladie                    |   |
| <input type="checkbox"/> Numéro d'identification d'une carte d'hôpital |   |
| <input type="checkbox"/> Médecin généraliste ou en médecine familiale  |   |
| <input type="checkbox"/> Autre (préciser svp) _____                    |   |

V-5. Dans quelle mesure est-ce difficile pour les professionnels de la santé publique, dont vous-même, d'avoir accès aux renseignements nominatifs dont vous disposez et autres données sur la santé qui y sont liées lorsque ces renseignements ou données sont situées à l'extérieur du périmètre immédiat de votre équipe de travail, mais à l'intérieur ...

		Impossible <span style="display: inline-block; width: 100px; border-bottom: 1px solid black; position: relative; top: -5px;"> <span style="position: absolute; right: -10px; top: -5px;">→</span> </span> Très facile									
	Ne sais pas	1	2	3	4	5	6	7	8	9	10
De votre organisation ?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Du gouvernement fédéral	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
De votre gouvernement provincial	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
D'un gouvernement provincial autre que le vôtre?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
D'une régie régionale ou autre autorité chargée de la santé publique ?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
D'une université ou établissement de recherche?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
De l'administration d'un gouvernement étranger (p. ex. le CDC aux États-Unis, le NHS au Royaume-Uni, etc.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
De l'organisation de santé mondiale	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

La recherche proposée s'intéressera à l'application d'une méthode (que l'on appelle *transformation*) à des données relatives à la santé publique de manière à ce que les liens importants entre les données et parmi celles-ci soient préservés, tout en rendant anonyme l'identité de l'individu. Par exemple, si l'on étudie le phénomène d'une éclosion d'une maladie infectieuse chez les enfants, vous vous intéresseriez à savoir où sont situés les enfants infectés les uns par rapport aux autres, ainsi qu'où sont situés les écoles, arénas, centres communautaires, etc.. Vous pourriez alors préserver *les liens* entre ces points d'intérêt et changer tous les autres éléments, de manière à ce que les points d'intérêt ainsi ciblés ne puissent être retracés vers l'identité des individus en jeu. Ainsi, vous avez réussi à *transformer* les données de telle sorte que les données soumises à l'analyse soient visibles au niveau individuel comme tel, mais sans pouvoir établir précisément à quelle personne elles se rapportent (les renseignements deviennent alors anonymes). Ceci suppose que les dépositaires des renseignements et des données permettent que les données issues de cette *transformation* soient mises à la disposition des professionnels de la santé publique .

Dans l'échelle des choix de réponse, encrer « 0 » si votre réponse est « Ne sais pas ».

VI-1. Dans quelle mesure une telle transformation vous serait-elle utile dans vos fonctions actuelles ?

Encrer le chiffre correspondant à votre réponse, 1 = « Pas du tout utile » et 10 = « Très utile – aideraient grandement les rôles et responsabilités qui m'incombent »

0	1	2	3	4	5	6	7	8	9	10
Pas du tout utile		→								Très utile

VI-2. Dans quelle mesure estimez-vous qu'une telle transformation pourrait être utile au domaine de la santé publique en général ?

Encrer le chiffre correspondant à votre réponse, 1 = « Pas du tout utile » et 10 = « Très utile – aideraient grandement les rôles et responsabilités qui m'incombent »

0	1	2	3	4	5	6	7	8	9	10
Pas du tout utile		→								Très utile

VI-3. Imaginez que vous êtes un dépositaire de données et qu'une méthode a été mise au point pour récupérer les renseignements nominatifs, les « masquer » ou les modifier d'une certaine façon, tout en les préservant au niveau individuel, individu par individu. Est-ce que vous autoriseriez que l'on utilise une telle méthode sur vos données, pour que le résultat puisse être partagé avec d'autres professionnels de la santé publique aux fins de la recherche et de la pratique en santé publique ?

OUI       NON → Préciser pourquoi :     Peut-être → Préciser pourquoi :

Une maladie ou un état de santé spécifique serviront à mettre à l'essai et à évaluer la ou les méthodes élaborées. La maladie ou l'état de santé choisi devra posséder une étiologie connue, avec des modèles et des liens bien connus, à fin de servir comme point de départ à la recherche. La maladie ou l'état de santé choisi devra également être d'un intérêt certain pour la communauté de la santé publique.

VI-4. Quelle maladie, quel état de santé, ou quelle base de données vous viennent spontanément à l'esprit à titre de sujets potentiels aux fins de cette recherche ?

VI-5. Suivant vos connaissances au sujet de l'état de santé ou de la maladie que vous avez signalés dans votre réponse à la question précédente, quels liens à l'environnement physique une *transformation* (telle que définie au paragraphe introductif de la présente section) devra-t-elle conserver à fin que les données demeurent pertinentes et utiles à vos fins. Par exemple : Où se trouvent les patients ou les cas les uns par rapport aux autres, par rapport aux écoles, restaurants, etc.; où les écoles se trouvent par rapport à un type d'industrie, etc. ?

Une autre solution proposée à la problématique consiste à construire ce que l'on appelle des *agents logiciels* automatisés. Il s'agit d'applications qui s'insèrent dans l'ensemble de données visé, là où il est hébergé (i.e., à l'établissement du dépositaire des données), exécutent les analyses pour vous (*sur les renseignements nominatifs*) et vous retournent uniquement les résultats agrégés, lesquels sont ainsi rendus anonymes. En d'autres termes, vous ne verrez jamais les données comme tel, mais « l'agent » effectuera les analyses pour vous, directement sur les renseignements nominatifs, et vous obtenez uniquement les résultats de l'analyse, dans la mesure où ils ne compromettent pas la protection des renseignements personnels. Par analogie, c'est comme si vous me donniez une équation ou une fonction à exécuter avec les données que je possède, et que je vous remettrais ensuite le résultat de cette fonction sans que vous ayez besoin de voir mes données. En supposant que les dépositaires de données permettent qu'un tel *agent logiciel* analyse leurs données et rendent disponibles les résultats de l'analyse à la communauté des professionnels de la santé publique ...

VI-6. Dans quelle mesure un tel agent logiciel vous serait-il utile dans vos fonctions actuelles ?  
Encercler le chiffre correspondant à votre réponse, 1 = « Pas du tout utile » et 10 = « Très utile – aideraient grandement les rôles et responsabilités qui m'incombent »

0	1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	---	----

Pas du tout utile ▶ Très utile

VI-7. Dans quelle mesure estimez-vous qu'une tel agent logiciel pourrait être utile au domaine de la santé publique en général ?  
Encercler le chiffre correspondant à votre réponse, 1 = « Pas du tout utile » et 10 = « Très utile – aideraient grandement les rôles et responsabilités qui m'incombent »

0	1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	---	----

Pas du tout utile ▶ Très utile

VI-8. Si vous étiez (ou êtes) un dépositaire de données, est-ce que vous autoriseriez un tel agent logiciel à accéder à vos données, à procéder à leur analyse, et à en remettre les résultats à la communauté des professionnels de la santé publique à des fins d'analyse et de recherche ?

OUI                       NON → Préciser pourquoi :     Peut-être → Préciser pourquoi :

VI-9. En résumé, si une solution était trouvée de telle sorte que la protection des renseignements personnels ne serait plus problématique, quelle méthode est-ce que vous préféreriez utiliser ?

(Veuillez ne cocher qu'un seul choix)

- Je préférerais pouvoir travailler directement avec les données brutes, à fin de pouvoir accéder aux renseignements au cas par cas.
  
- Je n'ai pas besoin de voir les données brutes, et je préférerais alors avoir accès aux informations et aux résultats en leur forme agrégée.

SECTION VII – Aspects qualitatifs

(~ 5 minutes)

VII-1. Comment considérez-vous l'état de vos connaissances sur les questions et les lois se rapportant à la protection et la confidentialité des renseignements personnels ?

Encercler le chiffre correspondant à votre réponse, 1 = « Aucune connaissance particulière à ce sujet » et 10 = « Expert »

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

Aucune connaissance → Expert

VII-2. Que pensez-vous de l'impact des lois en matière de protection des renseignements personnels et de confidentialité sur la santé publique – en particulier en ce qui concerne les restrictions à l'accès aux renseignements nominatifs (*Loi sur la protection des renseignements personnels, Loi sur la protection des renseignements personnels et les documents électroniques, etc.*) ?

VII-3. Que pensez-vous de la recherche proposée (mise au point d'une méthode de *transformation*) ?

VII-4. Que pensez-vous de l'idée de l'*agent logiciel* ?

VII-5. Avez-vous des idées ou des commentaires à formuler au sujet de cette problématique, de la recherche proposée ou de ce questionnaire ?

## SECTION VIII – Participation future et communications

Veillez indiquer ici le degré d'anonymat que vous souhaitez et votre intérêt à participer à la présente initiative. Vous pouvez cocher plusieurs cases au besoin. Veillez noter qu'en laissant la présente section en blanc, votre réponse sera considérée par défaut comme étant « anonymat absolu », rendant vos réponses au présent questionnaire non identifiables au plan individuel et vous soustrayant à toute communication ou participation future à ce sujet.

- Vous pouvez lier mon identité à mes réponses au présent questionnaire aux fins de clarification et de suivi
- Veuillez me transmettre un sommaire des résultats du présent questionnaire une fois l'analyse achevée
- Veuillez me transmettre des suivis périodiques sur la progression de cette recherche, par courriel à l'adresse électronique indiquée ci-dessous
- Je suis intéressé (e) à participer aux essais pilote des résultats de cette recherche

Veillez indiquer vos coordonnées à l'endroit prévu ci-après. Vos renseignements personnels seront conservés dans un répertoire protégé par mot de passe au sein de l'Agence de santé publique du Canada, et seront protégés conformément aux dispositions de la *Loi sur l'accès à l'information* et de la *Loi sur la protection des renseignements personnels*.

**Nom:** \_\_\_\_\_

**Titre:** \_\_\_\_\_

**Organisation:** \_\_\_\_\_

**Adresse:** \_\_\_\_\_

**Courriel:** \_\_\_\_\_

**Téléphone:** \_\_\_\_\_

**Méthode de communication préférée :**

- Téléphone       Télécopieur       Courriel       Par la poste

Êtes-vous en possession de renseignements nominatifs que vous pourriez utiliser aux fins des essais et de l'évaluation de la méthode de transformation mise au point dans le cadre de ce projet ?

- OUI       NON

Cher(e) professionnel(le) de santé publique,

Grand merci d'avoir pris le temps de remplir ce questionnaire. Vos réponses aideront à évaluer l'impact de la législation de protection de renseignements personnels et de confidentialité sur la recherche en santé publique, et seront employées pour étudier et développer des solutions spécifiques à des maladies précises. En bout de ligne, vous nous aidez à augmenter la qualité des décisions stratégiques et de la recherche en la santé publique.

Encore une fois, je tiens à vous remercier et j'anticipe d'avoir l'occasion dans le futur d'explorer à fond ces questions avec vous et d'ainsi améliorer la pratique de la santé publique. N'hésitez pas à m'appeler ou à m'envoyer un courriel, pour toute question ou commentaire.

Bien à vous,

Philip AbdelMalik

### **B.3. PUBLIC HEALTH PROFESSIONAL QUESTIONNAIRE UNITED KINGDOM**

**ESTIMATED TIME TO COMPLETE THIS QUESTIONNAIRE IS 20 TO 30 MINUTES**

Thank you for taking the time to complete this questionnaire – your contribution is highly valued and critical to the research.

This questionnaire is divided into eight short sections to better categorize and organize the type of information being captured. You may skip any question(s) you would rather not answer, however please keep in mind that this compromises the quality of the research. Please also note that, unless you choose to complete section VIII of the questionnaire, your responses will remain anonymous and cannot be linked back to you. Therefore, please ensure that all responses are clearly marked.

Please note that your responses will vest in, and remain the property of Her Majesty the Queen in Right of Canada, and as such, may also be made available to and used by the Public Health Agency of Canada to improve their operations and service provision to the public health community. As an Office of Public Health Practice, Public Health Agency of Canada survey, all data collected, including your personal information, will be protected according to the Access to Information Act and the Privacy Act in Canada.



SECTION I – A little about you...

(~ 5 minutes)

I-1 What would you say is the scope of the **bulk** of your involvement in public health?  
(If more than one, please select only your **main** involvement)

- International     European     National     Regional     Local

I-2 With which public health organization are you currently employed / affiliated?  
(If more than one, please select only your **main** organization)

- Department of Health  
 Health Protection Agency  
 National Health Service [NHS; include Strategic Health Authorities (SHA) and other regional or central organizations]  
 NHS Trust (including Care Trust, Hospital Trust, Mental Health Trust, etc.)

Please specify: \_\_\_\_\_

- Special Health Authority  
 Association of Public Health Observatories (APHO)  
 Public Health Observatory (please specify):  
 EMPHO     LHO     ScotPHO     WCH  
 ERPHO     NEPHO     SEPHO     WMPHO  
 INIsPHO     NWPHO     SWPHO     YHPHO  
 Public Health Faculty  
 University / Academia  
 Other

Please specify: \_\_\_\_\_

I-3. Please indicate your current specific area(s) of expertise: (Check as many as apply)

- |  |  |
|--|--|
| <input type="checkbox"/> Chronic Diseases (cancer, diabetes, etc.)                   | <input type="checkbox"/> Genetics  |
| <input type="checkbox"/> Child / Paediatric Public Health                            | <input type="checkbox"/> Health Services (needs, delivery, etc.)   |
| <input type="checkbox"/> Communicable Diseases                                       | <input type="checkbox"/> Injuries / Disability   |
| <input type="checkbox"/> Dental Public Health  | <input type="checkbox"/> Mental Health & Substance Misuse  |
| <input type="checkbox"/> Emergency Preparedness & Response                           | <input type="checkbox"/> Occupational Health   |
| <input type="checkbox"/> Environment (pollution, climate, water & food safety, etc.) | <input type="checkbox"/> Social Determinants of Health (e.g. poverty, education, social exclusion, etc.) |
| <input type="checkbox"/> Ethics, Public Health Law, Privacy, etc                     | <input type="checkbox"/> Surveillance  |
| <input type="checkbox"/> Food & Nutrition  |  |
| <input type="checkbox"/> Other   |  |

Please specify: \_\_\_\_\_

I-4. Which of the following **best** describe your roles or functions as a public health professional? (If more than one, please select only your **main** roles)

- |  |   |
|--|---|
| <input type="checkbox"/> Strategic decision / policy maker | <input type="checkbox"/> Research and Analysis                          |
| <input type="checkbox"/> Manager or Coordinator            | <input type="checkbox"/> Front-line responder / patient care / clinical |
| <input type="checkbox"/> Consultant                        |   |
| <input type="checkbox"/> Other                             |   |

Please specify: \_\_\_\_\_

I-5. Thinking of your regular activities, how much of your time (roughly, as a percentage) would you typically spend doing each of the following?

Strategic decision / policy making	_____	%
Management / Coordination	_____	%
Consultation	_____	%
Research / Analysis	_____	%
Front-line response / patient care / clinical	_____	%
Other (as specified in I-4)	_____	%

I-6. In which of the roles you identified above are you **most likely** to use or require the use of personally identifiable data?

- |  |   |
|--|---|
| <input type="checkbox"/> Strategic decision / policy maker | <input type="checkbox"/> Research and Analysis                          |
| <input type="checkbox"/> Manager or Coordinator            | <input type="checkbox"/> Front-line responder / patient care / clinical |
| <input type="checkbox"/> Consultant                        |   |
| <input type="checkbox"/> Other                             |   |

Please specify: \_\_\_\_\_

I-7. Do you have or foresee a need for including geographic location of health data in your roles or organization?

YES

NO

I-8. Geographic Information Systems (GIS) are tools that allow you to visualise and analyse your data spatially – that is, using their geographical location on earth. In which of the roles you identified above would GIS be useful?

- |  |   |
|--|---|
| <input type="checkbox"/> Strategic decision / policy maker | <input type="checkbox"/> Research and Analysis                          |
| <input type="checkbox"/> Manager or Coordinator            | <input type="checkbox"/> Front-line responder / patient care / clinical |
| <input type="checkbox"/> Consultant                        |   |
| <input type="checkbox"/> Other                             |   |

Please specify: \_\_\_\_\_

I-9. What GIS application(s) do you currently use, or have you used in the past?

Web-based: specify \_\_\_\_\_

Desktop GIS products:

- ESRI ArcGIS products
- MapInfo
- AutoDesk products
- PCI Geomatics products
- Intergraph products
- Other

Please specify: \_\_\_\_\_

- I have never used any GIS applications, and have no use for them
- I have never used any GIS applications, but am interested in learning more

I-10 At what level(s) of geography do you visualise your data and/or conduct spatial analyses for each product you use?

	Web-Based	Desktop GIS
Latitude and Longitude	<input type="checkbox"/>	<input type="checkbox"/>
Street address	<input type="checkbox"/>	<input type="checkbox"/>
Postcode	<input type="checkbox"/>	<input type="checkbox"/>
Community Name	<input type="checkbox"/>	<input type="checkbox"/>
City / Town / Village	<input type="checkbox"/>	<input type="checkbox"/>
Region / Geographic Area	<input type="checkbox"/>	<input type="checkbox"/>
Urban – Rural	<input type="checkbox"/>	<input type="checkbox"/>
Other		

Please specify: \_\_\_\_\_

I-11 Are you or have you been restricted in your use of GIS for any public health activity because of privacy concerns (i.e. map or data might identify an individual or community)?

YES → Please explain

No → Please explain

I-12 Setting privacy issues aside and in light of your response to the previous question, at what level(s) of geography would you **ideally** like to visualise your data and/or conduct spatial analyses for each product you use?

	Web-Based	Desktop GIS
Latitude and Longitude	<input type="checkbox"/>	<input type="checkbox"/>
Street address	<input type="checkbox"/>	<input type="checkbox"/>
Postcode	<input type="checkbox"/>	<input type="checkbox"/>
Urban – Rural	<input type="checkbox"/>	<input type="checkbox"/>
Other		

Please specify: \_\_\_\_\_

SECTION II – Current access to data

(~ 4 minutes)

The questions in this section all pertain to the role you identified in question I-6 in Section I. If you do **not** have access to any of the *Personally Identifiable Data (PID)* listed in question II-1, please mark the last option in question II-1 and skip to Section III – No current access to data. Otherwise, please complete this section, and then skip to Section IV – Privacy Issues.

**NOTE:** The term “access” as used in this survey implies the ability to actually acquire individual level data so you can work with it directly.

II-1. What PID do you currently have access to? (Check as many as apply)

- |   |   |
|---|---|
| <input type="checkbox"/> Forename(s)                      | <input type="checkbox"/> Street Address           |
| <input type="checkbox"/> Surname                          | <input type="checkbox"/> Postcode                 |
| <input type="checkbox"/> Initials                         | <input type="checkbox"/> Community Name           |
| <input type="checkbox"/> Sex                              | <input type="checkbox"/> City / Town / Village    |
| <input type="checkbox"/> Date of Birth / Age              | <input type="checkbox"/> Region / Geographic Area |
| <input type="checkbox"/> Date of Death                    | <input type="checkbox"/> Latitude & Longitude     |
| <input type="checkbox"/> NHS Number (OLD)                 |   |
| <input type="checkbox"/> NHS Number (NEW)                 |   |
| <input type="checkbox"/> Registered GP / Family Physician |   |

Other  
Please specify:

\_\_\_\_\_

I do NOT currently have access to any of the above (please skip to Section III)

The following questions all pertain to the *PID* you have access to, as identified in the previous question. For all scales, circle ‘0’ if you “Don’t know”

II-2 From a privacy and organisational bureaucracy perspective, how easy would you say it is for you to access this *PID* when you need it?

Please circle the appropriate number, 1 being “Extremely difficult”, and 10 being “Very easy”

	Don't Know	Extremely difficult → Very easy									
		1	2	3	4	5	6	7	8	9	10
Forename(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Surname	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- Old	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- New	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Post Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

II-3. On average, how often do you access the *PID* you identified above?

		Rarely <span style="float: right;">→</span> All the time									
	Not Applicable	1	2	3	4	5	6	7	8	9	10
Forename(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Surname	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- Old	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- New	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Post Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

II-4. How useful / important is this *PID* to you and your roles and responsibilities?  
Please circle the appropriate number, with 1 being “Not at all useful”, and 10 being “Critical to my roles and responsibilities”

		Not useful <span style="float: right;">→</span> Critical									
	Don't Know	1	2	3	4	5	6	7	8	9	10
Forename(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Surname	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- Old	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- New	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Post Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

II-5. What impact would removal of your access to this *PID* have on the quality of your work and resulting public health decisions?  
 Please circle the appropriate number, with 1 being “No impact – quality would not suffer”, and 10 being “Severe Impact - results and decisions would be severely compromised”

	Don't Know	No impact  Severe impact									
		1	2	3	4	5	6	7	8	9	10
Forename(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Surname	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- Old	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- New	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Post Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

II-6. What *PID* do you currently **NOT** have access to, but believe would be beneficial to you to further enhance your work and resulting public health decisions?  
 (Check as many as apply)

- |   |   |
|---|---|
| <input type="checkbox"/> Forename(s)                      | <input type="checkbox"/> Street Address           |
| <input type="checkbox"/> Surname                          | <input type="checkbox"/> Postcode                 |
| <input type="checkbox"/> Initials                         | <input type="checkbox"/> Community Name           |
| <input type="checkbox"/> Sex                              | <input type="checkbox"/> City / Town / Village    |
| <input type="checkbox"/> Date of Birth / Age              | <input type="checkbox"/> Region / Geographic Area |
| <input type="checkbox"/> Date of Death                    | <input type="checkbox"/> Latitude & Longitude     |
| <input type="checkbox"/> NHS Number (OLD)                 |   |
| <input type="checkbox"/> NHS Number (NEW)                 |   |
| <input type="checkbox"/> Registered GP / Family Physician |   |

Other  
 Please specify: \_\_\_\_\_

None

Please skip to Section IV – Privacy Issues

If you have access to *Personally Identifiable Data (PID)* and completed Section II above, then please skip to Section IV – Privacy Issues.

**NOTE:** The term “access” as used in this survey implies the ability to actually acquire individual level data so you can work with it directly.

III-1. Having access to which of the following *PID* would facilitate your roles and responsibilities, or enhance your work and improve resulting public health decisions? (Check as many as apply)

- |   |   |
|---|---|
| <input type="checkbox"/> Forename(s)                      | <input type="checkbox"/> Street Address           |
| <input type="checkbox"/> Surname                          | <input type="checkbox"/> Postcode                 |
| <input type="checkbox"/> Initials                         | <input type="checkbox"/> Community Name           |
| <input type="checkbox"/> Sex                              | <input type="checkbox"/> City / Town / Village    |
| <input type="checkbox"/> Date of Birth / Age              | <input type="checkbox"/> Region / Geographic Area |
| <input type="checkbox"/> Date of Death                    | <input type="checkbox"/> Latitude & Longitude     |
| <input type="checkbox"/> NHS Number (OLD)                 |   |
| <input type="checkbox"/> NHS Number (NEW)                 |   |
| <input type="checkbox"/> Registered GP / Family Physician |   |
| <input type="checkbox"/> Other                            |   |
| Please specify:   |   |
| _____   |   |
| <input type="checkbox"/> None                             |   |

III-2. How useful to you and your roles and responsibilities (as identified in Section I) would access to the *PID* you identified above be?  
Please circle the appropriate number, with 1 being “Not at all useful”, and 10 being “Very useful – would greatly enhance by roles and responsibilities”

		Not useful <span style="float: right;">→</span> Very useful									
	Don't Know	1	2	3	4	5	6	7	8	9	10
Forename(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Surname	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- Old	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- New	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Post Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

III-3 How easy would it be for you to access the *PID* you identified above, if you were to need it?

Please circle the appropriate number, with 1 being “Impossible”, and 10 being “Very easy”

	Don't Know	Impossible <span style="float: right;">→</span> Very easy									
		1	2	3	4	5	6	7	8	9	10
Forename(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Surname	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- Old	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- New	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Post Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

III-4. What impact has your lack of access to this *PID* had on the quality of your work and resulting public health decisions?

Please circle the appropriate number, with 1 being “No impact – quality has not suffered”, and 10 being “Severe Impact - results and decisions have been severely compromised”

	Don't Know	No impact <span style="float: right;">→</span> Severe impact									
		1	2	3	4	5	6	7	8	9	10
Forename(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Surname	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Initials	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of Birth / Age	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date of death	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- Old	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
NHS Number- New	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Registered GP / Family Physician	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Street Address	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Post Code	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Community Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
City / Town / Village	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Region / Geographic Area	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latitude / Longitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



SECTION IV – Privacy Issues

(~ 3 minutes)

This section pertains to the field of public health in general, and uses the term “*public health practice*” to refer to its various activities, including research, surveillance, health service delivery, strategic policy and decision making, etc. The goal is to get your opinion, as a public health professional, on the overall impact of restricted access to *PID* on *public health practice* in the United Kingdom. These questions ask for your opinion; if you’re not sure how to answer a question in this section, please just hazard a guess!

**NOTE:** The term “access” as used in this survey implies the ability to actually acquire individual level data so you can work with it directly.

IV-1. In your opinion, do current restrictions on access to *PID* pose an obstacle to any aspects of *public health practice* (e.g. research, surveillance, etc.)?

Please circle the appropriate number, with 1 being “Not an obstacle at all”, and 10 being “Yes, they pose a serious threat to accurate public health practice”

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

Not an issue → Serious Threat

IV-2. How amenable would **you** be to other professionals in the public health field having access to **your** *PID* for public health research and analyses (e.g. **your** address, date of birth, etc.) to improve public health delivery, service, etc.?

Sure, go ahead                       No Way                       Not Sure → Please Explain

IV-3a. In your opinion, what proportion of the public is aware of the impact of restricted access to *PID* on public health practice? (Please just guess!)

Please circle the approximate proportion

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
-----	-----	-----	-----	-----	-----	-----	-----	-----	------

No one → Everyone

IV-3b. How do you think we could increase this proportion?

IV-4a. In your opinion, what proportion of the public would allow the use of *PID* for public health practice if they were asked and educated on the usefulness of such data to public health practice? (Please just guess!) - Please circle the approximate proportion

10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
-----	-----	-----	-----	-----	-----	-----	-----	-----	------

No one → Everyone

IV-4b. How do you think we could increase this proportion?

SECTION V – Current data holdings and provision to others...

(~ 2 minutes)

This section gathers information on the sharing of PID within and between organizations.

V-1. What would you say is the one **most** critical obstacle in the sharing or acquisition of PID linked to health data? (Please select only one; give your opinion!)

- National legislation     Public disapproval     Organizational bureaucracy  
 Lack of knowledge     Public Paranoia     Practitioner Paranoia  
 Other (please specify): \_\_\_\_\_

V-2. Do you or your organization currently collect individual-level health data for any purpose (e.g. research, surveillance, service delivery, etc.), or act as the custodian of such data?

- YES → Continue                       NO → Go to Section VI

V-3. For what specific purpose(s) is this data collected? (check as many as apply)

- Research                       Surveillance                       Service Delivery  
 Other  
 Please specify: \_\_\_\_\_

V-4. What data is collected?

- Forename(s)                       Street Address  
 Surname                       Postcode  
 Initials                       Community Name  
 Sex                       City / Town / Village  
 Date of Birth / Age                       Region / Geographic Area  
 Date of Death                       Latitude & Longitude  
 NHS Number (OLD)  
 NHS Number (NEW)  
 Registered GP / Family Physician  
  
 Other  
 Please specify: \_\_\_\_\_

V-5. How difficult is it for other public health professionals such as yourself to acquire access to your PID and linked health data holdings if they are outside your immediate working team, but...

		Impossible <span style="float: right;">→</span>							Very Easy		
	D/K	1	2	3	4	5	6	7	8	9	10
Within your organisation?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Within the NHS or Department of Health?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Within the UK but outside of the NHS or Dept. of Health?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Within the European Union outside the UK?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Outside the European Union?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Within the World Health Organization?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

The proposed research will seek to apply a method (called a *transformation*) to public health data such that important relationships within and between the data are preserved, but the actual identity of the individual is anonymized. So, for example, if you were looking at an infectious outbreak in children, you might be interested in where the infected children are relative to one another, as well as where the schools are, arenas, community centres, etc. You would then preserve the *relationship* between these points of interest, and change everything else, so that the original points can no longer be identified back to their original owners. In this way, you have *transformed* the data so that you're still looking at individual-level data, but can't determine who it belongs to (i.e., it has become anonymous). Assuming the data custodians allow the data derived from such a *transformation* to be made available to the public health professional community:

For all scales, circle '0' if you "Don't know"

VI-1. How useful would such a transformation be to you in your current role?

Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

0	1	2	3	4	5	6	7	8	9	10	
Not useful		→									Very useful

VI-2. How useful do you think such a transformation would be to the field of public health in general?

Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

0	1	2	3	4	5	6	7	8	9	10	
Not useful		→									Very useful

VI-3. Imagine you are a data custodian, and that a method has been developed to take your individual level data and mask it or change it somehow, while still keeping it at an individual-by-individual level. Would you allow such a method to be conducted on your data so that it can be shared with other public health professionals for public health research and practice?

YES

NO → Please explain why not

MAYBE → Please explain

A specific disease or health condition will be used to test and evaluate the developed method(s). This condition must have a known aetiology, with well-known patterns and relationships, to serve as a starting point for the research. It must also be a disease of interest to the public health community.

VI-4. What diseases, health conditions, or databases most immediately come to mind as potential subjects for this research?

VI-5. Based on your knowledge of the condition you identified in the previous question, what relationships to the physical environment would a *transformation* as defined in the opening paragraph of this section have to retain in order for the data to be meaningful and useful to you (e.g. where patients or cases are relative to each other, to schools, to restaurants, etc; where schools are relative to a type of industry; etc...)?

Another proposed solution to the issue at hand is to build what are called automated *software agents*. You can think of these as applications that would go into a dataset wherever it is housed (i.e. at the custodian's location), perform the analyses for you (on the *personally identifiable data*) and return only the aggregated, and therefore anonymised, results. In other words, you would never see the actual data, but would have this "agent" do the analyses for you, directly on the *PID*; you simply get the results of the analyses, as long as, of course, they don't compromise privacy. As a simple analogy, it would be like you giving me an equation or function to perform on my data, and I giving you back the result of that function without you ever needing to see my actual data. Assuming the data custodians allow such a *software agent* to analyse their data and make the results available to the public health professional community:

VI-6. How useful do you think such a *software agent* would be to you in your current role?  
Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

0	1	2	3	4	5	6	7	8	9	10	
Not useful										Very useful	

VI-7. How useful do you think such a *software agent* would be to the field of public health in general?  
Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

0	1	2	3	4	5	6	7	8	9	10	
Not useful										Very useful	

VI-8. If you were (or are) a data custodian, would you allow such a *software agent* to access your data, conduct the analyses, and return the results to the public health professional community for research and analysis?

- YES     
  NO → Please explain why not     
  MAYBE → Please explain

VI-9. To summarise, if a solution is found such that privacy is no longer an issue, which of the following would you prefer? (Please select only one)

- I would prefer to be able to work directly with the raw data, so I can access information on a case-by-case basis.
- I have no need to see the raw data, and would prefer to access information and results on an aggregate basis.

SECTION VII – Qualitative Component

(~ 5 minutes)

VII-1. How knowledgeable do you consider yourself on privacy and confidentiality issues / legislation?

Please circle the appropriate number, 1 being “Not at all knowledgeable”, and 10 being “Expert”

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

Not Knowledgeable → Expert

VII-2. How do you feel about the impact of privacy and confidentiality legislation – in particular the restrictions on access to personally identifiable data (e.g. The Data Protection Act, The Caldicott Guidelines) – on public health?

VII-3. What do you think of the proposed research (development of a *transformation*)?

VII-4. What do you think of the “*software agent*” idea?

VII-5. Do you have any other thoughts or comments regarding this issue, the proposed research, or this questionnaire that you would like to share?

## SECTION VIII – Further Participation and Contact

Please indicate your desired level of anonymity and interest in further participation; you may check multiple boxes as applicable. Please note that leaving this section empty will default your response to “absolute anonymity”, rendering your answers on this questionnaire personally unidentifiable and removing yourself from any further contact or participation.

- You may link my identity to my responses on this questionnaire for clarifications and follow-up
- Please send me a summary of the findings of this questionnaire, once completed
- Please send me periodic updates on the progress of this research by email to the address given below
- I am interested in piloting the results of this research

If you checked any of the above boxes, please complete your details below. Your personal information will be stored in a password-protected directory within the Public Health Agency of Canada, and will be protected according to the Access to Information Act and the Privacy Act.

**Name:** \_\_\_\_\_

**Title:** \_\_\_\_\_

**Organization:** \_\_\_\_\_

**Address:** \_\_\_\_\_  
\_\_\_\_\_

**E-Mail:** \_\_\_\_\_

**Phone:** \_\_\_\_\_

**Preferred Method of Contact:**  Phone  Fax  E-Mail  Mail

Are you in possession of any personally identifiable data that you can use for testing and evaluation of the developed transformation?

YES  NO

Dear Public Health Professional,

Thank you so much for taking the time to complete this questionnaire; your responses will help assess the impact of privacy and confidentiality legislation on public health research, and will be used to investigate and develop a disease-specific solution, which will, in turn, enhance strategic decisions and research in public health.

Once again, many thanks for your time, and I look forward to enhancing public health practice by exploring this issue, and its solution, further with you. Should you have any comments, questions or concerns, please feel free to send me an email or give me a ring.

Best wishes,

Philip AbdelMalik

## C. Full survey findings

### SECTION I – A little about you...

I-1 What would you say is the scope of the **bulk** of your involvement in public health?  
(If more than one, please select only your **main** involvement)

#### CANADA

Scope	Frequency	Percent	Cumulative Frequency	Cumulative Percent
National	19	28.79	19	28.79
North American	2	3.03	21	31.82
Provincial / Territorial	16	24.24	37	56.06
Regional / Local	29	43.94	66	100.00

#### UK

Scope	Frequency	Percent	Cumulative Frequency	Cumulative Percent
European	1	3.57	1	3.57
Local	7	25.00	8	28.57
National	2	7.14	10	35.71
Regional	17	60.71	27	96.43
Skip	1	3.57	28	100.00

I-2 With which public health organization are you currently employed / affiliated?  
(If more than one, please select only your **main** organization)

#### CANADA

Organisation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Health Canada	1	1.52	1	1.52
Independent planning consultant	1	1.52	2	3.03
Newfoundland & Labrador Centre for Health Information	1	1.52	3	4.55
Other federal government agency	1	1.52	4	6.06
Other non-government association	5	7.25	9	13.64
Primary care networks & a regional health authority contract	1	1.52	10	15.15
Private consultant	1	1.52	11	16.67
Provincial government	10	15.15	21	31.82
Public Health Agency of Canada	15	22.73	36	54.55
Regional / local health authority	27	40.91	63	95.45
University / Academia	3	4.55	66	100.00

**UK (a)**

Organisation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Association of Public Health Observatories	2	7.14	2	7.14
Department of Health	1	3.57	3	10.71
Health Board	1	3.57	4	14.29
NHS	2	7.14	6	21.43
PCT	5	17.86	11	39.29
Public Health Observatory*	13	46.43	24	85.71
Special Health Authority	1	3.57	25	89.29
University / Academia	3	10.71	28	100.00

**\*UK (b)**

Observatory	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ERPHO	3	23.08	3	23.08
NEPHO	1	7.69	4	30.77
NWPHO	1	7.69	5	38.46
SEPHO	2	15.38	7	53.85
SWPHO	5	38.46	12	92.31
YHPHO	1	7.69	13	100.00

**I-3** Please indicate your current specific area(s) of expertise: (Check as many as apply)

**CANADA**

Area of Expertise	Freq	% (n=64)
Aboriginal Health	10	15.63
Chronic Diseases	18	28.13
Child / Paediatric Public Health	14	21.88
Communicable / Infectious	30	46.88
Dental Public Health	3	4.69
Emergency Prep & Response	12	18.75
Environment	18	28.13
Ethics, PH Law, Privacy, etc	2	3.13
Food & Nutrition	3	4.69
Genetics	0	0.00
Health Services	16	25.00
Injuries / Disability	6	9.38
Mental Health & Sub Misuse	4	6.25
Occupational Health	0	0.00
Social Determinants of Health	15	23.44
Surveillance	29	45.31
Other	11	17.19
Not Specified	2	3.13

**UK**

Area of Expertise	Freq	% (n=27)
Aboriginal Health	0	0.00
Chronic Diseases	6	22.22
Child / Paediatric Public Health	5	18.52
Communicable / Infectious	2	7.41
Dental Public Health	2	7.41
Emergency Prep & Response	1	3.70
Environment	4	14.81
Ethics, PH Law, Privacy, etc	1	3.70
Food & Nutrition	2	7.41
Genetics	0	0.00
Health Services	4	14.81
Injuries / Disability	1	3.70
Mental Health & Sub Misuse	3	11.11
Occupational Health	0	0.00
Social Determinants of Health	12	44.44
Surveillance	4	14.81
Other	13	48.15
Not Specified	1	3.70



**I-4** Which of the following best describe your roles or functions as a public health professional?  
(If more than one, please select only your main roles)

**CANADA**

Main Role	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Consultant	5	7.58	5	7.58
Educator	1	1.52	6	9.09
Front Line	2	3.03	8	12.12
Health Protection	1	1.52	9	13.64
Manager or Coordinator	16	24.24	25	37.88
Public Educator, GIS Project Manager, Health Researcher	1	1.52	26	39.39
Research & Analysis	32	48.48	58	87.88
Senior Epidemiologist	1	1.52	59	89.39
Strategic Decision / Policy Maker	6	9.09	65	98.48
Surveillance & Content Expert Advisor	1	1.52	66	100.00

**UK**

Main Role	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Consultant	2	7.14	2	7.14
Manager or Coordinator	2	7.14	4	14.29
Research & Analysis	18	64.29	22	78.57
Specialist in Public Health Intelligence	1	3.57	23	82.14
Strategic Decision / Policy Maker	5	17.86	28	100.00

**I-5** Thinking of your regular activities, how much of your time (roughly, as a percentage) would you typically spend doing each of the following?

**CANADA**

Main Role	N, Mean, (STD) & Range percentage of time spent on activity					
	1	2	3	4	5	6
1 – Strategic Decision / Policy Maker	6 44.17 (29.23) 15 – 100	5 41.00 (13.42) 30 – 60	5 16.80 (4.60) 10 – 20	5 8.00 (4.47) 5 – 15	2 3.00 (2.83) 1 – 5	0
2 – Manager / Coordinator	15 18.67 (12.60) 10 – 55	16 51.25 (17.84) 20 – 90	12 18.00 (14.10) 1 – 50	10 19.00 (13.70) 5 – 50	6 5.33 (4.03) 1 – 10	2 10.50 (13.44) 1 – 20
3 – Consultant	2 17.50 (10.61) 10 – 25	2 32.50 (24.75) 15 – 50	5 47.00 (25.40) 10 – 80	5 31.00 (16.73) 10 – 50	1 10.00	0
4 – Research & Analysis	16 7.63 (3.10) 5 – 15	16 20.06 (12.94) 1 – 50	24 17.83 (8.55) 5 – 33	32 68.22 (22.05) 33 – 100	4 6.50 (4.36) 1 – 10	6 14.17 (15.83) 1 – 35
5 – Front-Line	1 5.00	2 22.50 (17.68) 10 – 35	2 15.00 (7.07) 10 – 20	2 15.00 (0)	2 30.00 (0)	1 30.00
6 – Other	4 10.00 (4.08) 5 – 15	3 30.00 (17.32) 20 – 50	4 15.00 (7.07) 10 – 25	4 18.75 (6.29) 10 – 25	2 12.50 (10.61) 5 – 20	4 48.75 (27.20) 20 – 85

UK

Main Role	N, Mean, (STD) & Range percentage of time spent on activity					
	1	2	3	4	5	6
1 – Strategic Decision / Policy Maker	5 56.00 (5.48) 50 – 60	5 28.00 (8.37) 20 – 40	3 10.00 (0) 10 – 20	4 10.00 (0) 10 – 20	1 10.00 (0) 10 – 20	0
2 – Manager / Coordinator	2 17.50 (17.68) 5 – 30	2 55.00 (21.21) 40 – 70	2 15.00 (7.07) 10 – 20	2 10.00 (0) 10 – 20	1 5.00 (0) 10 – 20	0
3 – Consultant	2 30.00 (14.14) 20 – 40	2 25.00 (7.07) 20 – 30	2 20.00 (14.14) 10 – 30	2 15.00 (7.07) 10 – 20	1 10.00 (0) 10 – 20	1 10.00 (0) 10 – 20
4 – Research & Analysis	5 22.00 (11.51) 10 – 40	11 23.64 (14.85) 5 – 50	5 6.00 (2.24) 5 – 10	17 66.18 (29.61) 10 – 100	1 5.00 (0) 10 – 20	2 22.50 (17.68) 10 – 35
5 – Front-Line	0	0	0	0	0	0
6 – Other	1 0.00	1 10.00	1 20.00	1 30.00	0	1 40.00

I-6 In which of the roles you identified above are you **most likely** to use or require the use of personally identifiable data?

CANADA

Main Role	Role in which most likely to use PID (n, %)								
	0	1	2	3	4	5	6	99	
1 – Strategic Decision / Policy Maker		2 33.33		1 16.67	2 33.33	1 16.67			
2 – Manager / Coordinator		1 6.25	3 18.75	2 12.50	6 37.50	2 12.50		2 12.50	
3 – Consultant				3 60.00	2 40.00				
4 – Research & Analysis	1 3.13				27 84.38	1 3.13		3 9.38	
5 – Front-Line					1 50.00	1 50.00			
6 – Other					4 80.00		1 20.00		
<b>TOTALS</b>	<b>1</b> 1.52	<b>3</b> 4.55	<b>3</b> 4.55	<b>6</b> 9.09	<b>42</b> 63.64	<b>5</b> 7.58	<b>1</b> 1.52	<b>5</b> 7.58	

UK

Main Role	Role in which most likely to use PID (n, %)							
	0	1	2	3	4	5	6	77
1 – Strategic Decision / Policy Maker		2 40.00			3 60.00			
2 – Manager / Coordinator			1 50.00		1 50.00			
3 – Consultant					1 50.00	1 50.00		
4 – Research & Analysis					16 88.89		1 5.56	1 5.56
5 – Front-Line								
6 – Other					1 100.00			
<b>TOTALS</b>		<b>2</b> 7.14	<b>1</b> 3.57		<b>22</b> 78.57	<b>1</b> 3.57	<b>1</b> 3.57	<b>1</b> 3.57

**I-7** Do you have or foresee a need for including geographic location of health data in your roles or organization?

**CANADA**

Response	Frequency	Percent	Cumulative Frequency	Cumulative Percent
YES	66	100.00	66	100.00
NO	0	0.00	0	0.00

**UK**

Response	Frequency	Percent	Cumulative Frequency	Cumulative Percent
YES	27	96.43	27	96.43
NO	0	0	27	96.73
No Response	1	3.57	28	100.00

**I-8** Geographic Information Systems (GIS) are tools that allow you to visualise and analyse your data spatially – that is, using their geographical location on earth. In which of the roles you identified above would GIS be useful?

**CANADA**

Role	N
1 – Strategic Decision / Policy Maker	35
2 – Manager / Coordinator	19
3 – Consultant	26
4 – Research & Analysis	55
5 – Front-Line	17
6 – Other	5

**UK**

Role	N
1 – Strategic Decision / Policy Maker	16
2 – Manager / Coordinator	5
3 – Consultant	3
4 – Research & Analysis	24
5 – Front-Line	3
6 – Other	2

**I-9** What GIS application(s) do you currently use, or have you used in the past?

**CANADA**

GIS Application	N
PHMG	17
ESRI	25
MapInfo	19
Other	14
No GIS	15

**UK**

GIS Application	N
PHMG	0
ESRI	10
MapInfo	18
Other	8
No GIS	6

**I-10** At what level(s) of geography do you visualise your data and/or conduct spatial analyses for each product you use?

**CANADA**

Level	PHMG	Other Web-based	Desktop GIS
Latitude & Longitude	5	10	12
Street Address	7	8	12
Dissemination Area	9	7	14
Postal Code	14	14	19
Census Subdivision	12	7	16
Census Division	11	10	14
Forward Sortation Area	7	5	12
Urban – Rural	8	6	13
Provincial	6	11	12
Don't Use	4	1	2
Skipped / Did not respond	37	27	26

**UK**

Level	Other Web-based	Desktop GIS
Latitude & Longitude	3	4
Street Address	2	3
Post Code	14	11
Community Name	7	5
City / Town / Village	5	7

Region / Geographic Area	12	11
Urban – Rural	7	9
Skipped / Did not respond	6	7

**Other identified geographies used:**

CANADA	UK
<ul style="list-style-type: none"> <li>- Nursing Districts</li> <li>- Location/Public Health Units</li> <li>- Emergency Locator Numbers</li> <li>- Census Tracts</li> <li>- Locally defined / custom neighbourhoods and regions</li> <li>- International geographies</li> <li>- Municipalities</li> <li>- Watershed</li> </ul>	<ul style="list-style-type: none"> <li>- Local Authorities</li> <li>- Counties</li> <li>- Output Areas</li> <li>- Primary Care Trusts</li> <li>- Wards</li> <li>- Census Areas</li> <li>- British National Grid</li> <li>- Other administrative boundaries</li> </ul>

**I-11** Are you or have you been restricted in your use of GIS for any public health activity because of privacy concerns (i.e. map or data might identify an individual or community)?

**CANADA**

Main Role	Restricted in GIS use because of privacy? (n, %)		
	NO	YES	Not Specified
1 – Strategic Decision / Policy Maker	1 16.67	3 50.00	2 33.33
2 – Manager / Coordinator	1 6.25	6 37.50	9 56.25
3 – Consultant	0 0.00	3 60.00	2 40.00
4 – Research & Analysis	5 15.63	18 56.25	9 28.13
5 – Front-Line	2 100.00	0 0.00	0 0.00
6 – Other	1 20.00	4 80.00	0 0.00
<b>TOTALS</b>	<b>10 15.15</b>	<b>34 51.52</b>	<b>22 33.33</b>
<b>TOTALS (of those responding only)</b>	<b>10 22.73</b>	<b>34 77.27</b>	

**UK**

Main Role	Restricted in GIS use because of privacy? (n, %)		
	NO	YES	Not Specified
1 – Strategic Decision / Policy Maker	0 0.00	3 60.00	2 40.00
2 – Manager / Coordinator	1 50.00	0 0.00	1 50.00
3 – Consultant	0 0.00	1 50.00	1 50.00
4 – Research & Analysis	2 11.11	10 55.56	6 33.33
5 – Front-Line	0 0.00	0 0.00	0 0.00
6 – Other	0 0.00	1 100.00	0 0.00
<b>TOTALS</b>	<b>3 10.71</b>	<b>15 53.57</b>	<b>10 35.71</b>
<b>TOTALS (of those responding only)</b>	<b>3 16.67</b>	<b>15 83.33</b>	

I-12 Setting privacy issues aside and in light of your response to the previous question, at what level(s) of geography would you *ideally* like to visualise your data and/or conduct spatial analyses for each product you use?

**CANADA**

Minimum granularity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
BLANK	16	24.24	16	24.24
CSD	1	1.52	17	25.76
CT	1	1.52	18	27.27
Cadastre	1	1.52	19	28.79
Custom	1	1.52	20	30.30
DA	5	7.58	25	37.88
DEPENDS	1	1.52	26	39.39
ELN	1	1.52	27	40.91
LatLong	21	31.82	48	72.73
Municipality	1	1.52	49	74.24
PC	3	4.55	52	78.79
PHU	1	1.52	53	80.30
SKIP	3	4.55	56	84.85
Street	9	13.64	65	98.48
SubRegion	1	1.52	66	100.00

LatLong or Street or ELN	31	46.97
--------------------------	----	-------

Of those responding only (n=47):

LatLong or Street or ELN	31	65.96
--------------------------	----	-------

**UK**

Minimum granularity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
BLANK	4	14.29	4	14.29
DEPENDS	1	3.57	5	17.86
Household	1	3.57	6	21.43
LatLong	3	10.71	9	32.14
OTHER	4	14.29	13	46.43
PC	11	39.29	24	85.71
SKIP	1	3.57	25	89.29
Street	13	10.71	28	100.00

LatLong or Street or Household	17	60.71
--------------------------------	----	-------

Of those responding only (n=23):

LatLong or Street or Household	17	73.91
--------------------------------	----	-------

SECTIONS II & III – Access to data

**NOTE:** The term “access” as used in this survey implies the ability to actually acquire individual level data so you can work with it directly.

**II-1** Do you currently have access to PID?

**CANADA**

Response	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	15	22.73	15	22.73
YES	48	72.72	63	95.45
No Response	3	4.55	66	100.00

**UK**

Response	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	4	14.29	4	14.29
YES	23	82.14	27	96.43
No Response	1	3.57	28	100.00

The following results show, where applicable, responses of both, those with access to PID and those without.

**II-2** From a privacy and organisational bureaucracy perspective, how easy would you say it is for you to access this *PID* when you need it?  
Please circle the appropriate number, 1 being “Extremely difficult”, and 10 being “Very easy”

**III-3** How easy would it be for you to access the *PID* you identified above, if you were to need it?  
Please circle the appropriate number, with 1 being “Impossible”, and 10 being “Very easy”

### CANADA

	0	1	2	3	4	5	6	7	8	9	10	77	99
First Name	3	19	6	3	3	2	1	4	4	4	5	0	12
Last Name	3	19	7	2	3	2	1	4	2	3	8	0	12
Initials	10	14	4	5	2	2	2	2	2	2	6	0	15
Sex	5	1	1	1	0	3	0	9	8	8	21	0	9
Date of Birth / Age	4	4	2	3	0	6	1	9	8	5	15	0	9
Date of death	6	6	3	1	4	8	1	8	6	3	9	0	11
Provincial Health Insurance Plan Number	9	20	3	2	1	5	1	5	3	3	2	0	12
Hospital ID	9	11	7	2	2	7	1	5	2	1	7	1	11
Registered GP / Family Physician	3	2	0	1	1	1	0	1	0	0	0	47	10
Street Address	0	3	1	1	1	2	0	0	1	0	1	47	9
Postal Code	0	1	0	0	1	4	0	2	2	1	1	47	7
Community Name	1	1	0	0	0	0	1	1	3	2	3	47	7
City / Town / Village	0	0	0	0	0	1	1	1	2	3	4	47	7
Region / Geographic Area	0	0	0	0	0	1	0	2	1	2	6	47	7
Latitude / Longitude	3	1	3	0	0	3	0	0	0	0	1	47	8

### UK

	0	1	2	3	4	5	6	7	8	9	10	77	99
Forename(s)	2	8	4	2	0	1	0	1	1	0	1	0	8
Surname	2	8	5	1	0	1	0	1	1	0	1	0	8
Initials	2	7	4	2	1	1	0	1	1	0	1	0	8
Sex	0	1	1	2	0	1	0	1	3	1	12	0	6
Date of Birth / Age	0	1	2	2	0	2	0	3	3	1	8	0	6
Date of death	0	1	2	1	0	3	0	1	2	0	12	0	6
NHS Number (OLD)	5	7	2	1	0	4	0	1	1	0	10	0	6
NHS Number (NEW)	2	5	1	2	1	5	0	0	2	0	3	0	7
Registered GP / Family Physician	0	0	0	2	0	3	0	0	0	0	1	16	6
Street Address	1	3	1	1	0	0	0	0	0	0	0	16	6
Postcode	0	0	0	0	0	1	1	1	1	1	1	16	6
Community Name	2	0	0	0	0	0	1	0	2	0	0	16	7
City / Town / Village	1	0	0	0	0	0	0	0	1	3	1	16	6
Region / Geographic Area	1	0	0	0	0	0	0	0	2	1	2	16	6
Latitude / Longitude	2	0	0	0	0	0	1	1	0	1	1	16	6

## II-3

On average, how often do you access the *PID* you identified above?

Please circle the appropriate number, 1 being "Rarely", and 10 being "All the time"

**CANADA**

	0	1	2	3	4	5	6	7	8	9	10	77	88	99
First Name	14	10	6	2	2	1	3	1	0	1	4	0	17	5
Last Name	14	8	6	3	2	2	3	1	0	1	4	0	17	5
Initials	20	5	5	3	1	1	0	1	1	1	2	0	17	9
Sex	3	1	0	1	1	5	4	6	3	5	16	0	17	4
Date of Birth / Age	4	2	0	2	1	6	3	6	5	6	10	0	17	4
Date of death	8	5	1	5	1	6	1	5	1	4	8	0	17	4
Provincial Health Insurance Plan Number	22	7	2	2	1	6	2	1	1	0	1	0	17	4
Hospital ID	14	10	2	3	2	2	3	1	2	0	5	0	17	5
Registered GP / Family Physician	5	3	0	0	1	0	0	0	0	0	0	36	17	4
Street Address	3	0	1	0	0	2	1	1	0	0	1	36	17	4
Postal Code	0	0	0	0	1	0	2	2	1	1	2	36	17	4
Community Name	0	0	0	1	0	0	2	1	1	1	3	36	17	4
City / Town / Village	0	0	0	0	0	0	1	2	0	3	3	36	17	4
Region / Geographic Area	0	0	0	0	0	0	0	2	1	1	5	36	17	4
Latitude / Longitude	3	2	0	1	1	0	0	0	0	0	2	36	17	4

**UK**

	0	1	2	3	4	5	6	7	8	9	10	77	88	99
Forename(s)	10	4	4	0	0	1	0	1	0	0	0	0	4	4
Surname	10	4	3	1	0	1	0	1	0	0	0	0	4	4
Initials	9	6	2	1	0	1	0	1	0	0	0	0	4	4
Sex	1	2	2	0	0	0	1	3	0	3	8	0	4	4
Date of Birth / Age	1	2	2	1	0	0	1	5	1	3	4	0	4	4
Date of death	1	2	2	1	1	3	0	3	0	2	5	0	4	4
NHS Number (OLD)	7	6	3	3	0	0	0	0	0	0	1	0	4	4
NHS Number (NEW)	5	3	3	2	1	3	0	1	1	0	0	0	4	5
Registered GP / Family Physician	0	0	1	0	1	0	1	1	1	0	0	15	4	4
Street Address	0	3	1	0	0	1	0	0	0	0	0	15	4	4
Postcode	0	0	0	0	0	0	0	1	1	0	3	15	4	4
Community Name	1	1	1	0	0	0	0	1	0	0	0	15	4	5
City / Town / Village	0	1	1	0	0	0	0	0	1	1	1	15	4	4
Region / Geographic Area	0	0	0	0	0	0	1	0	1	2	1	15	4	4
Latitude / Longitude	2	0	1	1	0	0	0	0	0	0	0	15	4	5



**II-4** How useful / important is this *PID* to you and your roles and responsibilities?  
Please circle the appropriate number, with 1 being “Not at all useful”, and 10 being “Critical to my roles and responsibilities”

**III-2** How useful to you and your roles and responsibilities (as identified in Section I) would access to the *PID* you identified above be?  
Please circle the appropriate number, with 1 being “Not at all useful”, and 10 being “Very useful – would greatly enhance by roles and responsibilities”

**CANADA**

	0	1	2	3	4	5	6	7	8	9	10	77	99
First Name	6	15	6	7	1	4	2	3	3	2	3	0	14
Last Name	6	15	4	5	2	4	1	3	4	3	5	0	14
Initials	9	15	5	4	2	6	2	0	1	1	3	0	18
Sex	1	2	2	1	1	3	1	5	10	5	24	0	11
Date of Birth / Age	1	2	1	1	0	1	4	4	10	8	24	0	10
Date of death	3	2	1	3	0	6	4	3	6	8	18	0	12
Provincial Health Insurance Plan Number	11	12	6	3	2	4	2	4	3	2	7	0	10
Hospital ID	6	12	8	2	2	5	4	1	2	2	7	0	15
Registered GP / Family Physician	1	5	3	0	0	1	0	0	0	0	0	47	9
Street Address	1	2	1	0	1	0	1	0	1	2	1	47	9
Postal Code	0	0	0	0	0	0	1	1	2	2	3	47	10
Community Name	0	0	0	0	0	2	0	2	1	2	3	47	9
City / Town / Village	0	0	0	0	0	0	0	1	3	2	4	47	9
Region / Geographic Area	0	0	0	0	0	0	1	1	2	2	4	47	9
Latitude / Longitude	1	1	2	0	0	1	1	0	1	2	1	47	9

**UK**

	0	1	2	3	4	5	6	7	8	9	10	77	99
Forename(s)	2	11	3	3	0	1	0	2	0	0	0	0	6
Surname	2	11	3	3	0	1	0	2	0	0	0	0	6
Initials	3	13	1	3	0	1	0	1	0	0	0	0	6
Sex	0	1	2	0	0	0	0	1	2	3	13	0	6
Date of Birth / Age	0	1	0	1	0	1	0	1	4	4	10	0	6
Date of death	0	1	0	2	0	3	0	4	2	1	9	0	6
NHS Number (OLD)	3	8	2	2	1	3	2	1	0	0	0	0	6
NHS Number (NEW)	1	2	0	2	1	5	3	2	2	0	3	0	7
Registered GP / Family Physician	0	0	0	2	0	0	0	0	1	1	1	16	7
Street Address	0	2	1	1	0	1	0	0	1	0	0	16	6
Postcode	0	0	0	0	0	0	0	0	0	1	5	16	6
Community Name	0	2	2	0	0	0	0	1	0	0	0	16	7
City / Town / Village	0	1	2	0	0	0	1	0	0	1	1	16	6
Region / Geographic Area	0	0	0	0	0	0	0	0	2	1	3	16	6
Latitude / Longitude	0	4	1	0	0	1	0	0	0	0	0	16	6

**II-5** What impact would removal of your access to this *PID* have on the quality of your work and resulting public health decisions?  
Please circle the appropriate number, with 1 being “No impact – quality would not suffer”, and 10 being “Severe Impact - results and decisions would be severely compromised”

**III-4** What impact has your lack of access to this *PID* had on the quality of your work and resulting public health decisions?  
Please circle the appropriate number, with 1 being “No impact – quality has not suffered”, and 10 being “Severe Impact - results and decisions have been severely compromised”

### CANADA

	0	1	2	3	4	5	6	7	8	9	10	77	99
First Name	4	19	2	5	3	3	4	3	3	1	5	0	14
Last Name	4	19	2	4	3	3	3	2	3	3	6	0	14
Initials	6	19	3	4	6	2	2	3	1	1	2	0	17
Sex	2	4	1	1	0	3	4	3	4	9	24	0	11
Date of Birth / Age	2	4	1	1	1	3	1	4	6	8	24	0	11
Date of death	3	5	0	2	5	2	3	4	5	6	18	0	13
Provincial Health Insurance Plan Number	5	18	5	5	1	4	1	3	4	1	5	0	14
Hospital ID	4	14	10	3	1	4	1	2	4	1	6	0	16
Registered GP / Family Physician	1	5	4	0	0	0	0	0	0	0	0	44	12
Street Address	1	2	1	1	0	1	0	1	1	0	2	44	12
Postal Code	0	1	0	0	0	1	0	1	2	2	3	44	12
Community Name	0	1	0	0	1	1	0	1	2	1	4	44	11
City / Town / Village	0	1	0	0	0	0	0	2	2	1	4	44	12
Region / Geographic Area	0	1	0	0	0	0	1	1	2	0	6	44	11
Latitude / Longitude	1	4	1	0	0	1	2	0	0	0	0	44	12

### UK

	0	1	2	3	4	5	6	7	8	9	10	77	99
Forename(s)	4	11	1	3	0	1	1	1	0	0	0	0	6
Surname	4	10	2	2	0	1	1	0	1	1	0	0	6
Initials	4	10	2	3	0	1	1	1	0	0	0	0	6
Sex	0	2	0	1	0	0	0	0	4	4	12	0	5
Date of Birth / Age	0	2	0	1	0	0	0	1	3	2	14	0	5
Date of death	0	2	0	2	0	0	1	5	2	2	9	0	5
NHS Number (OLD)	3	11	4	0	1	1	0	1	1	0	0	0	6
NHS Number (NEW)	3	4	1	1	2	2	0	1	4	1	3	0	6
Registered GP / Family Physician	1	0	0	1	0	0	0	0	1	1	1	18	5
Street Address	0	1	0	2	1	0	0	0	1	0	0	18	5
Postcode	0	0	0	0	0	0	0	0	0	1	4	18	5
Community Name	0	2	0	1	0	0	1	0	0	0	0	18	6
City / Town / Village	0	1	0	1	0	1	0	0	0	1	1	18	5
Region / Geographic Area	0	0	0	0	0	0	0	0	2	0	3	18	5
Latitude / Longitude	0	3	0	2	0	0	0	0	0	0	0	18	5

**II-6** What *PID* do you currently **NOT** have access to, but believe would be beneficial to you to further enhance your work and resulting public health decisions? (Check as many as apply)

**III-1** Having access to which of the following *PID* would facilitate your roles and responsibilities, or enhance your work and improve resulting public health decisions? (Check as many as apply)

### CANADA

	NO	YES	88	99	Missing
First Name	6	3	17	10	30
Last Name	6	4	17	10	29
Initials	6	1	17	10	32
Sex	6	1	17	10	32
Date of Birth / Age	6	2	17	10	31
Date of death	6	5	17	10	28
Provincial Health Insurance Plan Number	6	8	17	10	25
Hospital ID	6	6	17	10	27
Registered GP / Family Physician	6	16	17	10	17
Street Address	6	9	17	10	24
Postal Code	6	8	17	10	25
Community Name	6	5	17	10	28
City / Town / Village	6	5	17	10	28
Region / Geographic Area	6	3	17	10	30
Latitude / Longitude	6	10	17	10	23

### UK

	NO	YES	88	99	Missing
Forename(s)	5	0	4	7	12
Surname	5	1	4	8	10
Initials	5	1	4	8	10
Sex	5	0	4	8	11
Date of Birth / Age	5	0	4	8	11
Date of death	5	0	4	8	11
NHS Number (OLD)	5	3	4	8	8
NHS Number (NEW)	5	3	4	8	8
Registered GP / Family Physician	5	5	4	8	6
Street Address	5	2	4	8	9
Postcode	5	2	4	8	9
Community Name	5	0	4	8	11
City / Town / Village	5	0	4	8	11
Region / Geographic Area	5	0	4	8	11
Latitude / Longitude	5	0	4	8	11

## SECTION IV – Privacy Issues

This section pertains to the field of public health in general, and uses the term “*public health practice*” to refer to its various activities, including research, surveillance, health service delivery, strategic policy and decision making, etc. The goal is to get your opinion, as a public health professional, on the overall impact of restricted access to *PID* on *public health practice* in Canada. These questions ask for your opinion; if you’re not sure how to answer a question in this section, please just hazard a guess!

**NOTE:** The term “access” as used in this survey implies the ability to actually acquire individual level data so you can work with it directly.

**IV-1** In your opinion, do current restrictions on access to *PID* pose an obstacle to any aspects of *public health practice* (e.g. research, surveillance, etc.)? Please circle the appropriate number, with 1 being “Not an obstacle at all”, and 10 being “Yes, they pose a serious threat to accurate public health practice”

### CANADA

Restriction Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5	7.69	5	7.69
1	1	1.54	6	9.23
2	6	9.23	12	18.46
3	2	3.08	14	21.54
4	2	3.08	16	24.62
5	7	10.77	23	35.38
6	9	13.85	32	49.23
7	11	16.92	43	66.15
8	10	15.38	53	81.54
9	3	4.62	56	86.15
10	6	9.23	62	95.38
99	3	4.62	65	100.00
Missing	1		66	

### UK

Restriction Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	0	0.00	0	0.00
1	2	7.41	2	7.41
2	1	3.70	3	11.11
3	0	0.00	3	11.11
4	0	0.00	3	11.11
5	3	11.11	6	22.22
6	3	11.11	9	33.33
7	4	14.81	13	48.15
8	6	22.22	19	70.37
9	3	11.11	22	81.48
10	4	14.81	26	96.30
99	1	3.70	27	100.00
Missing	1		28	

**IV-2** How amenable would **you** be to other professionals in the public health field having access to **your PID** for public health research and analyses (e.g. **your** address, date of birth, etc.) to improve public health delivery, service, etc?

**CANADA**

	<b>NO</b>	<b>YES</b>	<b>Maybe</b>	<b>Skip</b>	<b>Missing</b>
Frequency	3	40	19	3	1
Percent	4.62	61.54	29.23	4.62	-

**RESPONSES**

Reputable public health professionals would not share the data with outside agencies, right? If it would be used only for the purpose of improving public health - by all means 'go ahead'

Depends on what else goes with the PID

Really depends on issue and level. For many things just need unique identifier so that various information sources can be linked. Can do anonymously. Sometimes large clusters are adequate. Other situations specifically need to know individual for contact tracing etc. Generally for research purposes only need to know that counting once, and depending on issue, to be able to identify geographical aggregation or demographics etc.

I would need to have some information about how the data was going to be used and what data elements were required.

Except for some conditions, I'd be OK with this

Depends - age, sex, clinical info, postal code okay.

NOT first/last name, address.

It would be no problem as long as the information did not include name, address for example but could include sex, date of birth, region or postal code

I think that each case has to be weighed up on its individual merits. If the data were misused, it could have severe negative ramifications for everyone. The policy in PH practice should be to use the minimum amount of private information as possible. Some researchers may feel a pressure to publish and forego this rule. Hence I think checks have to be put in place.

It depends on what conditions are set, what training they have, and what their needs are for the data.

Yes with the understanding that the information was secure and used only for stated purpose

I would want to know the purpose from a research perspective, from a surveillance or disease control I would be less averse

I would share my information but would like to have an opportunity to consent based on the type or purpose of study/research rather than just a blanket/implied consent. I would be comfortable with Public Health Organizations accessing data, less comfortable with academic curiosity-driven research and not at all comfortable for commercial companies to have access unless I gave explicit consent.

I would also like to know that an ethical review of research has been done prior to any release of information to the researcher.

These questions are very context specific. In the case of an outbreak and response - I think PID should be available to officials. In terms of research - it depends on who is doing the research and why...thus let an ethics committee have the ability to make this decision. If handled as such, I would have no problems with allowing for the use of my address and date of birth.

Would require the researcher to submit a research proposal to determine consistent use of data and measures to protect privacy and confidentiality

This is an awkward question because I keep picturing colleagues being able to identify me!! I think if it were still kept somewhat "unidentifiable" i.e. year of birth rather than full date, name removed, postal code, etc. that it would be completely fine.

If confidentiality of this information can be respected, I would be fine with it.

Yes, as long as results from the research were published without providing public access to raw data - i.e. give my PID to public health researchers to conduct research, but use it for aggregating results in published reports. Also - PID should be tracked - i.e. distribution of PID should be subject to appropriate data licenses.

I think we need to strike a balance. A great deal of solid work can be accomplished with nominal data that has a unique identifier attached but no personal identifier attached. It is important to be able to accurately attest that the nominal data are "unique". It is far less important to attach a personal identifier.

Depends on who is using the info how it is being used.

**IV-2** How amenable would **you** be to other professionals in the public health field having access to **your PID** for public health research and analyses (e.g. **your** address, date of birth, etc.) to improve public health delivery, service, etc?

**UK**

	NO	YES	Maybe	Skip	Missing
Frequency	0	18	9	0	1
Percent	0.00	66.67	33.33	0.00	-

**RESPONSES**

Only if appropriate confidentiality agreements and secure access arrangements were in place  
 No problem as long as it was explicit that it was to be used either for my benefit of that of others.  
 I would only need to be assured there were processes in place to monitor and regulate the governance of the organisations and processes.  
 It would depend on the situation  
 Only for information that is ESSENTIAL – e.g. I would not be happy for my name to be used and don't believe that address (except postcode) is required.  
 There is always a concern about one's personal details being given out because of identity theft, credit card fraud etc. I'd be happy as long as they didn't have my name.  
 I would only be happy if the security of the info was ensured and that the data was being used for proper public health analysis!  
 It depends on what assurances there are regarding the use to which the information will be put and whether specific individuals (or organisations) will be held responsible for how data is to be used.  
 Best practice would be to use NHS number plus DOB, sex, postcode and names of individuals should not be accessed  
 Depends on the system, function and purpose. GP registration data - no problem; clinical data should not need names; surveillance data may need address - need to know basis only. Date of birth no problem  
 We have very little PID. (My assumption at an earlier stage of this questionnaire was that any PID was being referred to even that of community clients and co-workers in other partner organisations... this is obviously not the case.)

**IV-3a** In your opinion, what proportion of the public is aware of the impact of restricted access to *PID* on public health practice?  
 (Please just guess!) Please circle the approximate proportion

**CANADA**

	Percentage of public aware of impact											
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	NS	.
Frequency	36	10	8	3	4	2	0	0	1	0	1	1
Percent	55.38	15.38	12.31	4.62	6.15	3.08	0.00	0.00	1.54	0.00	1.54	-

**UK**

	Percentage of public aware of impact											
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	NS	.
Frequency	18	6	1	0	0	1	0	0	0	0	1	1
Percent	66.67	22.22	3.70	0.00	0.00	3.70	0.00	0.00	0.00	0.00	3.70	-

**CANADA**

Television ads already reach such a broad range of the public

Demonstration of value through reports that are meaningful to decision makers and presented in a way that the average person (OK maybe "less than average") can understand. PLUS Demonstration of where in the preparation of the report that the PID were removed, along with controls in place to ensure protection of privacy.

Not sure it's a public issue as much as it is a public policy issue. As a principle the most anonymous least intrusive measures are what we utilize necessary to get info for the protection of public or individuals. Fundamental ethical principles, and true need, rather than investigator curiosity are what should drive it

We need to show the public what we do with data in general and then what we would do with confidential data in particular. Most members of the public haven't a clue as to what public health professionals do (not easily visible like doctors or other health care professionals), and it's too easy to lump us all as "big brother" government types.

This is not easy. Opponents to science are given every opportunity to publicize their position and they often do. Potential supporters of science [e.g., senior government officials, politicians, public health practitioners] never do. Why not? We have to answer that question first.

Case studies, what if stories

Public education about times that populations health is put at risk due to lack of information being made available that already exists. Instead of asking permission for every use of the data, ask people if they are aware that the data is not allowed to be used by these medical professionals who need it to protect their health and the health of the community!

When individual seeks health services, they are notified, in writing of the restrictions

Better education of the public on how this type of data is being used and why it's beneficial.

By putting in place rules that would protect the individuals but at the same time enable public health professionals to have meaningful access to this information.

Publicizing specific problems that arise because of lack of access - and letting people know who to contact and what to ask for. Need to make sure first that safeguards are well in place, as protection of personal info is very much an issue. Use of consent forms and information forms with individual contact e.g. with public health professionals and with other health system workers? (i.e. explaining uses of info for public health purposes as well as specifics of protecting privacy etc.) Might be interesting to find out about the proportion using omnibus surveys etc.

Better communication, understanding and education

Advertisements, informing front-line health practitioners.

There are two sides to this ... is the public aware of the restriction & do they agree in an increase ... I would only agree to a strategy to increase access if there was a better understanding of what was trying to be accessed. I find that this is not clear to me ... preceding questions have been very generic & all encompassing

I do not know the percentage so I feel that I have no useful opinion here.

Communicate importance of this information for health research and planning purposes.

Education of health care providers (all levels) who are responsible for the collection of the data - so they can accurately explain the need for the information.

Education of politicians and policy-makers about the impact of limiting access.

Education and awareness-building of general public about how the information can be used to improve their health care.

Education of data users about need for privacy, confidentiality, and the use of small numbers - and how to report them accurately for the public.

Educate on purposes of health promotion- clearly explain what the data is used for- alleviate concerns around reporting, e.g. reported data is aggregated

Through education and awareness. Demonstrate to the public how access to PID information can further research and analyses, and ultimately improve public health and the delivery of services.

media campaign to promote public health benefits to access to PID for research/analysis

I am not sure why this question is being asked.

Increasing public awareness of regarding the impact of restricted access to PID on Public Health Practice will not necessarily lead to increased access to PID for public health activities.

Commercial media is probably the best avenue, provide information at points of contact but unless there is some kind of "event" or crisis I suspect it will be very difficult and even if informed once it will probably not be retained in memory.

Provide examples of the benefits of providing PID

Education. Stress the need for access under certain conditions and that it is 'for the greater good'. AND ensure that people know that these data are not generally available and are maintained in a secure environment.

Public education. The general public does not understand how their information can be valuable in research and public health practice. Connections need to be made and this can only be achieved through education.

Some kind of comparative study comparing the health and public health practices of a community that restricted the release of PID versus a community that did not.

media

government lobbying

pamphlets at doctors' offices, PH clinics

on government websites

lay public "scientific" journals (e.g. Canadian Parent)

Increasing awareness. I just think individuals don't understand why it is important.

Reports such as that will likely follow this survey.

Encrypted personal ID and determine as small as possible geographic areas of which the sum of personal information will protect the privacy issues.

Public education during admission to hospital; media campaign

Highlight in media research produced as a result of PID.

Including rationale for the reasoning behind data-sharing authorization forms that the public signs for integrated health information systems to work properly. These explanations of the purpose and usefulness may create the level of understanding in the public and their subsequent buy in to the integrated systems and sharing of the personal health information.

some type of marketing / advertising

Publicise studies that are based on confidential data and stress the importance of how the information could not have been obtained without this individual level data

I'm not sure that it is necessary to increase the proportion. I think most people already assume that we share information within the health care system. I think the bigger problem is with the constant emphasis on active consent. For example, our screening program for speech, motor skills, etc. deficits only reaches about half of kids b/c getting formal, active, positive consent is so difficult. There are plenty of other examples which I'm sure I'll be able to share as the survey continues.

Good question! Demonstrate use of PID for research that benefits public - in terms of new policies, programs, drug coverage plans, or resources that target health issues impacting communities / health regions across Canada. Results of research could be published in reports and using web mapping services in support of decision-making. Public education in doctors' offices / hospitals and of public health professionals may be needed in order to show patients the benefit of controlled PID usage (i.e. through appropriate licenses / safeguards) to Canadian health in general. The less people get sick - the less it costs our tax payer.

Put blogs on health reports about such

Offer information through primary health care provider i.e. family physician, media campaigns

Promote the importance of research in public health and the benefits of prevention.

Demonstrate the impact of a lack of data

Information during Census data gathering

## UK

Better information about how personally identifiable data is used for research. Better reassurance that such data will only be used for research and not for other purposes

Better media understanding of the problem and support

By alerting people to the INDIVIDUAL benefit they might accrue (e.g. good health service planning, not being asked your name 10 times when being admitted to hospital. Not being asked you personal details by 10 different government departments (and all of them being slightly wrong. I.e. demonstrate tangible benefit to user).

Need to better inform patients about the legitimate uses for their data with good examples of decisions or knowledge that are of benefit them that couldn't have been made without their data.

We also need to explain how data is secured, how they can exercise their rights under article 8 of the Human Rights Act. All of this falls under various sections of the Data Protection Act.

We also need more research on public attitudes to sharing - who can/should access; what trade offs people are prepared to make and so forth.

By better dissemination of Public Health information and intelligence directly to the public

Media interesting in this topic is only covered in the health journals - needs a wider coverage as the implications are huge



Use examples of research where the public is directly benefiting from the researchers having had accessed to PID. Seeing is believing.  
 Be clear about the safety precautions we all take when using PID, but also be upfront about why we need it.  
 Clarity, honesty and visible benefits.

Media coverage?

Education

I read the question as meaning what proportion are aware that personal details could be used in a useful way for analysis and in targeting initiatives etc. People, often stirred by the media, are generally reluctant to have their details 'out there'. There would have to be many examples of how better data-sharing had changed things positively in order to win the hearts and minds of the public.

At the present, there appears to be an air of negativity around why personal data SHOULD'N'T be used. Maybe there needs to be an emphasis on the benefits that this could bring, perhaps through describing specific cases where there would be an advantage.

Clear examples of what valuable analysis might be undertaken if record linkage was possible

Not sure we need to

Why would we want to

Raising awareness when the public interact with the NHS, probably at primary care level. But first the GPs have to be aware!

By explaining at every opportunity the importance of this kind of data usage

**IV-3a** In your opinion, what proportion of the public would allow the use of *PID* for public health practice if they were asked and educated on the usefulness of such data to public health practice? (Please just guess!) - Please circle the approximate proportion

**CANADA**

	Percentage of public aware of impact											NS
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%		
Frequency	2	4	8	6	10	3	7	15	7	0	3	
Percent	3.08	6.15	12.31	9.23	15.38	4.62	10.77	23.08	10.77	0.00	4.62	

**UK**

	Percentage of public aware of impact											NS
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%		
Frequency	1	0	4	0	5	5	4	4	3	0	1	
Percent	3.70	0.00	14.81	0.00	18.52	18.52	14.81	14.81	11.11	0.00	3.70	

**IV-4b** How do you think we could increase this proportion?

**CANADA**

Guarantee security of information being held by one independent agency only

In general, people are quite happy to talk about their own health if they perceive a benefit. We need to demonstrate the benefits to population health and what that means to each individual (in terms of access to services, individual health benefit, etc.)

Hard to answer as kind of all or none like questions. Public and policy makers as well as practitioners need to engage in discussion of principles and purpose that respects both, the scientific need, public goods and personal protection and privacy.

Since these estimated percentages were wild guesses, it is difficult to postulate a method to further increase an increase in a hypothetical number!

Preparation of scenarios that demonstrate the utility of: - linked data for research, anonymous / encrypted data that are still considered personal health information and are used for surveillance and epidemiology purposes.

Make it clear to the public what they will NOT GET if they refuse access to data. The BC Pharmacare opt-out option experience is a good example.

Case studies, what if models

I think it would be important to distinguish between research and other aspects of Public health such as surveillance. I think that there is a lot of personal information needed for Public Health to do its job - including follow-up of communicable diseases and surveillance and monitoring. If a researcher from outside of Public Health wanted access to PID, that, I think would require additional consent.

Simply asking the question and making the privacy personnel aware of the wishes of the public so as to allow policy to match public opinion.

Need to demonstrate the value added in them providing access to the information as well as demonstrating how the data will be kept confidential (only available on a need to know basis - limited access)

I think this was probably an upper end guess - although again, educating people, explaining the uses, showing that there is adequate protection in place for protecting personal info. Putting real limits on uses except where they are truly necessary (e.g. if I don't need to know name, then I shouldn't have it. If I don't need postal code, only FSA, then I should only have FSA.)

Education programs, pieces on TV programs such as "Discovery Canada", "Quirks and Quarks", public workshops and town-hall meetings.

By making it clear how the data could be used, the security ensures in place to prevent exposure/loss of confidentiality; clearly state the kinds of information that would/would not be used; assure public that data will not be passed on to another party and that use will be based on anonymized records.

I think that having a review panel for scrutinizing all applications would go a long way to improving confidence. As well, the benefits of the access must be clearly defined and risks minimized.

You can't get everyone because they are somewhat paranoid about privacy issues and will never understand how the data is anonymized and protected.

Show examples of how information has been used successfully- improve access to reported data

Through successful initiatives that used PID information...some need to see it to believe it! Also, people tend to understand the issues better if it directly affects them or a loved one.

Address fears that public have regarding reasons for collecting data and security of data

I don't think you can increase this proportion, and if so, not by very much unfortunately.

My opinion is that the risk perception regarding the implications of disclosure of personal information in the general public is very high. The majority of the public will always be reluctant to enable access to their personal information unless they themselves/close family/friends are involved in a situation where the benefits of access to PID are very evident.

I think there will always be reluctance by some again information at point of contact, case studies demonstrating importance an relevance to public health might be useful

When the public were asked to let their PID be used in the Walkerton study 95 % said yes. I think that is about as good as we will ever get.

Again - simply through education and illustration of how these data might be used.

Regular and consistent public education, including how the information will be used and how it will be safeguarded.

See 4.3B - Note the results of this study should be made available to the public through various types of media

Same as 4.3b

Awareness

Showing some evidences that personal information can be used for public health practices but at the same time can be protected to be available without control.

Explain how it is used and why. Give examples of research and policy that have been advantaged by the use of PID. State explicitly how id is protected

Highlight case stories in which research using PID has helped identify information pertinent to managing /preventing disease.

One on one personalized requests with the more determined individuals. This level of personalized request may be required for some people, perhaps when they were to request health services. Also during this discussion they could be informed of the percentage of people that HAVE already agreed to this, some might find security in knowing that the majority of their community has done this already.

Put into law that their PID would be only used for research

This is an interesting question. I guess it would depend on the procedure. Existing consent procedures indicate that about 85% of new mothers give consent in hospital to a follow-up telephone call from public health. I suspect that the other 15% aren't trying to hide their personal health information; they simply do not want the follow-up service. Perhaps a small proportion doesn't trust public health b/c they may know someone who had a run-in with CAS that was precipitated by a public health home visit. In such cases, the needs of the child have to be weighed against the privacy rights of the parents.

1. Public Education
2. Signed Release Forms (nightmare in terms of paperwork)
3. Through Data License Agreements protecting PID during research - held by Privacy Commissioner.
4. Through Canadian Privacy Legislation

It will take a concerted effort especially if the data involves highly sensitive information such as sexually transmitted diseases and other such information. Public health will also have to demonstrate very tangible benefits of having access to such data and those "benefits" will have to mean something to the average lay person, not just the public health professionals.

Offer information through primary caregiver, media campaigns

Prove that information will remain confidential

\* Info during Census data gathering

\* address this question directly on government web site

\* Info at hospital during event

## UK

Better education/information about the consequences of not doing so

If they were given practical examples of how it might benefit them and their families. (e.g. vaccination recall, better access to targeted population health care interventions such as screening)

It should be a condition of treatment that data maybe used for research/ public health and implied or opt-out consent should be the default.

By better communication of the benefits and guarantees of the privacy of information not required - e.g. name

Clearly outline the importance of PIDs to tackling morbidity and mortality -

Not sure you could convince everybody...

Provide public with more information regarding how the data are used, what for and potential benefits of providing the data and negative impact of missing data for research and public health improvement

Popular media coverage

Get media backing of the way that this has helped public health practice and also publicise how not sharing PID can lead to tragedies and missed opportunities - I'm thinking of the examples of Holly and Jessica in the Soham murder case where information about the risk of offending posed by Ian Huntley was not shared between the relevant authorities and he was therefore able to take a post at the school.

Employ Jamie Oliver to promote the idea

OR

Generate a major media scare story about how the absence of data leads to ineffective treatment or health promotion.

Or, more moderately, some intelligent media coverage of studies such as the research done looking at the wide public acceptance of use of data in cancer registries when people had the purpose of holding/using such personal data explained to them.

In terms of publicity, the 'infringement of privacy' lobby definitely has the upper hand at the moment when it comes to getting a message across. It's unlikely that health professionals or public health professionals can come anywhere near to the impact of mass media scare stories about 'big brother'.

By providing assurances around how the data will be used as well as demonstrating a means of ensuring that malpractice can be dealt with.

However, I suspect that even if this were provided, people would still be a bit twitchy about personal information being used by others.

Demonstrate need

By tackling the loss of trust in public institutions

Better publicity, when it is clearly in the public interest and not too tenuous as in some research work

Not aware of the need to do this

I don't think it could go any higher.

Creating summaries that take away the most identifiable data (forename, surname)

SECTION V – Current data holdings and provision to others...

This section gathers information on the sharing of PID within and between organizations.

For all scales, circle '0' if you "Don't know"

**V-1** What would you say is the one **most** critical obstacle in the sharing or acquisition of PID linked to health data? (Please select only one; give your opinion!)

**CANADA**

Obstacle	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	18	27.27	18	27.27
2	11	16.67	29	43.94
3	22	33.33	51	77.27
4	4	6.06	55	83.33
5	2	3.03	57	86.36
6	4	6.06	61	92.42
7	1	1.52	62	93.94
77	1	1.52	63	95.45
99	3	4.55	66	100.00

**UK**

Obstacle	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5	17.86	5	17.86
2	3	10.71	8	28.57
3	9	32.14	17	60.71
4	2	7.14	19	67.86
5	5	17.86	24	85.71
6	0	0.00	25	85.71
7	1	3.57	25	89.29
77	1	3.57	26	92.86
99	2	7.14	28	100.00

**V-2** Do you or your organization currently collect individual-level health data for any purpose (e.g. research, surveillance, service delivery, etc.), or act as the custodian of such data?

**CANADA**

	NO	YES	77	99
Frequency	8	52	1	5
Percent	12.12	78.79	1.52	7.58

**UK**

	NO	YES	77	99
Frequency	3	23	1	1
Percent	10.71	82.14	3.57	3.57

**V-3** For what specific purpose(s) is this data collected? (check as many as apply)

**CANADA**

	NO	YES	77	88	99	Missing
Research		27	1	8	5	25
Surveillance		37	1	8	5	15
Services		24	1	8	5	28
Other		8	1	8	5	44

**Other (responses)**

Outbreak investigations in collaboration with organizations that have mandate to act and to identify individuals

Outbreak assistance (so we don't own the data but we have access to it)

We do not collect, but we use PHI for the above purposes and for program planning and evaluation

Planning

Planning programs and services

Utilization studies, audits, quality reviews and projects  
 Outbreak investigations  
 Identification of environmental and health hazards

**UK**

	NO	YES	77	88	99	Missing
Research		16	1	3	1	7
Surveillance		15	1	3	1	8
Services		11	1	3	1	12
Other		5	1	3	1	18

**Other (responses)**

Performance  
 Epidemiological Analysis  
 Information provision to stakeholders  
 Epidemiology  
 Performance management

**V-4** What data is collected?

**CANADA**

	NO	YES	77	88	99	Missing
First Name	18	31	1	8	7	1
Last Name	17	32	1	8	7	1
Initials	31	18	1	8	7	1
Sex	9	40	1	8	7	1
Date of Birth / Age	8	41	1	8	7	1
Date of death	21	28	1	8	7	1
Provincial Health Insurance Plan Number	28	21	1	8	7	1
Hospital ID	36	13	1	8	7	1
Registered GP / Family Physician	36	13	1	8	7	1
Street Address	18	31	1	8	7	1
Postal Code	14	35	1	8	7	1
Community Name	26	23	1	8	7	1
City / Town / Village	19	30	1	8	7	1
Region / Geographic Area	28	21	1	8	7	1
Latitude / Longitude	46	3	1	8	7	1

**UK**

	NO	YES	77	88	99	Missing
Forename(s)	15	6	1	3	3	0
Surname	14	7	1	3	3	0
Initials	12	9	1	3	3	0
Sex	2	19	1	3	3	0
Date of Birth / Age	2	19	1	3	3	0
Date of death	7	14	1	3	3	0
NHS Number (OLD)	17	4	1	3	3	0
NHS Number (NEW)	11	10	1	3	3	0
Registered GP / Family Physician	9	12	1	3	3	0
Street Address	16	5	1	3	3	0
Postcode	6	15	1	3	3	0
Community Name	17	4	1	3	3	0
City / Town / Village	11	10	1	3	3	0
Region / Geographic Area	9	12	1	3	3	0
Latitude / Longitude	18	3	1	3	3	0

**V-5** How difficult is it for other public health professionals such as yourself to acquire access to your PID and linked health data holdings if they are outside your immediate working team, but within...  
Please circle the appropriate number, with 1 being “Impossible”, and 10 being “Very easy”

**CANADA**

	Difficulty Score													
	0	1	2	3	4	5	6	7	8	9	10	77	88	99
Your own organisation?	3	4	6	5	4	5	5	5	6	2	3	1	8	9
The Federal Government?	13	10	10	7	0	2	2	1	0	1	1	1	8	10
Your Provincial Government?	10	6	3	6	5	2	3	3	3	3	3	1	8	10
A Provincial Government other than your own?	20	11	6	4	0	1	2	1	1	0	0	1	8	11
A regional or public health authority?	11	7	3	5	3	9	2	3	1	1	0	1	8	12
A University or Research Facility	12	9	8	6	2	3	2	3	1	0	0	1	8	11
Another national government (e.g. CDC in the US, NHS in the UK, etc.)	18	10	9	1	4	0	0	0	0	0	0	1	8	15
The World Health Organization	21	9	6	3	1	0	0	0	0	0	0	1	8	17

**UK**

	Difficulty Score													
	0	1	2	3	4	5	6	7	8	9	10	77	88	99
Within your own organisation?	0	1	1	0	0	4	0	4	4	2	3	1	3	5
Within the NHS or Department of Health?	1	4	2	3	2	2	0	1	1	2	1	1	3	5
Within the UK, but outside of the NHS or Dept. of Health?	2	7	5	2	0	2	0	0	1	0	0	1	3	5
Within the European Union, but outside the UK?	8	7	1	0	0	1	0	0	0	0	0	1	3	7
Outside the European Union?	9	7	1	0	0	0	0	0	0	0	0	1	2	7
Within the World Health Organization*	9	6	0	1	0	1	0	0	0	0	0	1	3	7

\*Missing n=1

SECTION VI– Solutions & Research

The proposed research will seek to apply a method (called a *transformation*) to public health data such that important relationships within and between the data are preserved, but the actual identity of the individual is anonymised. So, for example, if you were looking at an infectious outbreak in children, you might be interested in where the infected children are relative to one another, as well as where the schools are, arenas, community centres, etc. You would then preserve the *relationship* between these points of interest, and change everything else, so that the original points can no longer be identified back to their original owners. In this way, you have *transformed* the data so that you're still looking at individual-level data, but can't determine who it belongs to (i.e., it has become anonymous). Assuming the data custodians allow the data derived from such a *transformation* to be made available to the public health professional community:

For all scales, circle '0' if you "Don't know"

**VI-1** How useful would such a transformation be to you in your current role?  
Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

**CANADA**

	Usefulness Score											
	0	1	2	3	4	5	6	7	8	9	10	99
Frequency	6	2	3	5	1	6	0	4	4	8	21	6
Percent	9.09	3.03	4.55	7.58	1.52	9.09	0.00	6.06	6.06	12.12	31.82	9.1

**UK**

	Usefulness Score											
	0	1	2	3	4	5	6	7	8	9	10	99
Frequency	1	2	0	0	1	5	1	3	6	0	7	2
Percent	3.57	7.14	0.00	0.00	3.57	17.86	3.57	10.71	21.43	0.00	25.00	7.14

**VI-2** How useful do you think such a transformation would be to the field of public health in general?  
Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

**CANADA**

	Usefulness Score											
	0	1	2	3	4	5	6	7	8	9	10	99
Frequency	5	0	0	0	4	1	3	10	6	13	19	5
Percent	7.58	0.00	0.00	0.00	6.06	1.52	4.55	15.15	9.09	19.70	28.79	10.58

**UK**

	Usefulness Score											
	0	1	2	3	4	5	6	7	8	9	10	99
Frequency	2	0	0	1	0	7	1	5	4	1	5	2
Percent	7.14	0.00	0.00	3.57	0.00	25.00	3.57	17.86	14.29	3.57	17.86	7.14

**VI-3** Imagine you are a data custodian, and that a method has been developed to take your individual level data and mask it or change it somehow, while still keeping it at an individual-by-individual level. Would you allow such a method to be conducted on your data so that it can be shared with other public health professionals for public health research and practice?

**CANADA**

	<b>NO</b>	<b>YES</b>	<b>Maybe</b>	<b>77</b>	<b>99</b>
Frequency	0	42	18	1	5
Percent	0.00	63.64	27.27	1.52	7.58

**RESPONSES**

Depends on use. The key will be to understand the planned dissemination of DATA and DERIVED DATA beyond the requestor.

Transformation is essentially a variation on aggregate data - changing the parameters to hide the individual while conveying the context. The difficulty with that in my field of public health is that outbreak investigation depends on the risk factors of the individuals who are sick... changing the context could well change the exposure and therefore change the perception of risk. Solving a difficult outbreak is all about going beyond the usual and probing why THESE individuals got sick at this time in this place... change any of those variables even subtly and you've got a different outbreak.

If I understood how it worked, maybe. For question 6.1, I do not understand how transformations to make the data anonymous will address the fact that the data would still be considered personal health information, i.e., even if the transformation removes names or addresses. We ONLY work with data that has no names, phone numbers, initials, street addresses (I believe that is what is meant by anonymous, right?). Health card number data are encrypted in the files to which we have access. Hospital numbers may also be encrypted.

Would require formal assurances such as a privacy impact analysis to validate appropriateness

I think it would depend on the use of the data. Transforming the data and keeping it on an individual level could be useful for some research projects. The process for transformation would need to consider all possible relevant relationships before the transformation was done.

I would still need to know how the data was going to be used and exactly what data elements were required.

This is no an easy question to respond yes or no to. Although this may be able to be arranged, given legislation a lot of work would need to be completed to determine if this was even possible.

The necessary safeguards to prevent identification of individuals still need to be considered - e.g. release of small numbers or exact longitudes and latitudes.

It depends on the purpose or publication. I'm not in a position to authorize such use.

Again, all of these questions are very context specific. In certain circumstances (regular surveillance data) I think transformed data would be excellent and would allow for sharing of data. In unusual situations, or highly volatile/political (e.g. SARS) situations - I can imagine that this would be a problem during the outbreak (but perhaps not for follow up research afterwards). The problem is, once you have sex, gender and PC - you can determine who the person is even if the ID is masked. Not sure how to get around this.

Provided key elements are not masked. Age and sex for example

The users would need to understand the purpose and caveats of the data. the audience would also have to be informed on what is actually presented

People may see the identification of an individual in a highly precise and accurate GIS environment as no different than an identification of the individual themselves.

It would be important for individuals to understand that sharing as described above, with professionals and researchers, does not mean making the information public

There is still the difficulty of sharing such a novel dataset! However, after a certain period of time, it would greatly facilitate expanded investigation of a particular dataset from different angles, etc... if confidentiality issues could be resolved.

If the process was approved and I had faith that the masking process was done well.

Depends how much the transformation changes the data. If you have a set of addresses of clients with cancer and you wanted to see if there was clustering based on similar age groups and this age field has been 'transformed' then what use would this be to the researcher or analyst?



For some diseases it might make sense. For infectious disease investigations it wouldn't be particularly helpful because we already have nominal data on infectious disease cases. In this case, researchers should work with local health units to pursue specific lines of inquiry. For other diseases like asthma, say in relation to smokestack plumes, there might be some utility.  
Need for the appropriate authority's approval

## UK

	NO	YES	Maybe	Skip	Missing
Frequency	0	14	9	1	4
Percent	0.00	50.00	32.14	3.57	14.29

## RESPONSES

Subject to clear understanding and appropriate protocols being in place

In practice this is what we do and pseudonymisation is definitely a way forward. Two issues that are essential to consider is who, for what reason and with what outputs; and the issue of disclosure from statistical outputs

Sharing of research data would undermine our own activities and result in a loss of business

It depends how else the data are published for example because of geographical boundary issues, data published at local authority level and primary care trust level may potentially be disclosive because of disclosure by differencing.

For most research, individual counts are not required but aggregated counts are (e.g. by age band and sex). Even if data are anonymised you might still be able to identify someone if their record had clinical information combined with geographical information such as postal code.

I am custodian of ONS birth and mortality data only. These data are available to public health professionals direct from ONS according to need, so I would leave it to them to modify their data

What we need is to be able to link health records via NHS number or some consistently pseudonamised code.

There are some rare conditions, treatments or circumstances where geographic and demographic information would still leave a 'data point' as a uniquely identifiable person. As the custodian of the data I might still have concerns that I was releasing information that could be used inappropriately and violated the privacy of the individual.

Depending on the reason(s) for this work being conducted.

Only if approved by current data security, confidentiality & Caldicott guidance

If reassured that data protection was preserved

A specific disease or health condition will be used to test and evaluate the developed method(s). This condition must have a known aetiology, with well-known patterns and relationships, to serve as a starting point for the research. It must also be a disease of interest to the public health community.

**VI-4** | What diseases, health conditions, or databases most immediately come to mind as potential subjects for this research?

&

**VI-5** | Based on your knowledge of the condition you identified in the previous question, what relationships to the physical environment would a *transformation* as defined in the opening paragraph of this section have to retain in order for the data to be meaningful and useful to you (e.g. where patients or cases are relative to each other, to schools, to restaurants, etc; where schools are relative to a type of industry; etc...)?

**CANADA**

<b>VI-4</b>	<b>VI-5</b>
Chronic disease surveillance	Clinics, hospitals, industry, schools, recreation centres, public libraries, roadways/transportation lines
Immunization & vaccine-preventable disease in children	School or day care setting (without knowing where the school is even) Age-specific population density (e.g. within CSD)
Could choose from a range of infectious or chronic conditions with multiple level of associations. e.g. Diabetes	Building on existing surveillance with physician pharmacy and other data bases.
Heart disease (MI, CVA). Its natural history, risk factors (host and environmental) are well described in the literature and although the overall incidence is declining, with the aging baby boomers we will see a rise in the absolute numbers in the next few years (further obscuring the identity of the individual).	As heart disease is associated with some environmental factors (e.g., hard water) and SES factors (e.g., smoking, which is associated with SES), these must be preserved. The advantage of heart disease is that we know the important variables and confounders, so when performing the transformation you can control for them. The lessons learned here could then be applied to other chronic diseases.
TB or an STI	"Geographic", Time, Person to person
Infectious / communicable diseases	infectious / communicable - relationship to schools
Injuries such as motor vehicle collisions.	injury - where the collision occurred and the characteristics (e.g., age, sex, blood alcohol level) of the people involved (as drivers or passengers)
Tuberculosis	Everything. And over a long period of time.
Diabetes, Heart Disease, cancer	Skip
HIV, chlamydia, diabetes, suicide attempts	Where they are relative to one another, and relative to contacts, where they socialize, where they visit health services,
Communicable diseases - ones that are more common - could be enterics (IPHIS) - wouldn't want something so common that would cause challenges (such as heart disease)	if it were enteric, could be restaurants, locations of special events, schools, etc.
STIs	Where STI cases are exposed to multiple partners/contacts
HIV or any STIs in general	Skip
Ischemic Heart Disease	Skip
Diabetes	
Cancer incidence data	Don't know
Some sort of communicable disease data, e.g. ??	
Diabetes, heart conditions, infectious diseases	Skip
Lung cancer	Skip

Various cancers, heart conditions, diabetes, infectious diseases such as Avian Flu, SARS, 'flu	schools, restaurants, industry proximity to schools, water supplies, age demographics
Gastro enteritis	Where patient lives relative to restaurants
West Nile Virus	Proximity to water bodies, parks, woodlots. Climate and land cover data would have to be retained.
communicable diseases	Geographic location, relativity to other populations,
Cancer registry	Skip
Cardiovascular disease	
Diabetes	
Asthma	
Infectious diseases such as influenza - using iPHIS	You can't change them and keep the integrity of the data. Sorry.
HIV, STDs, Hepatitis C	Where cases are relative to each other Type of community e.g., high risk
Food borne, waterborne and disease outbreaks	Where cases were exposed relative to each other in outbreak situations
Delivery of health services (inpatient, ambulatory):	Where individuals are receiving treatment and care relative to others in their community
Cancer therapies	
Surgeries	
Diagnostic imaging	
Invasive meningococcal disease	Where cases and/or contacts of cases live in relation to each other, schools, restaurants, sports teams other gatherings (e.g. church)
Chronic disease such as CVD with links to built environments, access to services etc., cancers with environmental exposure links, waterborne diseases	How cases relate to recreation sources, walking trails, fast food outlets, secondary treatment facilities, rehab programs. For cancers industrial exposure or contaminated sites, for waterborne types of water supplies, known chemical contaminants
Most chronic diseases, injuries, infectious disease	Sex, age, incident and prevalent disease, large industry (nuclear power plants?) drinking water source,
E.g. diabetes, heart disease	All relationships to the physical environment are potential contributors to the condition of interest
Infectious illnesses - enteric and respiratory (Influenza)	Infectious GI
Heart Disease	-where cases live relative to each other, to work, to school,
Breast Cancer	
TB, Sexually Transmitted Diseases	Skip
Enteric diseases (e.g. salmonella)	For Salmonella, I would need to know where cases lived/ate/visited, by date and by subtype.
Influenza	
HIV	
HPV/cervical cancer	
TB, HEPC	Location of case, sex and age
Outbreak investigations, trace backs	A large number of ethnic/family, spatial, environmental, diet, nutritional, age, sex, behavioural and other variables
Case control or spatial studies determining/assessing risks to health	
Enteric outbreaks, pollution from environmental sources	where patients or cases are relative to each other and to nearby point sources
Communicable diseases (influenza) and chronic diseases (diabetes, adverse medical events)	Very important.

Injury, trauma. Other conditions can at least be linked to current location at least in a broad sense, i.e., Postal Code. Trauma often happens outside of home/work (which itself is rarely collected) so without precise locational information, there is no way to link to environment/circumstances. E.g. motor vehicle accident with road condition.	For my work none, but I would imagine in the case I stated above, link with home and place of work/school, known location of incident (if accident), activity space (shopping, dining out)
Obesity	Where obese and non-obese children are relative to parks, public transportation, grocery stores, fast food restaurants, etc.
AIDS	Where the cases are in relation to each other.
Asthma, copd, cardiac	Residence, school or place of work of the individual, maybe using lat/long. Need to assign proper exposure (ie. to air pollution for each individual)
Giardiasis, Campylobacter	Where cases are in relation to each other, food stores and water sources
As I said in the previous question, I don't see any benefit to local public health agencies in anonymizing infectious disease data b/c we already have a mandate to collect the nominal information. However, an infectious disease, like a food-borne outbreak, would probably be a good model for your system.	The problem with any general framework for transforming infectious disease outbreaks is that each outbreak is unique. For a food borne outbreak, you might choose to preserve the relationships between cases and schools, restaurants, community centres, and churches. But, if the outbreak was from a butcher shop, you'd miss it. You can't think of every single possible source of outbreaks and then try to maintain all of the possible relationships. Rather, I can see a transformation routine being useful for obfuscation purposes. That is, I would use such a transformation if I wanted to make a map of an outbreak available to the public, but I don't want to identify particular case addresses.
Respiratory Health (Asthma, Chronic Obstructive Pulmonary Disease, Emphysema, Lung Cancer)	Relationship would need to be maintained in terms of: direct exposures to point-source pollution (industry, transportation - such as busy intersections)  Relationship could be aggregated to a geographic area for: direct exposures to atmospheric air pollution and SMOG indoor air quality / smoking other environmental determinants  Relationship to each other could be aggregated to a geographic area. (i.e. # of cases in given area). It is still important to have specific location information in order to geospatially reference health data into geographic layers (dissemination areas, health regions, provinces, etc.)
Reportable infectious diseases	Postal code assignments
Diabetes 2	Social determinants of health, where patients live in terms of SES, industry, transportation, education, smoking by-laws
Cancer	Skip
AMI	community-level

Avian influenza	Where the patients or the cases are relative to others, relative to schools, restaurants, etc; where the schools are relative to a type of industry; the origin of the disease, in which country, the geographical movement of the disease in the world, etc.
Reportable enteric diseases such as E. coli, Campylobacteriosis ...	Relative position of cases one to other Relative position and access to municipal services (water, sewage) Relative position to agriculture
<b>UK</b>	
Lung cancer Childhood cancers	Different datasets would probably need to be created depending on the type of exposures of interest. Relative position of cases to each other, to schools, where schools and homes are in relation to roads, airports and other major sources of pollution. Also need to attach information about socioeconomic variables to individual records (e.g. area deprivation scores)
You would be best to select a disease which is relatively rare rather than a common disease. Your choice!	Geographical location in relation to other determinants of deprivation such as those used in the Index of Multiple deprivation
Coronary heart disease. Some cancers. Some infectious diseases	Skip
The most obvious are cancer, but heart disease may be a better option	Skip
Infectious disease surveillance systems - e.g. tuberculosis	Locality, family relationships, age, contact with other cases
Diseases: cancers, chd, mortality, hospital episodes	Where cases are relative to one another or sources of exposure
CHD, cancers, injuries, Obesity amongst others	Skip Where a child lives, (is the area deprived), what is the local infrastructure? Is their home close to parks, busy roads, community centres etc?
Physical or sensory impairment	Skip
Mesothelioma	Skip
Diabetes; HIV; hepatitis C;	Need to handle linkage - so individual interactions with health service, social services etc. can be identified as relating to the same person
Circulatory Disease or Obesity	Street, neighbourhood community. Socio economic status, housing conditions,
Coronary heart disease	Deprivation level, age and gender
Stroke, admission data	Skip
STIs	Postcode of residence and/or school/workplace
Diabetes	Where patients are relative to health centres

Another proposed solution to the issue at hand is to build what are called automated *software agents*. You can think of these as applications that would go into a dataset wherever it is housed (i.e. at the custodian's location), perform the analyses for you (on the *personally identifiable data*) and return only the aggregated, and therefore anonymised, results. In other words, you would never see the actual data, but would have this "agent" do the analyses for you, directly on the *PID*; you simply get the results of the analyses, as long as, of course, they don't compromise privacy. As a simple analogy, it would be like you giving me an equation or function to perform on my data, and I giving you back the result of that function without you ever needing to see my actual data. Assuming the data custodians allow such a *software agent* to analyse their data and make the results available to the public health professional community:

**VI-6** How useful do you think such a *software agent* would be to you in your current role? Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

**CANADA**

	Usefulness Score											77	99	.
	0	1	2	3	4	5	6	7	8	9	10			
Frequency	3	5	3	6	3	7	4	4	7	4	12	1	5	2
Percent	4.69	7.81	4.69	9.38	4.69	10.94	6.25	6.25	10.94	6.25	18.75	1.56	7.81	-

**UK**

	Usefulness Score											77	99	.
	0	1	2	3	4	5	6	7	8	9	10			
Frequency	0	1	3	3	0	1	2	3	5	1	2	1	6	0
Percent	0.00	3.57	10.71	10.71	0.00	3.57	7.14	10.71	17.86	3.57	7.14	3.57	21.43	0.00

**VI-7** How useful do you think such a *software agent* would be to the field of public health in general? Please circle the appropriate number, with 1 being "Not at all useful", and 10 being "Very useful"

**CANADA**

	Usefulness Score											77	99	.
	0	1	2	3	4	5	6	7	8	9	10			
Frequency	5	1	0	6	2	5	5	5	9	8	13	1	5	1
Percent	7.69	1.54	0.00	9.23	3.08	7.69	7.69	7.69	13.85	12.31	20.00	1.54	7.69	-

**UK**

	Usefulness Score											77	99	.
	0	1	2	3	4	5	6	7	8	9	10			
Frequency	0	1	1	1	1	2	3	4	3	3	3	1	5	0
Percent	0.00	3.57	3.57	3.57	3.57	7.14	10.71	14.29	10.71	10.71	10.71	3.57	17.86	-

**VI-8** If you were (or are) a data custodian, would you allow such a *software agent* to access your data, conduct the analyses, and return the results to the public health professional community for research and analysis?

**CANADA**

	<b>NO</b>	<b>YES</b>	<b>Maybe</b>	<b>77</b>	<b>99</b>	<b>Missing</b>
Frequency	0	27	31	1	5	2
Percent	0.00	42.19	48.44	1.56	7.81	-

**RESPONSES**

With some involvement in the process of interpretation, or I suppose I would have to know for what purpose the data was being used in order to ensure all variables are accurately reported - would also depend on the dissemination of the outcomes, who are the results going to?

If I designed and controlled the agent in some way.

Need to be satisfied of not only the privacy protection but the accuracy, and relevant applicability of what is extracted.

Want to avoid the classical bad associations of fishing expeditions.

This technique has been used by Stats Can and Cancer Care Ontario forever... essentially they give you aggregate data on request and if the number in the cell is less than five (or something similar), they suppress that result in the name of privacy. Whether you have them do it or a software package, the result is the same.

If I had oversight of its operation on the data and a look at the results being provided back to the requestor

We would like to develop a collaboration within the Ontario government to do report generation using SAS and our PHPDB.

Would require confidence in the reliability of the application to meet conditions for sharing information.

I think the first concern would be to have confidence that the 'software agent' was able to do the analysis correctly and appropriately. I think it would be hard to analyze data without having access to the data. Therefore the issue of trust is on both sides - that of the researcher as well as the custodian of the data.

Would depend on the disease -

Don't have the time to be running analyses for others and trouble-shooting errors or figuring out how to do it for them.

I would have some concerns about quality assurance issues with people not being able to see what's going on.

Provided there was sufficient opportunity to test the application and fully understand its potential privacy and confidentiality impact

As long as the license belonged to the specific jurisdiction in order to control appropriate access & running of the software against data holdings

I would need to assure myself that the system does indeed work and I would want to see the final reports before they are released.

We would want to ensure what is returned is done accurately although we would still want to see the record-level data as a check on the work done.

Unable to verify data analysis

As long as limitations of the data are understood and accounted for/reported in the analysis

Depends entirely on the expertise of the software and if I had confidence in it to give me good analysis

Context specific again - but in concept, yes.

Provided privacy conditions were met.

Would like to analyze myself, might see certain abnormalities that may be missed using the software agent

Provincial/territorial partners would likely not go for this

Depends on the geographic scale, time frame, disease type (enteric vs STD), where the software agent is located (onsite only)

With consent of privacy issues

Trusting the black box is difficult. It would require a great deal of validation testing to have confidence that the right formulae and even the correct data are accessed. Not sure how confident I would be in the results

If I had faith in the process, and appropriate use of the data, and if it was a process approved by my superiors.

It would depend on contracting and the type of data they would have access to. Some data is easier to obtain than others

In a sense we do a less sophisticated version of this already: we aggregate cases up to a higher level of geography, which effectively anonymizes the data. For example, in Ontario, the Population Health Planning Database includes, among other things, all hospital discharge abstracts. The records have names and addresses stripped, but they have some higher level geographic variables, like postal code. This allows us to look at local neighbourhood effects, but keeps the person's exact location secret. I'm not sure what your software agent would do that would be better than that.

Would need legal to approve it first

Would have to understand it better

Need for the appropriate authority's approval

## UK

	NO	YES	Maybe	77	99	Missing
Frequency	0	15	7	2	4	0
Percent	0.00	53.57	25.00	7.14	14.29	-

## RESPONSES

Again depends on what this really means and appropriate approvals being in place

We are more likely to provide the software agent function than provide the data to it but Yes, possibly we would also contribute some data as well

It depends whether the Software Agent can be altered to exactly do what the researcher requires or whether it is just one standard one. In the latter case the information/data that comes back may not be useful and people may keep requesting the data "as is" anyway thus duplicating effort.

I think that there is a danger of the Software Agent becoming a black box. Whilst it is important to formulate a priori hypotheses it is important to have access to the data in order to establish what the best methods of answering this question may be.

Little control of the process

As long as satisfied on data protection safeguards

Provided the safeguards are in place

**VI-9** To summarise, if a solution is found such that privacy is no longer an issue, which of the following would you prefer? (Please select only **one**)

## CANADA

Preference	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Transformation	39	59.09	39	59.09
Software Agent	18	27.27	57	86.36
Skip	8	12.12	65	98.48
Blank	1	1.52	66	100.00

Those responding to the question only:

Preference	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Transformation	39	68.42	39	68.42
Software Agent	18	31.58	57	100.00

## UK

Preference	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Transformation	14	50.00	14	50.00
Software Agent	9	32.14	23	82.14
Skip	4	14.29	27	96.43
Blank	1	3.57	28	100.00

Those responding to the question only:

Preference	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Transformation	14	60.87	14	60.87
Software Agent	9	39.13	23	100.00



SECTION VII- Qualitative Component

**VII-1** How knowledgeable do you consider yourself on privacy and confidentiality issues / legislation?  
Please circle the appropriate number, 1 being "Not at all knowledgeable", and 10 being "Expert"

**CANADA**

	Knowledge Score											77	99
	1	2	3	4	5	6	7	8	9	10			
Strategic Decision / Policy Maker						1	1	4					
Manager / Coordinator		2	1		1	1	3	2		2	1	3	
Consultant	1	1				1						2	
Research & Analysis	2		3	2	6	3	7	9					
Front-Line					1		1						
Other				1	2		1	1					
<b>TOTALS</b>	<b>3</b>	<b>3</b>	<b>4</b>	<b>3</b>	<b>10</b>	<b>6</b>	<b>13</b>	<b>16</b>		<b>2</b>	<b>1</b>	<b>5</b>	

Mean Scores

	N	Mean	STD	Min	Max	Median
Strategic Decision / Policy Maker	6	7.50	0.8367	6.00	8.00	8.00
Manager / Coordinator	12	6.25	2.7675	2.00	10.00	7.00
Consultant	3	3.00	2.6458	1.00	6.00	2.00
Research & Analysis	32	5.88	2.0752	1.00	8.00	6.50
Front-Line	2	6.00	1.4142	5.00	7.00	6.00
Other	5	5.80	1.6432	4.00	8.00	5.00
<b>TOTALS</b>	<b>60</b>	<b>5.97</b>	<b>2.2168</b>	<b>1.00</b>	<b>10.00</b>	<b>7.00</b>

**UK**

	Knowledge Score											77	99
	1	2	3	4	5	6	7	8	9	10			
Strategic Decision / Policy Maker				1	1	1	1			1			
Manager / Coordinator					1		1						
Consultant								1					1
Research & Analysis	1		3		1	1	6	2			1	3	
Front-Line													
Other								1					
<b>TOTALS</b>	<b>1</b>		<b>3</b>	<b>1</b>	<b>3</b>	<b>2</b>	<b>8</b>	<b>4</b>		<b>1</b>			

Mean Scores

	N	Mean	STD	Min	Max	Median
Strategic Decision / Policy Maker	5	6.40	2.3022	4.00	10.00	6.00
Manager / Coordinator	2	6.00	1.4142	5.00	7.00	6.00
Consultant	1	8.00				8.00
Research & Analysis	14	5.64	2.2398	1.00	8.00	7.00
Front-Line						
Other	1	8.00				8.00
<b>TOTALS</b>	<b>23</b>	<b>6.04</b>	<b>2.1209</b>	<b>1.00</b>	<b>10.00</b>	<b>7.00</b>

**VII-2** How do you feel about the impact of privacy and confidentiality legislation – in particular the restrictions on access to personally identifiable data (e.g. The Privacy Act, The Personal Information Protection and Electronic Documents Act, etc.) – on public health?

### **CANADA**

They are important to have in place to ensure ethical use of personal health information.

Acts are too restrictive and do not take into account population health needs and modern approaches to data access/linkage and transformation. Organizations and individuals working in public health do not appreciate potential solutions and scope for data access (e.g. different rules for anonymous access, PID access, use, sharing, dissemination, reporting, aggregation)

Mixed bag.

Since not able to comment on last question, it's important to note that need for individual level vs. aggregate data really depends on the issue under study. Often need a mix of methods to ensure aggregate is valid, or that individual is relevant at population level.

Privacy legislation has become increasingly restrictive over the 15 years that I have been practicing public health in Canada. It has adversely affected outbreak investigations that I have been involved in and also the placement of our Field Epis (e.g., the current unresolved debate is whether Field Epis as federal PHAC employees can access provincial data in Alberta since the Alberta legislation specifically prohibits all but provincial employees).

I support such legislation and am hesitant to see these acts watered down. Identity theft is the fastest growing crimes today, so I feel Privacy Acts are increasingly important and outweigh other needs.

They are serious, but my ministry is a health information custodian and so are our health units. The challenge is for our external researchers who would like access to the data. Also, we have trouble getting data from other agencies, e.g., Cancer Care Ontario.

The legislation is based on incomplete, fuzzy thinking and does more harm than good. Privacy and confidentiality are legitimate concerns but aren't the only ones. Defaulting to the most restrictive interpretations is simple minded.

A two edge sword, that is still being implemented. Still cases of over interpretation leading to delayed or reduced access through risk avoidance.

In my position, I am able to access data for monitoring and surveillance so I am not specifically concerned about limitations in relation to privacy.

I feel it has the potential to compromise the surveillance and response role of public health, and that in general, there is an overly restrictive interpretation of legislation in some cases, and a lack of understanding of the legislation by many data custodians leading to more restriction than is actually warranted.

It is useful and necessary but can cause challenges with identifying patterns, particularly with occurrences which overlap geographical boundaries

I think there are some substantial impacts that have not been adequately addressed - there have been limits placed on research and understanding of disease aetiology, and I'm not convinced that this is helping to protect public privacy, meanwhile there are areas where privacy breaches are still a real problem and these are not being addressed. I also think much of the problem relates to a lack of understanding of the legislation its application and its full impacts at all levels, from those who are drafting the legislation to those interpreting it as data guardians and through to those who want/need to use it.

Too complicated and cumbersome and not enough flexibility and trust in our own judgements

The legislation delays and restricts access

Am in agreement with the intent of legislation but it does challenge us to look for the development of new tools & methods to effectively continue our work.

I think the effects are not that great yet. I think we can get around the main concerns with appropriate safeguards. The main problem is in sharing data across agencies.

As long as the data is does not include names I think it's fine.

Data would need to be stored in a secure place (password and User ID protected).

It has had little impact. Much of the impact has come through the MIS-understanding of the legislation by physicians and other health care custodians/professionals.

It is important to find the correct balance between protecting personal information and supporting research & decision making in health care

I am not sure about this question. We deal with the public at large in most cases however there are PH programs that deal with individuals directly - so they may be impacted.

I'm glad we have it...it ensures everyone is working on a level playing field. It hasn't been an obstacle for my work.

I think it is necessary to protect the rights of the public, but there should be some allowance (at least) in the case of disease surveillance and outbreak management for (at least) partial access to identifiable data to those who need it.

I'm not really sure as some public health functions will be covered by public health legislation such as disease control, it could make planning functions more difficult

It is important to have some PID available to public health so that the health of the public can be protected. Not every disease needs such access but some do.

These are important - though should not necessarily be universally applied

I think it is critical that we remain very aware of the need and right for privacy. However - under certain conditions (e.g. SARS) the appropriate official (e.g. CMOH) should have the discretionary ability to decide on whether these individual level data be available

Not knowledgeable enough on the issue to comment.

Don't see it as an obstacle, more of an obligation.

In general, I find it a hindrance to my work and a hindrance to the improvement and efficiency of public health.

I believe the impact has required a rethinking of why and how we use data.

Disappointed

It is a good thing but at times can be restrictive

Too restricted. Should be flexible with appropriate technology that can protect the privacy

It's a response to public concerns. Understandable as there have been stories in media about leaks/blunders - remember the banking data that was faxed to an American junk yard.

It will need to accommodate the need for detailed multi-level analyses that require individual level data.

These restrictions to personal data make public health activities challenging.

In my case I need to assign air pollution exposures to individuals and it is very difficult when the smallest geographical code is a Census Subdivision (mortality dataset). So basically it's frustrating.

I think it's important to protect privacy, but there do seem to be excessive restrictions to certain organizations. Health Authorities should be able to get data from the province that they supplied to them in the first place, without hassle! I can see making it a little more difficult for universities, but it should still be something they should be able to access

I do not have a particular problem with access to personally identifiable data. For infectious disease follow-up, we have PID already. For a lot of other data, we may need record-level data, but we don't need PID. This is easily accomplished at the database level by replacing the PID with a numerical unique identifier (needed for the purposes of eliminating duplicates, etc.). My greater concern is with issues of consent and use of data. In the provincial and municipal privacy acts there are some strange sections on how data can be used for "research". It almost becomes a game of semantics. In usual parlance, any systematic study could be called "research", but we find ourselves dancing around the word and instead calling our work "surveillance" or "program evaluation" so as to avoid the word research and the needless additional complications that introduces. Of course, most people believe, on principle, that they should have full control over their personal information. But that just isn't practical. We could never do any "research" if we had to constantly go back to get consent every time we want to go back and look at the data from another angle. Also, this isn't specifically related to legislation per se, but I have an issue with some of the blanket rules from agencies like Statistics Canada when it comes to residual disclosure. For example, we are supposed to suppress counts of less than 5 cases. This makes sense if you are cross-tabulating data. But this rule has somehow extended itself to cover simple univariate counts. Clearly, a univariate count in and of itself cannot have any residual disclosure. If we followed this blanket rule, in smaller population centres or with rare diseases, we'd have to suppress all of the data, despite the fact that there isn't any residual disclosure.

Needs to be worked on - Legislation in Canada does not consider the many complexities involved in sharing health information among authorities, and to third party - especially for use in geographic information systems.

For the time being, the legislation is appropriate as privacy remains paramount. Over time it may become evident that there is a need for some modifications if it can be demonstrated that public health has been endangered as a result of privacy legislation.

Good

I believe that it is important to regulate access to these data, but individuals in decision-making positions should have access to improve health care.

Usually difficult to interpret - therefore manager are over protective

## UK

It is important to have such legislation, and difficult to achieve the appropriate balance between protection of the individual and provision of useful data for research. However, from a research perspective, it can be very frustrating, and waste an enormous amount of time in negotiating access to and obtaining confidential data. If such data were more easily accessible, I believe there is a lot of important new research in public health that could be carried out that is not currently being done.

Reasonable concerns, though sometimes data protection is used inappropriately

Very worried, as I think legislators are walking blindfold into an area due to public pressure and ignorance without thinking through some of the longer terms DISADVANTAGES to the public interest.

It has increased the time taken to complete analysis and restricts our ability to make information available.

Not properly understood - culture of leaning to individual privacy - confusion and conservatism

There need to be guidelines and protection processes for restricted access in order to prevent misuse of the data. These processes make people collecting the data more aware of the exact need for all the information they intend to collect or use (much of it is not necessary for Public Health analysis)

At a personal level I have been very frustrated that it took me 9 months to obtain Ethics and Governance approval for a project that only uses non-identifiable A&E records even though I work for the NHS.

The reasons behind the DPA etc. are good but there should be exceptions made for research that will improve public health. Surely the greater good should override the Privacy principle especially if there is no intention of actually "investigating" any individuals but just the overall pattern of a "disease" or problem.

It has led to a lot of confusion especially when freedom of information act is added to the equation. I feel worried that restrictions are over protective.

Analysts locally are often not able to access data to the lowest level of geography which then enables analysis at any higher levels by aggregating the data up and are restricted to pre-defined geographical analysis or are unable to access data beyond their local geographies so are unable to benchmark results.

The secondary uses service (SUS) which will restrict access to PID by pseudonymising records is adding to this issue as often postcode is not released even with anonymised records. Even if area codes are added to data this is not always sufficient e.g. data may need to be analysed by neighbourhoods/sure start areas which do not conform to a standard geography and without postcode no needs assessments/geographical profiles/surveillance of health problems can be carried out.

Limits potentially useful research projects, and therefore hinders knowledge

I think that organisations often hide behind data protection legislation and use this as an excuse to not share data. the legislation has made people so fearful of sharing data that often organisations refuse to give out any data rather than engaging in coming to data-sharing agreements with other organisations. The legislation does seem to be ambiguous so that it is possible for organisations to refuse to share it but this is very frustrating when you want to use their data to do your job.

Although I feel that the privacy & confidentiality requirements are restricting public health work I can appreciate why they are required and public concerns. Having worked in a PCT as well as a PHO I certainly prefer \*not\* to encounter data that identifies individuals when I only really require to be able to identify whether the records relate to someone who is the same as or different from other records being analysed.

Personally, I would strongly favour solutions involving transformed data and methods of interrogating datasets to produce aggregated results that don't violate 'small numbers' rules. However, this is because I am not involved in research which involves tracing or follow-up with individuals or doing complex linkage between datasets. In these cases pseudonymised data may not provide all the answers.

The general principles behind each of these guidelines is understandable. However, I suspect that they have not been developed with an integrated approach in mind.

This silo type of approach can potentially obstruct legitimate studies being conducted.

It will severely limit the analysis and research that can be done locally in public health and will be extremely detrimental to understanding the needs to a local population.

Important: the public needs to be aware that confidentiality is taken very seriously in the NHS and social services

They exist for good reasons but are often misunderstood, and are interpreted too rigidly  
Data-sharing and use is not forbidden by these, these are simply excuses used by many practitioners, so as to not share data.

DPA ok; Caldicott is sometimes used unnecessarily strictly, to protect paranoid practitioners who think they're being audited on the fly.

Restrictions are required due to the ease thro IT of misuse of PID

I think it is appropriate as long as there are methods to anonymise data or to gain consent

I think that this type of legislation threatens to make access to routinely-collected national datasets increasingly difficult.

I am worried that there will be serious implications on public health intelligence

The legislation made us aware of the need to consider the individual behind the data and to only do relevant analysis. We continue to analyse data

### VII-3 | What do you think of the proposed research (development of a *transformation*)?

#### CANADA

Need more information about what it would look like to comment

Sounds like a good idea, especially for doing accurate spatial analyses.

I am sceptical that it can be successful in a real-world application and that anyone will accept it.

Methods that facilitate Public Health practice, research and surveillance, while protecting appropriate privacy considerations can only be a good thing.

For routine surveillance of communicable diseases it may protect privacy, but I'm not sure effective interventions could be designed to address the underlying risk factors in a timely fashion. For chronic diseases like heart disease, this may have more promise, but this disease tends to be less stigmatized and I'm not sure people care whether their privacy is "compromised" - unlike mental illness, HIV/AIDS.

I see the utility, but feel that lessening the constraints imposed by privacy legislation can be dangerous.

I'd like to know more about it.

Well, this is already being done. I certainly support it.

Worthwhile. Concerns with secondary use of information / results after transformation ... there is so much secondary use of data, the transformation may have unanticipated (and unrecognized) effects on the secondary use.

Extremely important, and fits well with similar work we are doing in Saskatoon.

Would be a valuable resource but still need to be flexible in that need to be able to access the raw data when necessary

Excellent proposal

I think it sounds like a great idea

I think this is a great idea - depends on how it actually functions I guess, but the concept is excellent.

Sounds like a worthwhile endeavour

This would be extremely useful and should be pursued

It sounds very promising.

Great! Good luck!

Not sure, would need time to think about it and explore implications to practice.

It may have use for many health agencies across jurisdictional boundaries.

It would be okay IF you can have unique identifiers to search out duplicates, repeat visits, etc. That, I think, is the biggest hurdle you face.

I like it

I like it

Great idea, but many issues still need to be addressed!

Sounds like a good idea - not sure what the buy in will be like from the public's perspective

Great idea, I have seen the application of this in the use of previously collected research data which was anonymized to link to mortality data

Interesting and could be really useful in chronic disease surveillance.

Interesting

I would very much support this.

Deserves consideration.

Think it may be useful, hard to say since the use of PID is not in my day to day activities.

Interesting and promising. However, there are still many of us in the field who need to have access to PID for case follow-up.

For some applications.

Good idea

Has good potential but need to address the legislative hurdles (federal, provincial and organizational)

Good idea!

If it gives access to individual level data I think it's important and potentially very useful. The problem I see is as I stated before. Even when anonymizing the individual, when you render him/her in a GIS environment, the individual "de-anonymizes" identifies themselves.

Useful if it manages to alter privacy legislation such that it maintains privacy and allows access to more individual level data, including geographic locations.

I think it is an excellent concept. To be able to maintain individual level results and information yet masking the identifiable portion of the datasets that people are most concerned with.

Sounds promising

Great idea if it can be modified enough so that necessary data attributes are still intact

As I said earlier, I think the proposed research might be useful for obfuscation. At this very moment, for example, I have 5 cases of blastomycosis that have a very interesting geographic distribution, but I don't want to publish the map b/c it would identify the affected households. I would love to have a method that re-positioned the data points but maintained all of the important relationships in the data. I'm not sure that a transformation is the way to go in this particular case; aggregation might be better. But it would be worth looking at. I can see how your proposed research might be useful to university researchers who don't have access to data the way I do.

Excellent idea. Fortunately, technologies and standards exist that can be used to create web services that automatically transform data into aggregate statistical data.

Daunting task but please continue your efforts

Encouraging, I would like to see it succeed.

Possible

It is important and will help us to better target our public health activities

Essential - Hopefully the start of ongoing research and discussion in public health and government arenas.

## UK

Potentially useful, but it is difficult to see how it will be able to fully solve the problems, since each individual study will have particular variables (exposures and confounders) of interest that will need to be linked at an individual level.

Good idea - but seems to me a bit like pseudonymisation

Depends on how easily it can be sold to the press/public (as in pseudonymisation)

Sounds like a useful research project.

Interesting but would have limited application since most Public Health analysis requires the researchers to have a detailed knowledge of the underlying data system. This research appears to assume that no Public Health professional would be able to access detailed datasets and get out meaningful analysis.

Transformation if it is feasible sounds like something that the Public might agree to. Not sure why but it feels more of an option than the Software agent.

Good idea but needs to be properly thought through in terms of whether it fulfills all the potential public health analytical needs.

Solves the problems of small numbers in a sensitive area, but not that of record linkage

I think it sounds like an excellent idea and definitely something that is needed to enable many datasets to be brought together in useful ways.

I am definitely in favour of this sort of approach.

A good starting point - but I think there may be some issues that might need some thought.

Preserving relative information whilst anonymising specific cases is a good tack. However, one needs to guard against the possibility of a black box type approach of analysis.

Further light on answering a research question might be gained through a consideration of specific information.

Although spatial relationships may be preserved in an analysis we also need to bear in mind other factors which may not be distributed evenly through space - eg, pockets of deprivation.

Can not see the usefulness to my practice

Sensible solution

Worth exploring

Not enough detail on how it would work for me to comment sensibly

Good luck

Sounds like a useful idea

I don't know, frankly.

Good idea that's already done to some extent

#### VII-4 What do you think of the "software agent" idea?

##### CANADA

Also a great idea for those that are able to what they want using the tools provided by the software agent.

Brilliant, but depends on everyone agreeing on standards and agreeing to properly manage data at the local level.

It will depend on how accurate, useful and verifiable

Stats Can has been doing it forever and we've been paying big bucks for the privilege.

Again, I see the utility and would prefer data at the aggregate level only.

Sure.

Fine. But it's only as good as the data it's working on and the sophistication of the algorithm.

Lots of work needed yet.

Same "envelope" as the "canned reports" approach to databases. For research, interesting questions are often detailed and unanticipated. Not sure that an agent could capture the full flexibility to perform a complex sequence of steps .. Much like a complicated "data" step in SAS followed by a series of PROCs which yield statistics which are themselves transformed through subsequent DATA and PROC type steps. How much of the research community/research community's needs could be coded into an agent?

Same as previous answer - extremely important

Valuable resource

Transformation would be the priority for me and if that's not feasible then the software agent.

I think this is also a good idea, and in many cases probably ideal. I do have some hesitation in terms of quality assurance - I know when I am analyzing data I tend to catch my mistakes by examining the data closely - at the same time, in the end I don't need to see the personal information and I just want the aggregated data...

Great

Sounds interesting, but may require quite a bit of 'selling'

This could also be extremely useful and should be pursued. I don't see this as an "either/or" issue - why not have both?

This sounds like an example of a tool which would facilitate the work required in this field

It also sounds promising. I can see each having its own place

Not sure - I would have to see it in practice. I don't like the idea of splitting up the data, research, and methodology. I wonder if it would increased opportunity for error and reduce rigor/validation.

Again, would need to think further about the implications to practice.

A good idea

Better than nothing but concerns with inability to verify data

I like it

Another great idea, but needs further development to address data limitations.

This is how we currently access data from the province. Works quite well in most cases.

Good idea

Another good idea that may be more acceptable

It depends on if the analysis is 'canned' to 'most users'. Then it becomes much less useful. If the user can specify what they what analysed it would be great as long as the analysis is done well.

Also interesting - I would like to test it out

Support this, but more keen on transformation as I would prefer to have access to data at individual level for my analyses.

Probably most useful to policy makers and planners. Researchers usually prefer to have raw data.

See, 7.3

We already use something similar and it is of use to unsophisticated users who do not need to manipulate the data but need to be aware of trends.

For what we do in our area, this is not an ideal situation.

Good idea

Good idea. Need to establish a common ground location. UBC CHSPR has a non-spatial version of this database linkage idea

Not sure, but worth trying

Sounds like a black box that I have little/no opportunity to validate. It would take regular data/program validation for me to have confidence that what comes out is a function of what went in and is correct.

Similar to previous answer. May be more acceptable to some data custodians.

I am not sure that I understood the concept fully.

I'm not sure about this idea. There is always the possibility that a software agent could corrupt the data.

Not as enthusiastic about this idea. What type of techniques would they use and how would they get there. Would like to see the raw data and be able to pick out anomalies.

I am just trying to imagine a situation where it would be useful. I guess the researcher would have to tell you ahead of time what geographic relationships he was interested in, and then you would program those particular things into the software agent, and then it would sort through all of the data, identify all of the relevant cases, and then obfuscate their geographic location while retaining all of the a priori geographic relationships. I can see the abstract beauty of the idea. And I think it does mitigate the privacy problem. But as you well know, researchers tend to hunt for low p-values; I think you'll have to be so, so, so careful of identifying spurious relationships. You are putting a tremendously powerful tool for finding correlations into the hands of people who are disconnected from the actual local conditions and occurrence of disease. This is potentially dangerous because the data become decontextualized.

Again, excellent idea. see last comment (7.3)

Security would still be an issue to be explored

Such a Software Agent could be deployed internally within a health organization, or externally such as in the Canadian Geospatial Data Infrastructure, or hosted on a web server by the Public Health Agency of Canada. Any organization with access to raw PID could set up a Software Agent for serving information to defined clients / purposes.

A little "scary" at this point. Even the term "software agent" may have a negative connotation from the perspective that it may be perceived as "intrusive".

I think it has potential

Unsure

Think it would be useful for about 75% of the work I do

Brilliant, if it is possible!

Conceptually great - think is not as simple as it sounds.

The "Buy-in" might be not good.

## UK

Again, potentially useful in situations where the analysis is relatively straightforward. However, for more complex analyses (e.g. multilevel modelling, inclusion of measurement error models, models to account for missing data etc.) it may be very difficult to achieve in practice. The analysis stage is often an interactive process of model building which it is difficult to fully prescribe in advance for a third party to carry out.

See previous question

Sounds useful for people who are not used to analysing or viewing data themselves

This is a good idea but all the caveats about the need to understand the underlying data in detail still apply.

Not so sure about this one.

Fine so long as it is appropriately quality assured and stakeholders consulted to ensure the right types of aggregates are available

I would want to have access to record level data for some purposes, but for most I already act as the Software Agent in supplying aggregate data to other people for further analysis

If it can be produced as a tool which it is easy to use and/or training is available then I think it will be extremely valuable. In my own area of work I would rarely need to hold record level data \*if\* I can easily and quickly produce aggregated and cross-tabulated data. However, I foresee that for many people they will always prefer to use their own tools (stats packages, databases, spreadsheet etc.) to perform the analysis. So I don't think that a Software Agent will necessarily provide a solution to people wanting to download large chunks of record-level data.

A good starting point - but there may be some issues that might need some thought.

Answering research questions appropriately will require an iterative approach to handling data. For instance, in applying diagnostic methods to ascertain whether proposed analyses are appropriate.



Completely disagree with it - access to the raw data is absolutely required  
Not good as there will be a lot of time wasted and expertise will be lost at local level  
Good idea but may be unnecessary given the work being done by the Information centre on pseudonymisation  
Not much  
Worth trialing as for Public Health policy planning etc. does not require identification of individuals  
Sounds interesting but I would need to know more about it  
I think that access to many types of databases is already organised in this way.  
Theory sounds good but I think it would be extremely hard to get working because of data quality and format issues  
Good idea too

**VII-5** Do you have any other thoughts or comments regarding this issue, the proposed research, or this questionnaire that you would like to share?

**CANADA**

Good luck!

Good luck with your research, Philip. I think the most important thing with your transformation idea is choosing one disease of public health interest that you can apply the technique to and make it work successfully. This would build confidence and make more people interested in the method.

If privacy legislation is loosened in any manner it can open the door to increased identity thefts. Systems must be put in place to ensure private information is safeguarded. There is growing evidence that criminal organizations (such as the Hell's Angels) are highly active in identity theft - even bonded government agents and healthcare workers can cave to the pressures that criminal organizations will exert to get personal information. How can we ensure that personal information will remain private?

This questionnaire was somewhat troublesome to use.

Question 1.3 was cut off at the bottom so I don't know the option after Mental Health & Substance Misuse.

Question 2.6 was cut off after Street Address

I am not sure that I understand question 5.5. We are permitted to disclose our personal health information to health units, who are also health information custodians.

For question 6.1, I do not understand how transformations to make the data anonymous will address the fact that the data would still be considered personal health information, i.e., even if the transformation removes names or addresses. We ONLY work with data that has no names, phone numbers, initials, street addresses (I believe that is what is meant by anonymous, right?). Health card number data are encrypted in the files to which we have access. Hospital numbers may also be encrypted.

Re 6.6. We already have software for our Provincial Health Planning DataBase. But the issue is it's not customised to suppress small values. And sometimes we need to see those small values to review the data.

I expect that if I knew more about transformation and PHMG, my responses could have been refined.

With the proposed research, transforming and making the data anonymous or a software agent, do you anticipate that all research would go through an ethics review process? I hope so.

I found it frustrating answering this survey online. I often went to the previous question in the midst of answering the current question and then my answer was not saved. I also like to see the layout of a questionnaire so I can organize my thoughts rather than going sequentially though the questions one at a time.

Please connect with my CCIS team in Saskatoon to compare notes and share your results sounds like a valuable resource that could meet challenges we currently have in the field

I found it a little difficult to answer some of the questions about my use of personal information, and particularly when it came to an organizational level, I really don't know how other parts of this very large organization work.

Collecting information that is already collected and available is ideal if the sharing info problems can be worked out.

I have noted this in a couple of responses - but it is hard to answer a lot of these questions 'generally' as my concerns and opinions on privacy very much depend on the situation.

Privacy is critical - but equally, if we have to aggregate our data too much for the sake of it, the results and information that come out of this are not always very useful, and in many cases do not give us a very accurate picture of what is really happening.

Great initiative. Long overdue. Best of luck.

Challenging issues, but a very worthy topic of concern ...privacy.

Very well done web based survey and quite user friendly. I also like that my privacy as a participant has been respected.

Good luck on these endeavours!

In my research, I don't need to know the individual. I just need to know the different groups of individuals that live in the same geographical location (whether by postal code, FSA, DA etc.).

Data that is very useful to our organization also includes:

Most Responsible Diagnosis / ICD 9-10 codes

Admission and Discharge data

Length of Stay data

Great research project!

## UK

This questionnaire has been quite lengthy and not terribly easy to use

I think it sounds like an excellent idea and I wish you the best of luck with the research. I am interested to know more.

I'm not clear how this research links with the work that is being done on providing pseudonymised data through SUS, or to what extent proposed systems will already provide some of these concepts of 'transformed' data and a 'software agent' for performing analysis.

This research is important but this questionnaire tool is quite frustrating to use

Many questions were hard to answer as my policy role related to PID but I personally don't work with it.

The notion of pseudonymisation is a more promising prospect - where individuals are anonymised but there is a way of getting back to the detail if needed.

## D. Multidimensional Point Transform algorithm: SAS code

```
/******  
/*          MIDAS TRANSFORM                               */  
/*          NEW YORK, US                                 */  
/******  
/*          CREATED BY PHILIP ABDELMALIK FOR popDATA DATASET */  
/*          SEPTEMBER 08, 2010                             */  
/*          ADAPTED FOR NEW YORK DATA OCTOBER 14, 2010    */  
/******  
/* REVISIONS: 08092010; 10092010; 13092010 (added age categorisation) */  
/*          15092010; 29092010 (AUS); 03102010; 12102010; 13102010 */  
/*          21102010; 22102010; 25102010; 26102010; 27102010; */  
/*          28102010; 23112010; 11122010                 */  
/*          (modified to allow setting of "k" anonymity level) */  
/******  
  
/******  
/* The following code generates simulated patient "lists" from the base synthetic */  
/* population, and then calculates the great-circle distance between each "patient"*/  
/* and all other individuals in the specified population who match on age and sex */  
/* The code first converts the lat-long coordinates to radians, and then uses the */  
/* quadratic mean radius of the earth and spherical trigonometry to calculate the */  
/* distance.                                             */  
/*                                                     */  
/* Note the error associated with this calculation:      */  
/* i. It assumes that the earth is spherical, which it is not */  
/* ii. Because the earth is a geoid, not a sphere, the radius will vary */  
/*     depending on where it is measured to. The quadratic mean radius was */  
/*     used as the best estimate.                        */  
/******  
  
PROC PRINTTO LOG = "C:\RESEARCH\NY\NewYork_LOG100b";  
RUN;  
  
%PUT BEGIN...;  
*SASFILE MIDAS.popDATA CLOSE;  
/* CLEAR POPULATION DATA FROM RAM IN CASE PRIOR EARLY TERMINATION */  
  
/******  
/*          HOUSEKEEPING                               */  
/*          CLEAR TEMP WORK LIBRARY                   */  
/*          CLEAR ALL MACROS                          */  
/*          CLEAR ALL MACRO VARIABLES                 */  
/*          ADAPTED FROM PAPER 082-2009 BY CHUCK BININGER */  
/*          RETRIEVED FROM THE SAS WEBSITE            */  
/******  
  
%MACRO HOUSEKEEPING;  
  
PROC DATASETS  
LIBRARY = work  
KILL;  
QUIT;  
  
%LET syscc = 0; /* Operating environment condition code */
```

```

%LET sysrc = 0; /* Operating system condition code */
%LET syslibrc = 0; /* Libname statement condition code */
%LET sysfilrc = 0; /* Filename statement condition code */
%LET syslckrc = 0; /* SAS Shre lock statement condition code */
%LET syslast = ; /* Contains last created dataset */

```

```

OPTIONS NONOTES; /* Requests whether or not notes are output to the log */
OPTIONS obs = max; /* Resets the number of observations to process */
OPTIONS THREADS=YES;

```

```
%MEND HOUSEKEEPING;
```

```
%MACRO FORMATTING;
```

```

/*****/
/* DEFINE FORMATTING */
/*****/

```

```
/* FORMATTING FOR THE VARIOUS AGE GROUPS */
```

```
PROC FORMAT;
```

```

VALUE secondAG 1="0-1" 2="2-3" 3="4-5" 4="6-7" 5="8-9" 6="10-11"
7="12-13" 8="14-15" 9="16-17" 10="18-19"
11="20-21" 12="22-23" 13="24-25" 14="26-27"
15="28-29" 16="30-31" 17="32-33" 18="34-35"
19="36-37" 20="38-39" 21="40-41" 22="42-43"
23="44-45" 24="46-47" 25="48-49" 26="50-51"
27="52-53" 28="54-55" 29="56-57" 30="58-59"
31="60-61" 32="62-63" 33="64-65" 34="66-67"
35="68-69" 36="70-71" 37="72-73" 38="74-75"
39="76-77" 40="78-79" 41="80-81" 42="82-83"
43="84-85" 44="86-87" 45="88-89" 46="90-91"
47="92-93" 48="94-95" 49="96-97" 50="98-99"
51="100-101" 52="102-103" 53="104-105" 54="106-107"
55="108-109" 56="110-111" 57="112-113"
58="114-115" 59="116-117" 60="118-119" 61="120-121";

```

```

VALUE thirdAG 1="0-2" 2="3-5" 3="6-8" 4="9-11" 5="12-14"
6="15-17" 7="18-20" 8="21-23" 9="24-26"
10="27-29" 11="30-32" 12="33-35" 13="36-38"
14="39-41" 15="42-44" 16="45-47" 17="48-50"
18="51-53" 19="54-56" 20="57-59" 21="60-62"
22="63-65" 23="66-68" 24="69-71" 25="72-74"
26="75-77" 27="78-80" 28="81-83" 29="84-86"
30="87-89" 31="90-92" 32="93-95" 33="96-98"
34="99-101" 35="102-104" 36="105-107" 37="108-110"
38="111-113" 39="114-116" 40="117-119" 41="120-122";

```

```

VALUE fourthAG 1="0-3" 2="4-7" 3="8-11" 4="12-15"
5="16-19" 6="20-23" 7="24-27" 8="28-31"
9="32-35" 10="36-39" 11="40-43" 12="44-47"
13="48-51" 14="52-55" 15="56-49" 16="60-63"
17="64-67" 18="68-71" 19="72-75" 20="76-79"
21="80-83" 22="84-87" 23="88-91" 24="92-95"
25="96-99" 26="100-103" 27="104-107" 28="108-111"
29="112-115" 30="116-119" 31="120-123";

```

VALUE fifthAG 1="0-4" 2="5-9" 3="10-14" 4="15-19"  
5="20-24" 6="25-29" 7="30-34" 8="35-39"  
9="40-44" 10="45-49" 11="50-54" 12="55-59"  
13="60-64" 14="65-69" 15="70-74" 16="75-79"  
17="80-84" 18="85-89" 19="90-94" 20="95-99"  
21="100-104" 22="105-109" 23="110-114" 24="115-119"  
25="120-124";

VALUE newAGE 1.1="1" 2.1="2" 3.1="3"  
4.1="4" 5.1="5" 6.1="6" 7.1="7"  
8.1="8" 9.1="9" 10.1="10" 11.1="11"  
12.1="12" 13.1="13" 14.1="14" 15.1="15"  
16.1="16" 17.1="17" 18.1="18" 19.1="19"  
20.1="20" 21.1="21" 22.1="22" 23.1="23"  
24.1="24" 25.1="25" 26.1="26" 27.1="27"  
28.1="28" 29.1="29" 30.1="30" 31.1="31"  
32.1="32" 33.1="33" 34.1="34" 35.1="35"  
36.1="36" 37.1="37" 38.1="38" 39.1="39"  
40.1="40" 41.1="41" 42.1="42" 43.1="43"  
44.1="44" 45.1="45" 46.1="46" 47.1="47"  
48.1="48" 49.1="49" 50.1="50" 51.1="51"  
52.1="52" 53.1="53" 54.1="54" 55.1="55"  
56.1="56" 57.1="57" 58.1="58" 59.1="59"  
60.1="60" 61.1="61" 62.1="62" 63.1="63"  
64.1="64" 65.1="65" 66.1="66" 67.1="67"  
68.1="68" 69.1="69" 70.1="70" 71.1="71"  
72.1="72" 73.1="73" 74.1="74" 75.1="75"  
76.1="76" 77.1="77" 78.1="78" 79.1="79"  
80.1="80" 81.1="81" 82.1="82" 83.1="83"  
84.1="84" 85.1="85" 86.1="86" 87.1="87"  
88.1="88" 89.1="89" 90.1="90" 91.1="91"  
92.1="92" 93.1="93" 94.1="94" 95.1="95"  
96.1="96" 97.1="97" 98.1="98" 99.1="99"  
100.1="100" 101.1="101" 102.1="102" 103.1="103"  
104.1="104" 105.1="105" 106.1="106" 107.1="107"  
108.1="108" 109.1="109" 110.1="110" 111.1="111"  
112.1="112" 113.1="113" 114.1="114" 115.1="115"  
116.1="116" 117.1="117" 118.1="118" 119.1="119"  
120.1="120" 121.1="121" 122.1="122" 123.1="123"  
124.1="124" 125.1="125"

1.2="0-1" 2.2="2-3" 3.2="4-5" 4.2="6-7"  
5.2="8-9" 6.2="10-11" 7.2="12-13" 8.2="14-15"  
9.2="16-17" 10.2="18-19" 11.2="20-21" 12.2="22-23"  
13.2="24-25" 14.2="26-27" 15.2="28-29" 16.2="30-31"  
17.2="32-33" 18.2="34-35" 19.2="36-37" 20.2="38-39"  
21.2="40-41" 22.2="42-43" 23.2="44-45" 24.2="46-47"  
25.2="48-49" 26.2="50-51" 27.2="52-53" 28.2="54-55"  
29.2="56-57" 30.2="58-59" 31.2="60-61" 32.2="62-63"  
33.2="64-65" 34.2="66-67" 35.2="68-69" 36.2="70-71"  
37.2="72-73" 38.2="74-75" 39.2="76-77" 40.2="78-79"  
41.2="80-81" 42.2="82-83" 43.2="84-85" 44.2="86-87"  
45.2="88-89" 46.2="90-91" 47.2="92-93" 48.2="94-95"  
49.2="96-97" 50.2="98-99" 51.2="100-101" 52.2="102-103"  
53.2="104-105" 54.2="106-107" 55.2="108-109" 56.2="110-111"  
57.2="112-113" 58.2="114-115" 59.2="116-117" 60.2="118-119"  
61.2="120-121"

1.3="0-2" 2.3="3-5" 3.3="6-8" 4.3="9-11"  
5.3="12-14" 6.3="15-17" 7.3="18-20" 8.3="21-23"  
9.3="24-26" 10.3="27-29" 11.3="30-32" 12.3="33-35"  
13.3="36-38" 14.3="39-41" 15.3="42-44" 16.3="45-47"  
17.3="48-50" 18.3="51-53" 19.3="54-56" 20.3="57-59"

```
21.3="60-62" 22.3="63-65" 23.3="66-68" 24.3="69-71"
25.3="72-74" 26.3="75-77" 27.3="78-80" 28.3="81-83"
29.3="84-86" 30.3="87-89" 31.3="90-92" 32.3="93-95"
33.3="96-98" 34.3="99-101" 35.3="102-104" 36.3="105-107"
37.3="108-110" 38.3="111-113" 39.3="114-116" 40.3="117-119"
41.3="120-122"
```

```
1.4="0-3" 2.4="4-7" 3.4="8-11" 4.4="12-15"
5.4="16-19" 6.4="20-23" 7.4="24-27" 8.4="28-31"
9.4="32-35" 10.4="36-39" 11.4="40-43" 12.4="44-47"
13.4="48-51" 14.4="52-55" 15.4="56-49" 16.4="60-63"
17.4="64-67" 18.4="68-71" 19.4="72-75" 20.4="76-79"
21.4="80-83" 22.4="84-87" 23.4="88-91" 24.4="92-95"
25.4="96-99" 26.4="100-103" 27.4="104-107" 28.4="108-111"
29.4="112-115" 30.4="116-119" 31.4="120-123"
```

```
1.5="0-4" 2.5="5-9" 3.5="10-14" 4.5="15-19"
5.5="20-24" 6.5="25-29" 7.5="30-34" 8.5="35-39"
9.5="40-44" 10.5="45-49" 11.5="50-54" 12.5="55-59"
13.5="60-64" 14.5="65-69" 15.5="70-74" 16.5="75-79"
17.5="80-84" 18.5="85-89" 19.5="90-94" 20.5="95-99"
21.5="100-104" 22.5="105-109" 23.5="110-114" 24.5="115-119"
25.5="120-124";
```

```
RUN;
```

```
%MEND FORMATTING;
```

```
/* IMPORT RAW CSV DATA AND SAVE TO MIDAS LIBRARY, DROPPING UNWANTED
FIELDS */
/* ONLY NEED TO RUN THIS THE VERY FIRST TIME WHEN IMPORT IS REQUIRED */
```

```
%MACRO IMPORT; /* ONLY NEED TO RUN THIS ONCE */
```

```
PROC IMPORT OUT= WORK.MIDAS_NY
DATAFILE= "C:\Research\EpiSim\MIDAS\NY\synth_people.csv"
DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;
```

```
DATA MIDAS.NY_full;
SET MIDAS_NY;
DROP pub_trans_id pnum st_serialno;
RUN;
```

```
%MEND IMPORT;
```

```
%MACRO SET_LIBRARIES;
```

```
/* VM LIBRARY SET */
/* NOTE: TO SET THE WORK LIBRARY TO THE RAM DISK, CLOSE SAS, RIGHT
CLICK THE SAS ICON, GOTO THE SHORTCUT TAB, AND ADD */
/* -work "V:\WORK" at the end of the entry in "TARGET", so that it reads as follows: */
/* "C:\Program Files\SAS\SAS 9.1\sas.exe" -CONFIG "C:\Program Files\SAS\SAS
9.1\nls\en\SASV9.CFG" -work "V:\WORK" */
```

```
/* LOCAL LIBRARY SET */
/* LIBNAME MIDAS "C:\Research\MODELS\US\NewYork";
RUN;
```

```

LIBNAME MIDAS "\\10.159.2.146\SASMain$\SASUSERS\PABDELMA\Transform\NewYork';
RUN;
*/

LIBNAME MIDAS "V:\RESEARCH\NY";

%MEND SET_LIBRARIES;

%MACRO PREP_DATA;                                /* ONLY NEED TO RUN THIS ONCE */

/* CONVERT LATITUDE AND LONGITUDE FROM DEGREES TO RADIANS */

DATA MIDAS.popDATA;
  SET MIDAS.NY_full;
  /* ASSIGN popDATA DATASET TO THE POPULATION BEING USED HERE*/

  latRad = lat/57.295779513082320876798154814105;
  /* Latitude in Radians (i.e. divide by (180/pi)) */

  longRad = long/57.295779513082320876798154814105;
  /* Longitude in Radians (i.e. divide by (180/pi))*/

  RECORDNUM = _n_;
  DROP hh_id school_id work_id;
RUN;

PROC SORT DATA=MIDAS.popDATA;
  BY person_id;
RUN;

/*CREATE AN EMPTY DATASET TEMPLATE TO BE USED BY VARIOUS OTHER
PROCEDURES */

DATA MIDAS.emptyDATASET;
  SET MIDAS.popDATA;
  WHERE RECORDNUM = 1;
  KEEP RECORDNUM;
RUN;

DATA MIDAS.prepTRANSFORMED;
  SET MIDAS.emptyDATASET;
RUN;

DATA MIDAS.popDATA;
  SET MIDAS.popDATA;
  DROP RECORDNUM;
RUN;

/* DATASETS THAT WILL BE USED REPEATEDLY ARE WRITTEN TO A
PERMANENT LIBRARY (MIDAS) IN ORDER TO ALLOW LOOP PURGING OF THE
WORK DIRECTORY */
/* TO SAVE SPACE EACH TIME THE LOOPS ARE RUN */

```

**%MEND** PREP\_DATA;

```
/* THE FOLLOWING MACRO WILL ADD FIVE AGEGROUPS TO EVERY RECORD */
/* THE FIRST AGEGROUP IS THE AGE, INCREMENTING IN ONE YEAR INTERVALS */
/* THE SECOND AGEGROUP INCREMENTS IN TWO-YEAR INTERVALS */
/* THE THIRD AGEGROUP INCREMENTS IN THREE-YEAR INTERVALS, etc. */
```

**%MACRO** CATEGORISE\_AGE; /\* ONLY NEED TO RUN THIS ONCE \*/

```
PROC SORT DATA=MIDAS.popDATA;
  BY AGE;
RUN;
DATA MIDAS.popDATA;
  SET MIDAS.popDATA;
  CALL SYMPUT ('MAXAGE',age);
RUN;
```

**%DO** i = 1 %TO 5;

```
  DATA MIDAS.popDATA;
    SET MIDAS.popDATA;
    IF AGE < 1 THEN tempAGE = 1;
    ELSE tempAGE = AGE;

    DO agegroup = 1 TO &MAXAGE;
      IF &i = 1 THEN AGEGROUP1 = AGE;
      ELSE DO;
        max = (&i * agegroup)-1;
        min = max - (&i-1);
        IF tempAGE LE max AND tempAGE GE min THEN
          AGEGROUP&i = agegroup;
      END;
    END;
    DROP min max agegroup tempAGE;

    /*FORMAT AGEGROUP2 secondAG. AGEGROUP3 thirdAG.
    AGEGROUP4 fourthAG. AGEGROUP5 fifthAG.*/
  RUN;
```

**%END**;

```
PROC SORT DATA=MIDAS.popDATA;
  BY person_id;
RUN;
```

**%MEND** CATEGORISE\_AGE;

**%MACRO** CHOOSE\_CASES(PATIENTNUM);

```
/* PATIENTNUM IS THE NUMBER OF PATIENTS TO SIMULATE */
```

```
SASFILE MIDAS.popDATA LOAD;
/* LOAD POPULATION DATA INTO RAM TO SPEED UP SAMPLE
SELECTION */
```

```
PROC SURVEYSELECT DATA=MIDAS.popDATA
  OUT=Cases
  METHOD=SRS
  N=&PATIENTNUM
  NOPRINT
  ;
```

```
RUN;
```



```

DATA Cases;
  SET Cases;
  RECORDNUM = _n_;
  patient_LatRad = latRad;
  patient_LongRad = longRad;
  patient_lat = lat;
  patient_long = long;
  patientID = person_ID;
  patientAGE = AGE;
  patientSEX = SEX;
RUN;

/* REMOVE ALL THE CASES FROM THE BASE POPULATION TO PREVENT
TRANSFORMING ONE CASE WITH ANOTHER CASE */

PROC SQL NOPRINT;
  CREATE TABLE POP_noCASES AS
    SELECT pop.*, Cases.patientID
      FROM MIDAS.popDATA pop LEFT JOIN Cases
        ON pop.person_ID = Cases.patientID;
QUIT;

DATA POP_noCASES;
  SET POP_noCASES;
  IF patientID NE "" THEN DELETE;
  FLAG = 0;
RUN;

SASFILE MIDAS.popDATA CLOSE;
/* CLEAR POPULATION DATA FROM RAM */

```

**%MEND** CHOOSE\_CASES;

**%MACRO** LIMIT\_EXTENT;

```

%LET d = 1;
/* DISTANCE THRESHOLD; HERE, 1KM */

%LET nd = &d/(SQRT((3*6378.137**2 + 6356.9**2)/4));
/* NORMALISE LINEAR DISTANCE TO EARTH'S QUADRATIC RADIUS TO
FACILITATE WORKING IN RADIANS */

PROC SQL NOPRINT;
  SELECT
    ARSIN((SIN(MIN(patient_LatRad))*COS(&nd))+
(COS(MIN(patient_LatRad))*SIN(&nd)*COS(constant('pi'))))
      INTO : MIN_LatRad
    FROM Cases;
  SELECT
    ARSIN((SIN(MAX(patient_LatRad))*COS(&nd))+
(COS(MAX(patient_LatRad))*SIN(&nd)*COS(0)))
      INTO : MAX_LatRad
    FROM Cases;
  SELECT MIN(patient_LongRad) +
    ATAN2((SIN((270*(constant('pi')))/180)*SIN(&nd)*COS(MIN(patient_LatRad))),
(COS(&nd)-SIN(MIN(patient_LatRad))*SIN(MIN(patient_LatRad))))
      INTO : MIN_LongRad
    FROM Cases;
  SELECT MAX(patient_LongRad) +
    ATAN2((SIN((90*(constant('pi')))/180)*SIN(&nd)*COS(MAX(patient_LatRad))),
(COS(&nd)-SIN(MAX(patient_LatRad))*SIN(MAX(patient_LatRad))))
      INTO : MAX_LongRad
  FROM Cases;

```

```
QUIT;

DATA POP_noCASES;
  SET POP_noCASES;
  IF LatRad < &MIN_LatRad OR LatRad > &MAX_LatRad OR LongRad <
    &MIN_LongRad OR LongRad > &MAX_LongRad THEN DELETE;
  DROP patientID;
RUN;
```

```
%MEND LIMIT_EXTENT;
```

```
%MACRO TRANSFORM_CASES;
```

```
%LET k=5; /* SET THE DESIRED K-ANONYMITY LEVEL */
```

```
PROC SQL NOPRINT;
  CREATE TABLE POP_useME AS
    SELECT *
      FROM POP_noCASES;

  SELECT COUNT(*)
    INTO : RECNUMS
      FROM Cases;
```

```
QUIT;
```

```
/* AT THE BEGINNING OF THIS MACRO, prepTRANSFORMED IS AN EMPTY DATASET */
```

```
DATA TRANSFORMED;
  SET prepTRANSFORMED;
RUN;
```

```
/* BEGIN LOOP FOR EACH PATIENT RECORD */
```

```
/******  
/*          LOGIC FOR EACH PATIENT RECORD:          */  
/*          */  
/* 1. SELECT BASE POPULATION WITH ANY ALREADY CHOSEN POINTS */  
/*    REMOVED */  
/* 2. SELECT ALL RECORDS FROM BASE POP. MATCHING CASE ON    */  
/*    AGE AND SEX */  
/* 3. COUNT NUMBER OF MATCHES */  
/* 4. IF FEWER THAN "k-1" MATCHES (I.E. COUNT INCLUDING CASE */  
/*    LESS THAN "k") THEN */  
/*       NEW AGE BECOMES CATEGORISED AGE AND REPEAT */  
/*       FROM STEP 2 ON, UNTIL AT LEAST "k-1" MATCHES, OR */  
/*       MAXIMUM AGE CATEGORISATION REACHED */  
/* 5. CALCULATE DISTANCE BETWEEN ALL MATCHES AND CASE, */  
/*    INCLUDING CASE (d=0) */  
/* 6. TAKE CLOSEST "k", AND CHECK TO SEE MAXIMUM IS WITHIN */  
/*    THRESHOLD. */  
/*    IF NOT, GOTO STEP 4. AS LONG AS MAXIMUM AGE */  
/*    CATEGORISATION NOT REACHED */  
/*    IF DISTANCE THRESHOLD REQUIREMENT MET, PROCEED */  
/* 7. STORE MAXIMUM DISTANCE AS PERTURBATION RADIUS */  
/* 8. SET EXTENT FOR SELECTION BASED ON SQUARE FORMED BY */  
/*    PERTURBATION RADIUS */  
/* 9. SELECT ALL RECORDS FROM (1.) WITHIN EXTENT */  
/* 10. CALCULATE DISTANCE BETWEEN RECORDS AND CASE */  
/* 11. SELECT ALL RECORDS WHERE DISTANCE LESS THAN OR EQUAL */  
/*     TO PERTURBATION RADIUS PLUS A RANDOMLY DEFINED BUFFER */  
/*     BETWEEN 1 AND 10 METRES */  
/* 12. STORE COUNT, MEAN (X,Y), STD (X,Y), MEDIAN (X,Y) */  
/* 13. RANDOMLY SELECT ONE RECORD; THIS IS THE TRANSFORMED */  
/*     RECORD */  
/******
```

```
%DO DCALC = 1 %TO &RECNUMS;
```

```
/******  
/* 1. SELECT BASE POPULATION WITH ANY ALREADY CHOSEN POINTS */  
/*    REMOVED */  
/******
```

```
DATA POP_useME;  
  SET POP_useME;  
  WHERE FLAG = 0;  
  /* FLAG IS USED TO PREVENT RE-SELECTION OF ANY GIVEN  
  RECORD */
```

```
  KEEP person_id age sex lat long latRad longRad AGEGROUP1  
  AGEGROUP2 AGEGROUP3 AGEGROUP4 AGEGROUP5 FLAG;
```

```
RUN;
```

```
DATA selectedCase;  
  SET Cases;  
  WHERE RECORDNUM = &DCALC;  
  CALL SYMPUT ('patientSEX',SEX);  
  CALL SYMPUT ('patientAGE',patientAGE);  
  CALL SYMPUT ('patient_Lat',patient_Lat);  
  CALL SYMPUT ('patient_Long',patient_Long);  
  CALL SYMPUT ('patient_LatRad',patient_LatRad);  
  CALL SYMPUT ('patient_LongRad',patient_LongRad);  
  CALL SYMPUT ('patientID', patientID);
```

```
RUN;
```

```
/* THIS LOOP ALLOWS TRANSFORMATION OF THE AGE DIMENSION
WHEN THE SPATIAL DIMENSION ALONE IS INSUFFICIENT TO ACHIEVE
DESIRED ANONYMISATION */
```

```
%DO AGEDIM = 1 %TO 5;
```

```
PROC SQL NOPRINT;
    SELECT AGEGROUP&AGEDIM INTO: pAGEGROUP FROM
    selectedCase;
QUIT;
```

```
/* *****
/* 2. SELECT ALL RECORDS FROM BASE POP. MATCHING CASE ON    */
/*    AGE AND SEX                                           */
/* *****
```

```
DATA matchingPOP;
    SET POP_useME selectedCase;
    WHERE SEX = &patientSEX AND AGEGROUP&AGEDIM =
    &pAGEGROUP;
RUN;
```

```
/* *****
/* 3. COUNT NUMBER OF MATCHES                               */
/* *****
```

```
/* CHECK TO MAKE SURE THAT THERE ARE AT LEAST "k" RECORDS IN TOTAL
THAT MATCH ON AGE AND SEX */
```

```
PROC SQL NOPRINT;
    SELECT COUNT(*) INTO : MATCHES1 FROM matchingPOP;
QUIT;
```

```
%IF &MATCHES1 GE &k %THEN %LET THRESHOLD1 = PASS;
%ELSE %LET THRESHOLD1 = FAIL;
```

```
/* *****
/* 4. IF FEWER THAN "k-1" MATCHES (I.E. COUNT INCLUDING CASE */
/*    LESS THAN "k") THEN                                   */
/*    NEW AGE BECOMES CATEGORISED AGE AND REPEAT         */
/*    FROM STEP 2 ON, UNTIL AT LEAST "k-1" MATCHES, OR   */
/*    MAXIMUM AGE CATEGORISATION REACHED                 */
/* *****
```

```
%IF &THRESHOLD1 = FAIL AND &AGEDIM NE 5 %THEN %GOTO
NEXT_AGEGROUP;
```

```
%ELSE %DO;
```

```
/* IF PASS (I.E. AT LEAST "k" RECORDS IN TOTAL THAT MATCH
ON AGE AND SEX) THEN CALCULATE DISTANCES */
/* NOTE THAT BY APPENDING THE PATIENT RECORD, WE
ALLOW IT TO BE SELECTED AS WELL AS THE "TRANSFORMED"
DATA POINT */
```

```
/* *****
/* 5. CALCULATE DISTANCE BETWEEN ALL MATCHES AND CASE,    */
/*    INCLUDING CASE (d=0)                                  */
/* *****
```

```

DATA Distance_Calc;
  SET matchingPOP;

  IF patient_Lat = . THEN patient_Lat = &patient_lat;
  IF patient_Long = . THEN patient_Long =
  &patient_Long;
  IF patientID = "" THEN patientID = "&patientID";
  IF patientAGE = . THEN patientAGE = &patientAGE;
  IF patientSEX = . THEN patientSEX = &patientSEX;
  IF patient_LatRad = . THEN patient_LatRad =
  &patient_LatRad;
  IF patient_LongRad = . THEN patient_LongRad =
  &patient_LongRad;

  PolarRadius = 6356.9;
  EquatorialRadius = 6378.137;
  /*Geod ref. sys. 1980 */
  MeanRadius = 6378.01;
  /*NASA: http://ssd.jpl.nasa.gov/phys\_props\_earth.html;
  +/-0.02Kms */
  QuadraticRadius = SQRT((3*EquatorialRadius**2 +
  PolarRadius**2)/4);

  A = SIN(LatRad);
  B = COS(LatRad);

  C = SIN(patient_LatRad);
  D = COS(patient_LatRad);

  E = COS(LongRad - patient_LongRad);

  F = A*C;
  G = B*D*E;

  H = F+G;

  IF LatRad = patient_LatRad AND LongRad =
  patient_LongRad THEN I = 0;

  ELSE I = ARCOS(H);

  distance = QuadraticRadius*I;

  DROP A B C D E F G H I PolarRadius
  EquatorialRadius MeanRadius QuadraticRadius;
RUN;

```

```

/*****/
/* 6. TAKE CLOSEST "k", AND CHECK TO SEE MAXIMUM IS WITHIN */
/* THRESHOLD. */
/* IF NOT, GOTO STEP 4. AS LONG AS MAXIMUM AGE */
/* CATEGORISATION NOT REACHED */
/* IF DISTANCE THRESHOLD REQUIREMENT MET, PROCEED */
/*****/

```

```

/* ASSESS WHETHER OR NOT AT LEAST "k" TOTAL MATCHING CASES WITHIN
SPECIFIED DISTANCE THRESHOLD */

```

```

PROC SQL NOPRINT;
    CREATE TABLE WITHIN_D AS SELECT * FROM
    Distance_Calc WHERE distance <= 1;
    SELECT COUNT(*) INTO :nWITHINDIST FROM
    WITHIN_D;

```

```

QUIT;

```

```

%IF &nWITHINDIST GE &k %THEN %LET THRESHOLD2 =
PASS;
%ELSE %LET THRESHOLD2 = FAIL;

```

```

%IF &THRESHOLD2 = FAIL AND &AGEDIM NE
5 %THEN %GOTO NEXT_AGEGROUP;
%ELSE %DO;

```

```

/*****/
/* 7. STORE MAXIMUM DISTANCE AS PERTURBATION RADIUS */
/*****/

```

```

/* IF ALL THRESHOLDS MAINTAINED, OR IF ON LAST AGE GROUP, THEN */
/* SELECT THE CLOSEST (i.e. FIRST) FIVE TO THE PATIENT */

```

```

PROC SQL NOPRINT OUTOBS = &k;
    CREATE TABLE TOP&k AS SELECT * FROM
    Distance_Calc ORDER BY distance;
    SELECT MAX(distance) INTO :pertRADIUS
    FROM TOP&k;
    /* Maximum radius for perturbation */

```

```

QUIT;

```

```

/*****/
/* 8. SET EXTENT FOR SELECTION BASED ON SQUARE FORMED BY */
/* PERTURBATION RADIUS PLUS A RANDOM BUFFER SIZE WITHIN */
/* PREDEFINED THRESHOLD (E.G. 1 TO 10 METRES) */
/*****/

```

```

%LET dbuff = &pertRADIUS +
(((RANUNI(0)*9)+1)/1000);

```

```

/* PERTURBATION RADIUS WITH AN ADDED RANDOM BUFFER BETWEEN 1 & 10
METRES*/

```

```

/* THIS ALSO MAKES RE-IDENTIFICATION MORE DIFFICULT */

```

```

/* SET EXTENT COORDINATES FOR SELECTION FOR ACTIVE SELECTED CASE
*/

```

```

/* VARIANT: Can be altered based on dimension priorities; for example, can refine distance
within an age loop instead of age within the distance loop if minimum age distortion is prioritised
over distance */

```

```

%LET d = &dbuf;
%LET nd = &d /(SQRT((3*6378.137**2 +
6356.9**2)/4)); /* NORMALISED DISTANCE */

PROC SQL NOPRINT;
  SELECT
    ARSIN((SIN(patient_LatRad)*COS(&nd))+COS
    S(patient_LatRad)*SIN(&nd)*COS(constant('pi'
    )))
      INTO : MIN_LatRad
      FROM selectedCase;

  SELECT
    ARSIN((SIN(patient_LatRad)*COS(&nd))+COS
    S(patient_LatRad)*SIN(&nd)*COS(0)))
      INTO : MAX_LatRad
      FROM SelectedCase;

  SELECT patient_LongRad +
    ATAN2((SIN((270*(constant('pi')))/180)*SIN(&n
    d)*COS(patient_LatRad)),(COS(&nd)-
    SIN(patient_LatRad)*SIN(patient_LatRad)))
      INTO : MIN_LongRad
      FROM SelectedCase;

  SELECT patient_LongRad +
    ATAN2((SIN((90*(constant('pi')))/180)*SIN(&nd
    )*COS(patient_LatRad)),(COS(&nd)-
    SIN(patient_LatRad)*SIN(patient_LatRad)))
      INTO : MAX_LongRad
      FROM selectedCase;

QUIT;

/*****
/* 9. SELECT ALL RECORDS FROM (1.) WITHIN EXTENT */
*****/

DATA limited_POP_useME;
  SET POP_useME selectedCase;

  IF LatRad < &MIN_LatRad OR LatRad >
  &MAX_LatRad OR LongRad <
  &MIN_LongRad OR LongRad >
  &MAX_LongRad THEN DELETE;

RUN;

PROC SQL NOPRINT;
  SELECT COUNT(*) INTO :
  LIMITED_squareCOUNT FROM
  limited_POP_useME;
QUIT;

```

```

/*****
/* 10. CALCULATE DISTANCE BETWEEN LIMITED EXTENT RECORDS */
/* AND CASE */
*****/

DATA limited_POP_useME;
SET limited_POP_useME;

IF patient_Lat = . THEN patient_Lat =
&patient_lat;

IF patient_Long = . THEN patient_Long =
&patient_Long;

IF patientID = "" THEN patientID =
"&patientID";

IF patientAGE = . THEN patientAGE =
&patientAGE;

IF patientSEX = . THEN patientSEX =
&patientSEX;

IF patient_LatRad = . THEN patient_LatRad =
&patient_LatRad;

IF patient_LongRad = . THEN
patient_LongRad = &patient_LongRad;

PolarRadius = 6356.9;
EquatorialRadius = 6378.137;
/*Geod ref. sys. 1980 */
MeanRadius = 6378.01;
/*NASA:
http://ssd.jpl.nasa.gov/phys\_props\_earth.html;
+/-0.02Kms */
QuadraticRadius =
SQRT((3*EquatorialRadius**2 +
PolarRadius**2)/4);

A = SIN(LatRad);
B = COS(LatRad);

C = SIN(patient_LatRad);
D = COS(patient_LatRad);

E = COS(LongRad - patient_LongRad);

F = A*C;
G = B*D*E;

H = F+G;

IF LatRad = patient_LatRad AND LongRad =
patient_LongRad THEN I = 0;

ELSE I = ARCOS(H);

distance = QuadraticRadius*I;

```



DROP A B C D E F G H I PolarRadius  
EquatorialRadius MeanRadius  
QuadraticRadius;

RUN;

```
/******  
/* 11. SELECT ALL RECORDS WHERE DISTANCE LESS THAN OR EQUAL */  
/* TO PERTURBATION RADIUS PLUS A RANDOMLY DEFINED BUFFER */  
/* BETWEEN 1 AND 10 METRES */  
/******
```

DATA limited\_POP\_useME;  
SET limited\_POP\_useME;  
WHERE distance LE &dbuff;

RUN;

```
/******  
/* 12. STORE COUNT, MEAN (X,Y), STD (X,Y) */  
/******
```

```
PROC SQL NOPRINT;  
SELECT COUNT(*) INTO : radiusCOUNT  
FROM limited_POP_useME;  
  
SELECT MEAN(Lat) INTO : radiusMEANLAT  
FROM limited_POP_useME;  
  
SELECT MEAN(Long) INTO :  
radiusMEANLONG FROM  
limited_POP_useME;  
  
SELECT STD(Lat) INTO : radiusSTDLAT  
FROM limited_POP_useME;  
  
SELECT STD(Long) INTO : radiusSTDLONG  
FROM limited_POP_useME;  
  
SELECT MEAN(distance) INTO :  
radiusMEANDIST FROM limited_POP_useME;  
  
SELECT STD(distance) INTO :  
radiusSTDDIST FROM limited_POP_useME;  
  
SELECT COUNT(*) INTO :  
nMATCHESINpertR FROM  
limited_POP_useME  
WHERE SEX = &patientSEX AND  
AGEGROUP&AGEDIM =  
&pAGEGROUP;
```

QUIT;

```

/*****
/* 13. RANDOMLY SELECT ONE RECORD; THIS IS THE TRANSFORMED */
/* RECORD */
*****/

```

```

PROC SURVEYSELECT
  DATA=limited_POP_useME
  OUT=TRANSFORMED_CASE
  METHOD=SRS
  N=1
  NOPRINT
  ;
RUN;

DATA TRANSFORMED_CASE;
  SET TRANSFORMED_CASE;
  THRESHOLD1 = "&THRESHOLD1";
  MATCHES1 = &MATCHES1;
  THRESHOLD2 = "&THRESHOLD2";
  nWITHINDIST = &nWITHINDIST;
  RECORDNUM = &DCALC;
  AGEGROUP = &AGEDIM;
  newAGE = &pAGEGROUP+(&AGEDIM/10);

  IF &AGEDIM = 1 THEN DO;
    maxAGE = patientAGE;
    minAGE = patientAGE;
    midAGE = patientAGE;
  END;

  ELSE DO;
    maxAGE =
      (&AGEDIM*&pAGEGROUP)-1;
    minAGE = maxAGE-(&AGEDIM-1);
    midAGE = (maxAGE+minAGE)/2;
  END;

  FLAG = 1;
  FORMAT newAGE newAGE.;
  pertR = &pertRADIUS;
  buffR = &d;
  nMATCHESINpertR = &nMATCHESINpertR;
  radiusCOUNT = &radiusCOUNT;
  radiusMEANLONG = &radiusMEANLONG;
  radiusMEANLAT = &radiusMEANLAT;
  radiusSTDLONG = &radiusSTDLONG;
  radiusSTDLAT = &radiusSTDLAT;
  radiusMEANDIST = &radiusMEANDIST;
  radiusSTDDIST = &radiusSTDDIST;

  KEEP RECORDNUM person_ID sex age
  patientAGE patientSEX patientID
  patient_Lat patient_Long Lat Long
  distance THRESHOLD1 MATCHES1
  THRESHOLD2 nWITHINDIST FLAG
  AGEGROUP1 AGEGROUP2
  AGEGROUP3 AGEGROUP4
  AGEGROUP5 AGEGROUP newAGE
  maxAGE minAGE midAGE pertR
  buffR nMATCHESINpertR
  radiusCOUNT radiusMEANLONG
  radiusMEANLAT radiusSTDLONG

```

```

radiusSTDLAT radiusMEANDIST
radiusSTDDIST;
RUN;

DATA TRANSFORMED;
SET TRANSFORMED
TRANSFORMED_CASE;
IF patientID = "" THEN DELETE;
RUN;

/* THIS BIT IS THE FOLLOW-UP THAT ALLOWS US TO PREVENT SELECTION OF OTHER
PATIENTS IN THE LIST */

DATA POP_useME;
MERGE POP_useME
TRANSFORMED_CASE;
BY person_ID;
RUN;
%GOTO NEXT_RECORD;
%END;
%END;
%NEXT_AGEGROUP:

%END;
%NEXT_RECORD:
%END;

%MEND TRANSFORM_CASES;

%MACRO OUTPUT_STATS;

DATA TRANSFORMED;
SET TRANSFORMED;
RUN;

/* CALCULATE DEMOGRAPHICS */

PROC MEANS NOPRINT DATA=TRANSFORMED;
VAR patientAGE;
OUTPUT OUT = CASES_AGE
N = CASES
MIN = MinpAGE
MAX = MaxpAGE
MEDIAN = MedianpAGE
MEAN = MeanpAGE
STD = stdpAGE;
RUN;
DATA CASES_AGE;
SET CASES_AGE;
RECORDNUM = _n_;
RUN;

PROC MEANS NOPRINT DATA=TRANSFORMED;
VAR midAGE;
OUTPUT OUT = tCASES_AGE
MIN = MintAGE
MAX = MaxtAGE
MEDIAN = MediantAGE
MEAN = MeantAGE
STD = stdtAGE;
RUN;

DATA tCASES_AGE;

```

```

        SET tCASES_AGE;
        RECORDNUM = _n_;
RUN;

PROC MEANS NOPRINT DATA=TRANSFORMED;
    VAR patientSEX;
    WHERE patientSEX = 1;
    OUTPUT      OUT    = oCASES_MALES
              N      = oMALES;

RUN;
DATA oCASES_MALES;
    SET oCASES_MALES;
    RECORDNUM = _n_;
RUN;

PROC MEANS NOPRINT DATA=TRANSFORMED;
    VAR patientSEX;
    WHERE patientSEX = 2;
    OUTPUT      OUT    = oCASES_FEMALES
              N      = oFEMALES;

RUN;
DATA oCASES_FEMALES;
    SET oCASES_FEMALES;
    RECORDNUM = _n_;
RUN;

PROC MEANS NOPRINT DATA=TRANSFORMED;
    VAR SEX;
    WHERE SEX = 1;
    OUTPUT      OUT    = tCASES_MALES
              N      = MALES;

RUN;
DATA tCASES_MALES;
    SET tCASES_MALES;
    RECORDNUM = _n_;
RUN;

PROC MEANS NOPRINT DATA=TRANSFORMED;
    VAR SEX;
    WHERE SEX = 2;
    OUTPUT      OUT    = tCASES_FEMALES
              N      = FEMALES;

RUN;
DATA tCASES_FEMALES;
    SET tCASES_FEMALES;
    RECORDNUM = _n_;
RUN;

/* CALCULATE STATS FOR TRANSFORMED POINT DISTANCES */

PROC MEANS NOPRINT DATA=TRANSFORMED;
    VAR distance;
    OUTPUT      OUT          = tCASES_dStats
              N              = CASES
              MIN             = MinDist
              MAX =          = MaxDist
              MEDIAN          = MedianDist
              MEAN =          = MeanDist
              STD              = stdDist;

RUN;
DATA tCASES_dStats;
    SET tCASES_dStats;
    RECORDNUM = _n_;

```

```

RUN;

/* CALCULATE STATS FOR ORIGINAL LATITUDE (Y) */

PROC MEANS NOPRINT DATA=TRANSFORMED;
  VAR patient_Lat;
  OUTPUT      OUT          =      tCASES_oyStats
             MIN          =      minOY
             MAX          =      maxOY
             MEDIAN       =      medianOY
             MEAN =      meanOY
             STD          =      stdOY;

RUN;
DATA tCASES_oyStats;
  SET tCASES_oyStats;
  RECORDNUM = _n_;

RUN;

/* CALCULATE STATS FOR ORIGINAL LONGITUDE (X) */

PROC MEANS NOPRINT DATA=TRANSFORMED;
  VAR patient_Long;
  OUTPUT      OUT          =      tCASES_oxStats
             MIN          =      minOX
             MAX          =      maxOX
             MEDIAN       =      medianOX
             MEAN =      meanOX
             STD          =      stdOX;

RUN;
DATA tCASES_oxStats;
  SET tCASES_oxStats;
  RECORDNUM = _n_;

RUN;

/* CALCULATE STATS FOR TRANSFORMED LATITUDE (Y) */

PROC MEANS NOPRINT DATA=TRANSFORMED;
  VAR lat;
  OUTPUT      OUT          =      tCASES_tyStats
             MIN          =      minTY
             MAX          =      maxTY
             MEDIAN       =      medianTY
             MEAN =      meanTY
             STD          =      stdTY;

RUN;
DATA tCASES_tyStats;
  SET tCASES_tyStats;
  RECORDNUM = _n_;

RUN;

/* CALCULATE STATS FOR TRANSFORMED LONGITUDE (X) */

PROC MEANS NOPRINT DATA=TRANSFORMED;
  VAR long;
  OUTPUT      OUT          =      tCASES_txStats
             MIN          =      minTX
             MAX          =      maxTX
             MEDIAN       =      medianTX
             MEAN =      meanTX
             STD          =      stdTX;

RUN;
DATA tCASES_txStats;
  SET tCASES_txStats;
  RECORDNUM = _n_;

```

```

RUN;

/* ADD NUMBER OF FAILS */

PROC MEANS NOPRINT DATA=TRANSFORMED;
  VAR RECORDNUM;
  WHERE THRESHOLD2 = "FAIL";
  OUTPUT      OUT      =      tCASES_FAILS
             N          =      FAILS;

RUN;
DATA tCASES_FAILS;
  SET tCASES_FAILS;
  RECORDNUM = _n_;
RUN;

/* ADD MATCH NUMBER STATS */

PROC MEANS NOPRINT DATA=TRANSFORMED;
  VAR nWITHINDIST;
  OUTPUT      OUT      =      tCASES_MATCHES
             MIN      =      minMATCHES
             MAX      =      maxMATCHES
             MEDIAN   =      medianMATCHES
             MEAN     =      meanMATCHES
             STD      =      stdMATCHES;

RUN;
DATA tCASES_MATCHES;
  SET tCASES_MATCHES;
  RECORDNUM = _n_;
RUN;

PROC MEANS NOPRINT DATA=TRANSFORMED;
  VAR nMATCHESINpertR;
  OUTPUT      OUT      =      tCASES_MATCHESpertR
             MIN      =      minMATCHESpertR
             MAX      =      maxMATCHESpertR
             MEDIAN   =      medianMATCHESpertR
             MEAN     =      meanMATCHESpertR
             STD      =      stdMATCHESpertR;

RUN;
DATA tCASES_MATCHESpertR;
  SET tCASES_MATCHESpertR;
  RECORDNUM = _n_;
RUN;

/* ADD NUMBER OF TIMES ORIGINAL LOCATION CHOSEN */

PROC MEANS NOPRINT DATA=TRANSFORMED;
  VAR RECORDNUM;
  WHERE person_ID = patientID;
  OUTPUT      OUT      =      tCASES_ORIGINAL
             N          =      ORIGINAL;

RUN;
DATA tCASES_ORIGINAL;
  SET tCASES_ORIGINAL;
  RECORDNUM = _n_;
RUN;

/* PUT THEM ALL TOGETHER */

DATA tSTATS;

```

```

MERGE      CASES_AGE tCASES_AGE oCASES_MALES
           oCASES_FEMALES tCASES_MALES tCASES_FEMALES
           tCASES_dStats tCASES_oySTATS tCASES_oxSTATS
           tCASES_tyStats tCASES_txStats
           tCASES_FAILS tCASES_MATCHES
           tCASES_MATCHESpertR tCASES_ORIGINAL;

BY RECORDNUM;
IF FAILS = . THEN FAILS = 0;
IF ORIGINAL = . THEN ORIGINAL = 0;
IF MALES = . THEN MALES = 0;
IF FEMALES = . THEN FEMALES = 0;
DROP _TYPE_ _FREQ_;

RUN;

%MEND OUTPUT_STATS;

%MACRO RUNIT;

  %HOUSEKEEPING;
  %FORMATTING;
  %SET_LIBRARIES;

  %PREP_DATA;
  %CATEGORISE_AGE;

  DATA _NULL_;
    CALL SYMPUT ('currentTIME',PUT (TIME(),TIME.));
    CALL SYMPUT ('currentDATE',PUT (DATE(),DATE.));
  RUN;
  %PUT BEGINNING ALGORITHM: &currentTIME ON &currentDATE;

  /* NUMBER OF ITERATIONS FOR EACH PATIENT-LIST SIZE (SUGGEST 1,000)*/
  %LET ITERATIONS = 1000;

  /* NUMBER OF DIFFERENT PATIENT-LIST SIZES (CAN ADD MORE / MODIFY AS
  DESIRED */
  %DO ploop = 1 %TO 5;

    %IF &ploop = 1 %THEN %LET SIZE = 25;
    %ELSE %IF &ploop = 2 %THEN %LET SIZE = 50;
    %ELSE %IF &ploop = 3 %THEN %LET SIZE = 100;
    %ELSE %IF &ploop = 4 %THEN %LET SIZE = 200;
    %ELSE %IF &ploop = 5 %THEN %LET SIZE = 400;

    DATA emptyDATASET;
      SET MIDAS.emptyDATASET;
    RUN;

    DATA prepTRANSFORMED;
      SET MIDAS.prepTRANSFORMED;
    RUN;

    DATA totalSTATS_&SIZE;
      SET emptyDATASET;
    RUN;

    DATA ALL_tCASES_&SIZE;
      SET emptyDATASET;
    RUN;

    %DO within = 1 %TO &ITERATIONS;

```

```
%PUT THIS IS RUN NUMBER &within IN SAMPLE GROUP SIZE  
&ploop;
```

```
%CHOOSE_CASES(&SIZE);  
%LIMIT_EXTENT;  
%TRANSFORM_CASES;  
%OUTPUT_STATS;
```

```
DATA tCASES;  
    SAMPLENUM = &within;  
    SET TRANSFORMED;  
RUN;
```

```
DATA tSTATS;  
    SAMPLENUM = &within;  
    SET tSTATS;
```

```
RUN;  
DATA totalSTATS_&SIZE;  
    SET totalSTATS_&SIZE tSTATS;  
    IF CASES = . THEN DELETE;  
    DROP RECORDNUM;  
RUN;
```

```
DATA ALL_tCASES_&SIZE;  
    SET ALL_tCASES_&SIZE tCASES;  
    IF patientID = "" THEN DELETE;  
RUN;
```

```
%END;
```

```
DATA MIDAS.totalSTATS_&SIZE;  
    SET totalSTATS_&SIZE;
```

```
RUN;  
DATA MIDAS.ALL_tCASES_&SIZE;  
    SET ALL_tCASES_&SIZE;  
RUN;
```

```
DATA _NULL_;  
    CALL SYMPUT ('currentTIME',PUT (TIME(),TIME.));  
    CALL SYMPUT ('currentDATE',PUT (DATE(),DATE.));
```

```
RUN;  
%PUT ALGORITHM COMPLETE: &currentTIME ON &currentDATE;
```

```
%HOUSEKEEPING;  
%FORMATTING;
```

```
%END;
```

```
%MEND RUNIT;  
%RUNIT;
```



## E. Example Code Modification to Allow for Cluster Insertion

To create a cluster within the patient dataset, make the following **TWO** modifications:

1. Replace the “**CHOOSE\_CASES**” MACRO in Appendix D with the one below

```
%MACRO CHOOSE_CASES(PATIENTNUM, CLUSTERSIZE);

/* PATIENTNUM IS THE NUMBER OF PATIENTS TO SIMULATE; CLUSTERSIZE IS THE
SIZE OF THE CLUSTER TO BE CREATED */

    %LET NONCLUSTERSIZE = %EVAL(&PATIENTNUM - &CLUSTERSIZE);
    %LET CLUSTERSIZEb = %EVAL(&CLUSTERSIZE - 1);

    SASFILE MIDAS.popDATA LOAD;
/* LOAD POPULATION DATA INTO RAM TO SPEED UP SAMPLE SELECTION */

    %PUT SELECTING &PATIENTNUM CASES OF WHICH &CLUSTERSIZE ARE
CLUSTERED AND &NONCLUSTERSIZE ARE NOT;

/*****
/*          SELECT INDEX CASE FOR CLUSTER CREATION          */
*****/

PROC SURVEYSELECT          DATA=MIDAS.popDATA
                           OUT=iCase
                           METHOD=SRS
                           N=1
                           NOPRINT
                           ;

RUN;

DATA iCase;
  SET iCase;
  patient_LatRad = latRad;
  patient_LongRad = longRad;
  patient_Lat = Lat;
  patient_Long = Long;
  patientID = person_ID;
  CALL SYMPUT ("iPATIENT",patientID);
  CALL SYMPUT ("patient_Lat", patient_Lat);
  CALL SYMPUT ("patient_Long", patient_Long);
  CALL SYMPUT ("patient_LatRad", patient_LatRad);
  CALL SYMPUT ("patient_LongRad", patient_LongRad);

RUN;

DATA _NULL_;
  CALL SYMPUT ('currentTIME',PUT (TIME(),TIME.));
RUN;
%PUT INDEX CASE SELECTED: &currentTIME;
```

```

/*****
/* REMOVE INDEX CASE FROM BASE POPULATION TO PREP FOR CLUSTER */
/* CREATION */
/*****

PROC SQL NOPRINT;
    CREATE TABLE POP_noCASES AS
        SELECT *
            FROM MIDAS.popDATA
                WHERE person_ID NE "&iPATIENT";

QUIT;

/*****
/* SELECT "CLUSTERSIZE" CASES WITHIN CLUSTER DISTANCE OF INDEX CASE */
/*****

DATA _NULL_;
    CALL SYMPUT ('currentTIME',PUT (TIME(),TIME.));
RUN;
%PUT BEGINNING CLUSTER EXTENT DEFINITION AND SELECTION:
&currentTIME;

/* DEFINE RECTANGULAR EXTENT FOR SELECTION BASED ON CLUSTER DISTANCE
THRESHOLD */

%LET d = 1;

/* CLUSTERING DISTANCE THRESHOLD; HERE, 1KM*/

%LET nd = &d/(SQRT((3*6378.137**2 + 6356.9**2)/4));
/* NORMALISE DISTANCE */

PROC SQL NOPRINT;
    SELECT
        ARSIN((SIN(MIN(patient_LatRad))*COS(&nd))+(COS(MIN(patient_Lat
Rad))*SIN(&nd)*COS(constant('pi'))))
            INTO : MIN_LatRad
                FROM iCase;

    SELECT
        ARSIN((SIN(MAX(patient_LatRad))*COS(&nd))+(COS(MAX(patient_La
tRad))*SIN(&nd)*COS(0)))
            INTO : MAX_LatRad
                FROM iCase;

    SELECT MIN(patient_LongRad) +
        ATAN2((SIN((270*(constant('pi')))/180)*SIN(&nd)*COS(MIN(patient_Lat
Rad))),(COS(&nd)-
SIN(MIN(patient_LatRad))*SIN(MIN(patient_LatRad))))
            INTO : MIN_LongRad
                FROM iCase;

    SELECT MAX(patient_LongRad) +
        ATAN2((SIN((90*(constant('pi')))/180)*SIN(&nd)*COS(MAX(patient_Lat
Rad))),(COS(&nd)-
SIN(MAX(patient_LatRad))*SIN(MAX(patient_LatRad))))
            INTO : MAX_LongRad
                FROM iCase;

QUIT;

/* FILTER OUT THOSE OUTSIDE OF THE RECTANGULAR EXTENT AND APPEND THE
INDEX CASE */

```

```

DATA clusterEXTENT;
  SET POP_noCASES iCase;
  IF LatRad < &MIN_LatRad OR LatRad > &MAX_LatRad OR LongRad
  < &MIN_LongRad OR LongRad > &MAX_LongRad THEN DELETE;
RUN;

```

*/\* CALCULATE DISTANCE BETWEEN EACH RECORD IN THE EXTENT AND THE INDEX CASE \*/*

```

DATA clusterEXTENT;
  SET clusterEXTENT ;

  IF patient_Lat = . THEN patient_Lat = &patient_lat;
  IF patient_Long = . THEN patient_Long = &patient_Long;
  IF patientID = "" THEN patientID = "&iPATIENT";
  IF patient_LatRad = . THEN patient_LatRad = &patient_LatRad;
  IF patient_LongRad = . THEN patient_LongRad = &patient_LongRad;

  PolarRadius = 6356.9;
  EquatorialRadius = 6378.137; /*Geod ref. sys. 1980 */
  MeanRadius = 6378.01; /*NASA:
  http://ssd.jpl.nasa.gov/phys\_props\_earth.html; +/-0.02Kms */
  QuadraticRadius = SQRT((3*EquatorialRadius**2 + PolarRadius**2)/4);

  A = SIN(LatRad);
  B = COS(LatRad);

  C = SIN(patient_LatRad);
  D = COS(patient_LatRad);

  E = COS(LongRad - patient_LongRad);

  F = A*C;
  G = B*D*E;

  H = F+G;

  IF LatRad = patient_LatRad AND LongRad = patient_LongRad THEN I
  = 0;
  ELSE I = ARCOS(H);

  distance = QuadraticRadius*I;

  DROP A B C D E F G H I PolarRadius EquatorialRadius MeanRadius
  QuadraticRadius;
RUN;

```

*/\* REMOVE THE INDEX CASE TO PREVENT RE-SELECTION AND ALL RECORDS OUTSIDE THE CLUSTER DISTANCE THRESHOLD \*/*

```

DATA clusterEXTENT;
  SET clusterEXTENT;
  WHERE distance LE 1 AND person_ID NE patientID;
RUN;

```

*/\* RANDOMLY SELECT CLUSTER RECORDS BASED ON THE CLUSTER SIZE STRING PASSED TO THIS MACRO ("CLUSTERSIZE") \*/*

```

PROC SURVEYSELECT      DATA=clusterEXTENT
                      OUT=clusterCases
                      METHOD=SRS
                      N=&CLUSTERSIZEb

```

NOPRINT

```

RUN;

DATA _NULL_;
    CALL SYMPUT ('currentTIME',PUT (TIME(),TIME.));
RUN;
%PUT CLUSTER EXTENT DEFINITION AND SELECTION COMPLETED:
&currentTIME;

```

```

/*****/
/* REMOVE ALL CLUSTER EXTENT CASES FROM THE BASE POPULATION FOR */
/* SELECTION OF NON-CLUSTERED CASES */
/*****/

```

```

DATA cluster;
    SET clusterEXTENT iCase;
    clusterFLAG = 1;
    clusterID = person_ID;
    KEEP clusterID clusterFLAG;
RUN;

PROC SQL NOPRINT;
    CREATE TABLE nonCLUSTER AS
        SELECT *
            FROM MIDAS.popDATA pop LEFT JOIN cluster
                ON pop.person_ID = cluster.clusterID;
QUIT;

```

```

DATA nonCLUSTER;
    SET nonCLUSTER;
    WHERE clusterFLAG NE 1;
RUN;

```

```

/*****/
/* SELECT REMAINDER OF CASES */
/*****/

```

```

DATA _NULL_;
    CALL SYMPUT ('currentTIME',PUT (TIME(),TIME.));
RUN;
%PUT BEGINNING NON-CLUSTER SELECTION: &currentTIME;

```

```

PROC SURVEYSELECT      DATA=nonCLUSTER
                      OUT=nonclusterCases
                      METHOD=SRS
                      N=&NONCLUSTERSIZE
                      NOPRINT
                      ;
RUN;

```

```

DATA _NULL_;
    CALL SYMPUT ('currentTIME',PUT (TIME(),TIME.));
RUN;
%PUT NON-CLUSTER SELECTION COMPLETE: &currentTIME;

```

```

/*****/
/* CREATE FINAL CASES DATASET */
/*****/

```

```

DATA Cases;
    RECORDNUM = _n_;
    SET iCase clusterCASES nonclusterCASES;

```

```

        patientAGE = AGE;
        patientSEX = SEX;
        patient_LatRad = latRad;
        patient_LongRad = longRad;
        patient_lat = lat;
        patient_long = long;
        patientID = person_ID;
        DROP distance clusterID clusterFLAG;
RUN;

/*****
/* REMOVE ALL THE CASES FROM THE BASE POPULATION TO PREVENT */
/* TRANSFORMING ONE CASE WITH ANOTHER CASE */
*****/

PROC SQL NOPRINT;
    CREATE TABLE POP_noCASES AS
        SELECT pop.*, Cases.patientID
            FROM MIDAS.popDATA pop LEFT JOIN Cases
                ON pop.person_ID = Cases.patientID;
QUIT;

DATA POP_noCASES;
    SET POP_noCASES;
    IF patientID NE "" THEN DELETE;
    FLAG = 0;
RUN;

SASFILE MIDAS.popDATA CLOSE;
/* CLEAR POPULATION DATA FROM RAM */

%MEND CHOOSE_CASES;

```

## 2. Make the following TWO changes to the “RUNIT” MACRO in Appendix D:

After specifying the number of iterations, specify the desired cluster size by adding the following lines:

```

/* SET DESIRED CLUSTER SIZE (e.g. 10) */
%LET CLUSTER = 10;

```

When calling the CHOOSE\_CASES macro, pass it the cluster size as follows:

```

%CHOOSE_CASES(&SIZE, &CLUSTER);

```

## F. (Partial) Synthetic Population Generation: SAS code

```
/******  
/*                               EPISIM 2010                               */  
/*                               Created by Philip AbdelMalik             */  
/*                               RAMDISK VERSION - RAM DRIVE SET TO "V"     */  
/*                               RUN START: SEPT 19, 2010                 */  
/******  
  
%PUT BEGIN...;  
  
PROC PRINTTO LOG = "V:\RESEARCH\EPISIM\LOGS\YOW_popLOG";  
RUN;  
  
/******  
/*                               SET LIBRARIES                               */  
/******  
%PUT SET LIBRARIES MACRO...;  
  
%MACRO SET_LIBRARIES;  
/* SETTING LIBNAMES FOR WORKING DIRECTORIES IN RAM DISK SPACE (i.e. ALL  
TEMP) */  
  
/* THIS DIRECTORY CONTAINS CLEANED CANADA 2006 CENSUS SIMULATION DATA */  
  
LIBNAME SIM 'V:\Research\EPISIM\CAN_CLEAN';  
RUN;  
  
/* THIS DIRECTORY CONTAINS OTTAWA-SPECIFIC DATA */  
LIBNAME YOW 'V:\Research\EPISIM\YOW';  
RUN;  
  
/* THIS DIRECTORY PROVIDES A "PLAYGROUND" FOR DATA */  
LIBNAME PLAY 'V:\Research\EPISIM\PLAYGROUND';  
RUN;  
  
/* SETTING LIBNAME FOR FINAL PERMANENT FILES */  
LIBNAME FINALS 'C:\Research\EPISIM\FINALS';  
RUN;  
  
%MEND SET_LIBRARIES;  
  
/* REMEMBER TO REMOVE PUT STATEMENTS WHERE APPROPRIATE WHEN WRITING  
TO LOG */
```

```

/*****
/*                               HOUSEKEEPING                               */
/*****
%PUT HOSKEEPING MACRO...;
/*****
/*                               HOUSEKEEPING                               */
/*                               */
/*                               CLEAR TEMP WORK LIBRARY                     */
/*                               CLEAR ALL MACROS                             */
/*                               CLEAR ALL MACRO VARIABLES                     */
/*                               */
/*   TAKEN FROM PAPER 082-2009 BY CHUCK BININGER                             */
/*   RETRIEVED FROM THE SAS WEBSITE                                         */
/*****

```

**%MACRO HOUSEKEEPING;**

```

PROC DATASETS
  LIBRARY = work
  KILL;
QUIT;

```

```

%LET syscc = 0; /* Operating environment condition code */
%LET sysrc = 0; /* Operating system condition code */
%LET syslibrc = 0; /* Libname statement condition code */
%LET sysfilrc = 0; /* Filename statement condition code */
%LET syslckrc = 0; /* SAS Shre lock statement condition code */
%LET syslast = ; /* Contains last created dataset */

```

```

OPTIONS NONOTES; /* Requests whether or not notes are output to the log */
OPTIONS obs = max; /* Resets the number of observations to process */

```

```

/*****
/*                               DEFINE FORMATTING                               */
/*****

```

```

PROC FORMAT;
  VALUE AGEGROUP 1 = "0-4 years"
    2 = "5-9 years"
    3 = "10-14 years"
    4 = "15-19 years"
    5 = "20-24 years"
    6 = "25-29 years"
    7 = "30-34 years"
    8 = "35-39 years"
    9 = "40-44 years"
    10 = "45-49 years"
    11 = "50-54 years"
    12 = "55-59 years"
    13 = "60-64 years"
    14 = "65-69 years"
    15 = "70-74 years"
    16 = "75-79 years"
    17 = "80-84 years"
    18 = "85-89 years"
    19 = "90-94 years"
    20 = "95-99 years"
    21 = "100+ years";

```

```

RUN;

```

**%MEND HOUSEKEEPING;**

```

/*****
/*                               ESTIMATE POPULATION COUNTS                               */

```

```

/*****/
%PUT ESTIMATE POPULATION COUNTS MACROS...;

/*****/
/* BEGIN POPULATION ESTIMATION AND OPTIMISATION */
/*****/

%MACRO ePOP;

/*****/
/* SET GLOBAL VARIABLES */
/*****/

%LET SEED = 0;
/* FOR RANDOM NUMBER GENERATION - ENSURES DIFFERENT WITH EACH RUN */
%LET RANGE = 8;
/* SET TO ERROR RANGE IN COUNTS DUE TO PRIVACY RULES - STATSCAN = 4 EACH
WAY, SO RANGE IS 8 */
%LET MIN = 4;
/* ERROR SIZE = 4 */

%LET OPTIMISE = 1000000;
/* MAXIMUM NUMBER OF RUNS FOR OPTIMISATION ATTEMPT */

/*****/
/* PREPARE THE DATASET */
/*****/

DATA temp_ePopulation;
SET SIM.POPULATION;
WHERE GEOGRAPHY > "35060000" AND GEOGRAPHY < "35070000" AND
population100 NE .;

/* NOTE: 93 records in Canada have no population information in the 2006 Census Profile */
/* NOTE: 6 records in Ottawa have no population and were therefore excluded through this
condition */

S_RECORDNUM = _n_;
CALL SYMPUT ('NUMDAS',_n_);
RUN;

PROC SORT DATA=temp_ePopulation;
BY GEOGRAPHY;
RUN;

/*****/
/* BEGIN WITH THOSE AGED 15 YEARS AND OVER */
/* (A ROUNDED COUNT FOR THIS GROUP IS GIVEN IN THE CENSUS*/
/*****/

DATA ePOP15OVER;

SET temp_ePopulation;
KEEP GEOGRAPHY eM15_19 eM20_24 eM25_29 eM30_34 eM35_39
eM40_44 eM45_49 eM50_54 eM55_59 eM60_64 eM65_69 eM70_74 eM75_79
eM80_84 eM85
eF15_19 eF20_24 eF25_29 eF30_34 eF35_39 eF40_44 eF45_49 eF50_54
eF55_59 eF60_64 eF65_69 eF70_74 eF75_79 eF80_84 eF85
mPop15Over fPop15Over ePop15Over Pop15Over LeaveFlag1 ARUNS;

/*****/
/* SET LOCAL VARIABLES */
/*****/

```



mPop15Over=0; fPop15Over=0; ePop15Over=0; ARUNS = 0;

```

/*****
/*      DEFINE ARRAYS      */
*****/

ARRAY totMALES{*}    M15_19 M20_24 M25_29 M30_34 M35_39 M40_44
M45_49 M50_54 M55_59 M60_64 M65_69 M70_74 M75_79 M80_84 M85;

ARRAY estMALES{*}    eM15_19 eM20_24 eM25_29 eM30_34 eM35_39 eM40_44
eM45_49 eM50_54 eM55_59 eM60_64 eM65_69 eM70_74 eM75_79 eM80_84 eM85;

ARRAY totFEMALES{*}  F15_19 F20_24 F25_29 F30_34 F35_39 F40_44 F45_49
F50_54 F55_59 F60_64 F65_69 F70_74 F75_79 F80_84 F85;

ARRAY estFEMALES{*}  eF15_19 eF20_24 eF25_29 eF30_34 eF35_39 eF40_44
eF45_49 eF50_54 eF55_59 eF60_64 eF65_69 eF70_74 eF75_79 eF80_84 eF85;

/*****
/*      PSEUDO-RANDOM NUMBER GENERATION      */
*****/

DO UNTIL (ARUNS = &OPTIMISE OR LeaveFlag1 = 1);

    mPop15Over=0; fPop15Over=0; ePop15Over=0;

    /* LOOP TO ESTIMATE MALE COUNTS */
    DO m = 1 TO 15;
        IF totMales(m) > 0 THEN estMales(m) = INT(RANUNI(&SEED)* &RANGE +
        (totMales(m)-&MIN));
        ELSE estMales(m) = INT(RANUNI(&SEED)*&MIN);
        mPop15Over = mPop15Over+estMales(m);
    END;

    /* LOOP TO ESTIMATE FEMALE COUNTS */
    DO f = 1 TO 15;
        IF totFemales(f) > 0 THEN estFemales(f) = INT(RANUNI(&SEED)* &RANGE
        +(totFemales(f)-&MIN));
        ELSE estFemales(f) = INT(RANUNI(&SEED)*&MIN);
        fPop15Over = fPop15Over+estFemales(f);
    END;

    /*****
    /*      CALCULATE TOTALS      */
    *****/

    ePop15Over = mPop15Over + fPop15Over;

    /*****
    /*      RUN COUNT VERIFICATION      */
    *****/

    IF ePop15Over < Pop15Over+4 AND ePop15Over > Pop15Over-4 THEN DO;
        LeaveFlag1 = 1;
    END;
    ELSE DO;
        LeaveFlag1 = 0;
    END;

    ARUNS = ARUNS + 1;

END;
```

```

RUN;

PROC FREQ DATA=ePop15Over NOPRINT;
  TABLES LeaveFlag1 / OUT=ePOP1CHECK;
RUN;

DATA TEMP_popCHECK1;
  SET ePOP1CHECK;
  WHERE LeaveFlag1 = 1;
  IF COUNT NE &NUMDAS THEN Check1 = "FAIL";
  ELSE Check1 = "PASS";
RUN;

/*****
/*  MERGE TO THE ORIGINAL POPULATION DATASET TO ALLOW FOR      */
/*  CALCULATION OF DESIRED NUMBER OF THOSE AGED UNDER 15 YEARS */
/*  SINCE WE HAVE THE EXACT 100% POPULATION COUNT IN THE CENSUS */
*****/

PROC SORT DATA=ePOP15Over;
  BY GEOGRAPHY;
RUN;
DATA interimPOP;
  MERGE temp_ePopulation ePOP15Over;
RUN;

/*****
/*  ESTIMATE POPULATION COUNTS FOR THOSE AGED UNDER 15 YEARS  */
*****/
DATA ePOPUNDER15;
  SET interimPOP;
  KEEP GEOGRAPHY Population100 TotalMale TotalFemale
  eM0_4 eM5_9 eM10_14 eF0_4 eF5_9 eF10_14 mPopUnder15 fPopUnder15
  ePopUnder15 Target_ePopUnder15 CRUNS LeaveFlag2;

  /*****
  /*  SET LOCAL VARIABLES  */
  *****/
  mPopUnder15=0; fPopUnder15=0; ePopUnder15=0; CRUNS=0;
  Target_ePopUnder15 = Population100-ePop15Over;

  /*****
  /*  DEFINE ARRAYS  */
  *****/

  ARRAY totMALES{*}  M0_4  M5_9  M10_14 ;
  ARRAY estMALES{*}  eM0_4  eM5_9  eM10_14 ;

  ARRAY totFEMALES{*}  F0_4  F5_9  F10_14 ;
  ARRAY estFEMALES{*}  eF0_4  eF5_9  eF10_14 ;

```

```

/*****
/* PSEUDO-RANDOM NUMBER GENERATION */
*****/

DO UNTIL (CRUNS = &OPTIMISE OR LeaveFlag2 = 1);

    mPopUnder15=0; fPopUnder15=0; ePopUnder15=0;

    /* LOOP TO ESTIMATE MALE COUNTS */
    DO m = 1 TO 3;
        IF totMales(m) > 0 THEN estMales(m) = INT(RANUNI(&SEED)* &RANGE +
            (totMales(m)-&MIN));
        ELSE estMales(m) = INT(RANUNI(&SEED)*&MIN);
        mPopUnder15 = mPopUnder15+estMales(m);
    END;

    /* LOOP TO ESTIMATE FEMALE COUNTS */
    DO f = 1 TO 3;
        IF totFemales(f) > 0 THEN estFemales(f) = INT(RANUNI(&SEED)* &RANGE
            +(totFemales(f)-&MIN));
        ELSE estFemales(f) = INT(RANUNI(&SEED)*&MIN);
        fPopUnder15 = fPopUnder15+estFemales(f);
    END;

    /*
    /* CALCULATE TOTALS */
    *****/

    ePopUnder15 = mPopUnder15 + fPopUnder15;

    /*
    /* RUN COUNT VERIFICATION */
    *****/

    IF ePopUnder15 = Target_ePopUnder15 THEN DO;
        LeaveFlag2 = 1;
    END;
    ELSE DO;
        LeaveFlag2 = 0;
    END;

    CRUNS = CRUNS + 1;

END;

RUN;

PROC FREQ DATA=ePopUnder15 NOPRINT;
    TABLES LeaveFlag2 / OUT=ePOP2CHECK;
RUN;

DATA TEMP_popCHECK2;
    SET ePOP2CHECK;
    WHERE LeaveFlag2 = 1;
    IF COUNT NE &NUMDAS THEN Check2 = "FAIL";
    ELSE Check2 = "PASS";
RUN;

```

```

%MEND ePOP;

/*****
/* LOOP THROUGH UNTIL ALL NUMBERS MATCH, OR 1,000,000 RUNS */
*****/

%MACRO ePOP_RUN;

    %LET OPTIMISE = 1000000;
/* MAXIMUM NUMBER OF RUNS FOR OPTIMISATION ATTEMPT */

    %DO ePOP_r = 1 %TO &OPTIMISE;
        %ePOP;
        DATA _NULL_;
            SET TEMP_popCHECK1;
            CALL SYMPUT ('CHECK1',CHECK1);
        RUN;
        DATA _NULL_;
            SET TEMP_popCHECK2;
            CALL SYMPUT ('CHECK2',CHECK2);
        RUN;

        %IF "&CHECK1" = "PASS" AND "&CHECK2" = "PASS" %THEN %GOTO
EXIT_ePOP_RUN;
        %END;
        %EXIT_ePOP_RUN:

        DATA PLAY.YOW_ePopulation_v2;
            MERGE ePopUnder15 ePOP15Over;
            BY GEOGRAPHY;
        RUN;
%MEND ePOP_RUN;

/*****
/*                               iGENESIS                               */
*****/

%PUT iGENESIS MACROS...;
/*****
/*                               BEGIN GENESIS                               */
*****/

%MACRO iGENESIS;

    /*****
    /*                               SET VARIABLES                               */
    *****/

    %LET SEED = 0;
        /* FOR RANDOM NUMBER GENERATION - ENSURES DIFFERENT WITH EACH
        RUN */

        /* PLEASE DEFINE THE FOLLOWING: */
        %LET tpM85_89 = 0.34;          /* TARGET PROPORTION OF THOSE AGED 85 TO 89
YEARS WHO ARE MALES */
        %LET tpM90_94 = 0.28;          /* TARGET PROPORTION OF THOSE AGED 90 TO 94
YEARS WHO ARE MALES */
        %LET tpM95_99 = 0.22;          /* TARGET PROPORTION OF THOSE AGED 95 TO 99
YEARS WHO ARE MALES */
        %LET tpM100 = 0.18;           /* TARGET PROPORTION OF THOSE AGED 100
YEARS PLUS WHO ARE MALES */
        %LET tmCENTENARIANS = 20;      /* TARGET NUMBER (APPROXIMATE) OF MALE
CENTENARIANS */

```

```

%LET tCENTENARIANS = 91; /* TARGET NUMBER (APPROXIMATE) OF
FEMALE CENTENARIANS */

```

```

/*****
/* PREPARE THE DATASET */
/* IN THE PROCESS, ALSO STORE THE FIRST AND LAST DAs (EXTENT) */
*****/

```

```

DATA iGENESIS_PREP;
  SET PLAY.YOW_ePopulation_v2;
  RECNUM = _n_;
RUN;
PROC SORT DATA=iGENESIS_PREP;
  BY DESCENDING GEOGRAPHY;
RUN;
DATA iGENESIS_PREP;
  SET iGENESIS_PREP;
  CALL SYMPUT ('FIRST_DA',GEOGRAPHY);
  /* SINCE CALL SYMPUT WILL STORE THE LAST RECORD'S VALUE */
RUN;
PROC SORT DATA=iGENESIS_PREP;
  BY GEOGRAPHY;
RUN;
DATA iGENESIS_PREP;
  SET iGENESIS_PREP;
  CALL SYMPUT ('LAST_DA',GEOGRAPHY);
RUN;
DATA SENIOR_CHECK;
  SET iGENESIS_PREP;
  WHERE RECNUM = 1;
  KEEP GEOGRAPHY;
RUN;
DATA SENIOR_CHECK;
  SET SENIOR_CHECK;
  FIRSTDA = &FIRST_DA; LASTDA = &LAST_DA;
  DROP GEOGRAPHY;
RUN;

```

```

/* NOTE: SENIOR CHECK IS USED TO CHECK THE PROPORTIONS OF MALES AND
FEMALES IN THE 85+ AGE GROUPS */

```

```

/* CALCULATE DESIRED PROPORTIONS BASED ON STATISTICS CANADA FOR
AGES 85 PLUS */

```

```

PROC MEANS DATA=PLAY.YOW_ePopulation_v2 NOPRINT;
  VAR eM85;
  OUTPUT OUT = MALES85plus
  SUM = m85plus;
RUN;
PROC MEANS DATA=PLAY.YOW_ePopulation_v2 NOPRINT;
  VAR eF85;
  OUTPUT OUT = FEMALES85plus
  SUM = f85plus;
RUN;
DATA SENIOR_SUMS;
  MERGE MALES85plus FEMALES85plus;
  KEEP m85plus f85plus;
RUN;
DATA targetSENIOR_PROPORTIONS;
  SET SENIOR_SUMS;

  fTOM85_89 = (1-&tpM85_89)/&tpM85_89;
  fTOM90_94 = (1-&tpM90_94)/&tpM90_94;
  fTOM95_99 = (1-&tpM95_99)/&tpM95_99;

```

```

tM85_99 = m85plus - &tmCENTENARIANS;
tF85_99 = f85plus - &tfCENTENARIANS;

/* LET a = NUMBER OF MALES AGED 85 to 89; b = NUMBER OF MALES AGED
90 to 94; c = NUMBER OF MALES AGED 95 to 99 */
/* d = NUMBER OF FEMALES AGED 85 to 89; e = NUMBER OF FEMALES
AGED 90 to 94; f = NUMBER OF FEMALES AGED 95 to 99 */
/* THEREFORE a+b+c = tM85_99; d+e+f = tF85_99 */
/* WITH SOME MANIPULATION, GET THAT a = xc + k where x is a coefficient and k is
a constant */
/* NEED TO CALCULATE x and k */

xCOEFF = (1/(fTOM90_94-fTOM85_89));
kCONST = -1*((tF85_99 - (fTOM90_94*tM85_99))/(fTOM90_94-fTOM85_89));

/* FIND MAXIMUM C BY ASSUMING THAT a>b>c AND b=c+1 */
maxM95_99 = (tM85_99 - kCONST - 1)/(2+xCOEFF);

/* ASSUME IN REALITY HAVE 80% OF THE MAXIMUM NUMBER OF MALES AGED
95 TO 99 */
tM95_99 = ROUND(maxM95_99*0.80);

/* CALCULATE THE REST AND THEIR RESPECTIVE PROPORTIONS */
tM85_89 = ROUND((xCOEFF*tM95_99) + kCONST);
tM90_94 = tM85_99 - tM85_89 - tM95_99;

tF85_89 = ROUND(fTOM85_89 * tM85_89);
tF90_94 = ROUND(fTOM90_94 * tM90_94);
tF95_99 = ROUND(fTOM95_99 * tM95_99);

ptM85_89=tM85_89/m85plus; ptM90_94=tM90_94/m85plus;
ptM95_99=tM95_99/m85plus;
ptF85_89=tF85_89/f85plus; ptF90_94=tF90_94/f85plus; ptF95_99=tF95_99/f85plus;

IF ptM85_89 >= ((m85plus-60)/m85plus) THEN DO;
/* IN CASE RESULTS IN TOO MANY MALES RELATIVE TO FEMALES */

CALL SYMPUT ('MP1',0.9860);
CALL SYMPUT ('MP2',0.9933);
CALL SYMPUT ('MP3',0.9959);
CALL SYMPUT ('FP1',0.9767);
CALL SYMPUT ('FP2',0.9863);
CALL SYMPUT ('FP3',0.9911);

END;
ELSE DO;
CALL SYMPUT ('MP1',ptM85_89);
CALL SYMPUT ('MP2',ptM85_89 + ptM90_94);
CALL SYMPUT ('MP3',ptM85_89 + ptM90_94 + ptM95_99);
CALL SYMPUT ('FP1',ptF85_89);
CALL SYMPUT ('FP2',ptF85_89 + ptF90_94);
CALL SYMPUT ('FP3',ptF85_89 + ptF90_94 + ptF95_99);

END;
RUN;

```

```

/*****
/*      BEGIN      */
*****/

DATA YOW_iGENESIS;
  SET iGENESIS_PREP;
  KEEP GEOGRAPHY POPULATION100 SEX tAGE AGE ADULT UID sRANDOM
  AGEGROUP;

  iSTRATIFIED = 0; /* Counter to allow as many records to be created as age and sex
  stratified population size for each DA */
  UIDCount = 0; /* Counter for appending to DA to create UID for each individual */
  LENGTH UID $14;

  ARRAY iSET (*) eM0_4 eM5_9 eM10_14 eM15_19 eM20_24 eM25_29 eM30_34
  eM35_39 eM40_44 eM45_49 eM50_54 eM55_59 eM60_64 eM65_69 eM70_74
  eM75_79 eM80_84 eM85
  eF0_4 eF5_9 eF10_14 eF15_19 eF20_24 eF25_29 eF30_34 eF35_39 eF40_44
  eF45_49 eF50_54 eF55_59 eF60_64 eF65_69 eF70_74 eF75_79 eF80_84 eF85;

  DO n = 1 TO 36;

    iSTRATIFIED = 0;
    IF iSET(n) > 0 THEN DO;

      DO UNTIL (iSTRATIFIED = iSET(n));

        iSTRATIFIED = iSTRATIFIED + 1;

        /* ASSIGN AGE GROUP */
        IF n <= 18 THEN AGEGROUP = n;
        ELSE AGEGROUP = n-18;

        /* ASSIGN A RANDOM NUMBER TO USE FOR GENERATING AGES 85+ */
        sRANDOM = RANUNI(&SEED);

        /* SET GENDER */
        IF n < 19 THEN SEX = "M";
        ELSE SEX = "F";

        /* ASSIGN AGE */
        IF n < 18 THEN tAGE = (RANUNI(&SEED)*5 + ((5*n)-5));
        /* RANDOM MALE AGE ASSIGNMENT BASED ON AGE RANGE */
        ELSE IF n = 18 THEN DO;
          IF sRANDOM <= &MP1 THEN tAGE = (RANUNI(&SEED)*5 + 85);
          ELSE IF sRANDOM > &MP1 AND sRANDOM <= &MP2 THEN DO;
            tAGE = (RANUNI(&SEED)*5 + 90);
            AGEGROUP = 19;
          END;
          ELSE IF sRANDOM > &MP2 AND sRANDOM <= &MP3 THEN DO;
            tAGE = (RANUNI(&SEED)*5 + 95);
            AGEGROUP = 20;
          END;
          ELSE DO;
            tAGE = (RANUNI(&SEED)*7 + 100);
            /* ASSUMES OLDEST MALE IN OTTAWA CAN BE 107 */
            AGEGROUP = 21;
          END;
        END;
      END;
    END;
  END;

```

```

ELSE IF n > 18 AND n < 36 THEN tAGE = (RANUNI(&SEED)*5 + ((5*(n-18))-5));
/* RANDOM FEMALE AGE ASSIGNMENT BASED ON AGE RANGE */
  ELSE IF n = 36 THEN DO;
    IF sRANDOM <= &FP1 THEN tAGE = (RANUNI(&SEED)*5 + 85);
    ELSE IF sRANDOM > &FP1 AND sRANDOM <= &FP2 THEN DO;
      tAGE = (RANUNI(&SEED)*5 + 90);
      AGEGROUP = 19;
    END;
    ELSE IF sRANDOM > &FP2 AND sRANDOM <= &FP3 THEN DO;
      tAGE = (RANUNI(&SEED)*5 + 95);
      AGEGROUP = 20;
    END;
    ELSE DO;
      tAGE = (RANUNI(&SEED)*12 + 100);
      /* ASSUMES OLDEST FEMALE IN OTTAWA CAN BE 112 */
      AGEGROUP = 21;
    END;
  END;
END;

/* ROUND AGE UNLESS LESS THAN ONE YEAR OLD */
IF tAGE < 0.08 THEN tAGE = 0.08;
/* MINIMUM AGE IS 1 MONTH */
IF tAGE < 1 THEN AGE = ROUND(tAGE,.2);
ELSE AGE = INT(tAGE);

/* SET ADULT FLAG FOR WHETHER OR NOT INDIVIDUAL IS AN ADULT */
IF AGE < 18 THEN ADULT = 0;
ELSE ADULT = 1;

/* UPDATE COUNTER AND ASSIGN INDIVIDUAL UNIQUE IDENTIFIER */
UIDCount = UIDCount + 1;
UIDCounter = COMPRESS(PUT(UIDCount, 5.));
UID = GEOGRAPHY || "." || UIDCounter;

  OUTPUT;
  END;
END;

END;

  FORMAT AGEGROUP AGEGROUP.;
  RUN;

%MEND iGENESIS;

%MACRO NUM_CHECK;

/* OUTPUT NUMBER OF MALES AND FEMALES FOR EACH AGE GROUP */
PROC MEANS NOPRINT DATA=YOW_iGENESIS;
  VAR AGEGROUP;
  CLASS AGEGROUP;
  WHERE SEX = "M";
  OUTPUT OUT = MAGEGROUP_NUMS
  N = MCOUNT;
RUN;
DATA MAGEGROUP_NUMS;
  SET MAGEGROUP_NUMS;
  WHERE AGEGROUP NE .;
  DROP _FREQ_;
RUN;
PROC SORT DATA=MAGEGROUP_NUMS;
  BY AGEGROUP;
RUN;

```



```

PROC MEANS NOPRINT DATA=YOW_iGENESIS;
  VAR AGEGROUP;
  CLASS AGEGROUP;
  WHERE SEX = "F";
  OUTPUT OUT = FAGEGROUP_NUMS
         N = FCOUNT;
RUN;
DATA FAGEGROUP_NUMS;
  SET FAGEGROUP_NUMS;
  WHERE AGEGROUP NE .;
  DROP _FREQ_;
RUN;
PROC SORT DATA=FAGEGROUP_NUMS;
  BY AGEGROUP;
RUN;

/* COMBINE MALES AND FEMALES AND CALCULATE GENDER PROPORTIONS FOR
EACH AGE GROUP */
DATA AGEGROUP_NUMS;
  MERGE MAGEGROUP_NUMS FAGEGROUP_NUMS;
  BY AGEGROUP;
  MPROP = MCOUNT/(MCOUNT+FCOUNT);
  FPROP = 1-MPROP;
RUN;

/* SET ALL PROPORTIONS IN SENIOR_CHECK TO 0 */
DATA SENIOR_CHECK;
  SET SENIOR_CHECK;
  ARRAY mprops (*) pM0_4 pM5_9 pM10_14 pM15_19 pM20_24 pM25_29 pM30_34
pM35_39 pM40_44 pM45_49 pM50_54 pM55_59
                pM60_64 pM54_69 pM70_74 pM75_79 pM80_84 pM85_89 pM90_94
pM95_99 pM100;
  DO p = 1 TO 21;
    mprops(p) = 0;
  END;
RUN;

/* POPULATE SENIOR_CHECK WITH THE MALE PROPORTIONS FOR EACH AGE
GROUP */

%DO sRUN = 1 %TO 21;
  DATA _NULL_;
    SET AGEGROUP_NUMS;
    WHERE AGEGROUP = &sRUN;
    CALL SYMPUT ('PASS_mPROP',MPROP);
  RUN;
  DATA SENIOR_CHECK;
    SET SENIOR_CHECK;
    ARRAY mprops (*) pM0_4 pM5_9 pM10_14 pM15_19 pM20_24 pM25_29
pM30_34 pM35_39 pM40_44 pM45_49 pM50_54 pM55_59
                pM60_64 pM54_69 pM70_74 pM75_79 pM80_84 pM85_89 pM90_94
pM95_99 pM100;

    mprops(&sRUN) = &PASS_mPROP;
  RUN;
%END;

```

```
/* CHECK TO SEE IF PROPORTIONS LIE WITHIN DESIRED RANGES */
```

```
DATA SENIOR_CHECK;  
SET SENIOR_CHECK;
```

```
/* SETTING S_FLAG, WITH PLUS OR MINUS 2% LENIENCY */
```

```
IF pM85_89 < 0.36 AND pM85_89 > 0.32 AND  
pM90_94 < 0.30 AND pM90_94 > 0.26 AND  
pM95_99 < 0.24 AND pM95_99 > 0.20 AND  
pM100 < 0.20 AND pM100 > 0.16 THEN S_FLAG=1;  
ELSE S_FLAG=0;
```

```
RUN;
```

```
%MEND NUM_CHECK;
```

```
/*  
GENESIS  
*/
```

```
%MACRO GENESIS;
```

```
%DO SENIORS_LOOP = 1 %TO 1000;
```

```
/* LOOP TO TRY TO ACHIEVE DESIRED SENIORS POPULATION PROPORTIONS */
```

```
%iGENESIS;  
%NUM_CHECK;
```

```
DATA SENIOR_CHECK;  
SET SENIOR_CHECK;  
CALL SYMPUT ('S_FLAG',S_FLAG);  
RUN;
```

```
%IF &S_FLAG = 1 %THEN %GOTO SENIORS_COMPLETE;  
%END;
```

```
%SENIORS_COMPLETE:
```

```
DATA PLAY.YOW_iGENESIS_v2;  
SET YOW_iGENESIS;  
RUN;
```

```
%MEND GENESIS;
```

```
/*  
RECONCILE FAMILY COUNTS  
*/  
%PUT RECONCILE FAMILY COUNTS...;
```

```
/*  
PREPARE FAMILY DATA  
*/
```

```
%MACRO PREP_FAM_DATA;
```

```
DATA YOW_iGENESIS;  
SET PLAY.YOW_iGENESIS_v2;  
RUN;  
PROC MEANS N DATA=YOW_iGENESIS NOPRINT;  
WHERE ADULT = 1;  
BY GEOGRAPHY;  
VAR ADULT;
```

```

OUTPUT OUT = YOW_aPOP
  N = aPOP;
RUN;
DATA YOW_aPOP;
  SET YOW_aPOP;
  KEEP GEOGRAPHY aPOP;
RUN;

PROC MEANS DATA=YOW_iGENESIS NOPRINT;
  VAR AGE;
  BY GEOGRAPHY;
  WHERE AGE > 17 AND SEX = "M";
  OUTPUT      N = iMADULTS
              OUT = MADULTS;
RUN;
PROC SORT DATA=MADULTS;
  BY GEOGRAPHY;
RUN;

PROC MEANS DATA=YOW_iGENESIS NOPRINT;
  VAR AGE;
  BY GEOGRAPHY;
  WHERE AGE > 17 AND SEX = "F";
  OUTPUT      N = iFADULTS
              OUT = FADULTS;
RUN;
PROC SORT DATA=FADULTS;
  BY GEOGRAPHY;
RUN;

PROC MEANS N DATA=YOW_iGENESIS NOPRINT;
  WHERE AGE < 6;
  BY GEOGRAPHY;
  VAR ADULT;
  OUTPUT OUT = YOW_cPOP1
         N = epopUNDER6;
RUN;
PROC MEANS N DATA=YOW_iGENESIS NOPRINT;
  WHERE AGE > 5 AND AGE < 15;
  BY GEOGRAPHY;
  VAR ADULT;
  OUTPUT OUT = YOW_cPOP2
         N = epop6TO14;
RUN;
PROC MEANS N DATA=YOW_iGENESIS NOPRINT;
  WHERE AGE >14 AND AGE < 18;
  BY GEOGRAPHY;
  VAR ADULT;
  OUTPUT OUT = YOW_cPOP3
         N = epop15TO17;
RUN;
DATA YOW_cPOP;
  MERGE YOW_cPOP1 YOW_cPOP2 YOW_cPOP3;
  BY GEOGRAPHY;
  KEEP GEOGRAPHY epopUNDER6 epop6TO14 epop15TO17 cPOP;
  IF epopUNDER6 = . THEN epopUNDER6 = 0;
  IF epop6TO14 = . THEN epop6TO14 = 0;
  IF epop15TO17 = . THEN epop15TO17 = 0;
  cPOP = epopUNDER6 + epop6TO14 + epop15TO17;
RUN;
PROC SORT DATA=YOW_cPOP;
  BY GEOGRAPHY;
RUN;

```

```

DATA YOW_acPOP;
  MERGE YOW_cPOP YOW_aPOP PLAY.YOW_ePOPULATION_v2 MADULTS FADULTS;
  BY GEOGRAPHY;
  totPOP = cPOP + aPOP;
  KEEP GEOGRAPHY epopUNDER6 epop6TO14 epop15TO17 cPOP aPOP totPOP
population100 iMADULTS iFADULTS;
RUN;

PROC SORT DATA=YOW_acPOP;
  BY GEOGRAPHY;
RUN;

DATA INTERIM_YOW_FAMILIES;
  SET SIM.FAMILIES;
  WHERE Geography > "35060000" AND Geography < "35070000";
RUN;
PROC SORT;
  BY GEOGRAPHY;
RUN;
DATA interimFAMILIES;
  MERGE YOW_acPOP INTERIM_YOW_FAMILIES;
  BY GEOGRAPHY;
  IF totPOP not> 0 THEN DELETE;
  SINGLES = 0; /* VARIABLE TO HOLD THE CALCULATED NUMBER OF
SINGLES */
RUN;
DATA interimFAMILIES;
  SET interimFAMILIES;
  FAM_RECORDNUM = _n_; /* ASSIGN EACH RECORD ITS RECORD NUMBER
FOR FUTURE ITERATIONS */
RUN;

DATA CHILDPOPCHECK;
  SET interimFAMILIES;
  KEEP GEOGRAPHY epopUNDER6 epop6TO14 epop15TO17 HomeKidsUnder6
HomeKids6to14 HomeKids15to17
  diffUNDER6 diff6to14 diff15to17 flagUNDER6 flag6to14 flag15to17 diffcTOTAL;

  IF HomeKidsUnder6 = . THEN HomeKidsUnder6 = 0; /* NOTE: SOME DAs SEEM TO
HAVE CONFLICTING DATA (e.g. 35060253) */
  IF HomeKids6to14 = . THEN HomeKids6to14 = 0;
  IF HomeKids15to17 = . THEN HomeKids15to17 = 0;

  diffUNDER6 = ABS(epopUNDER6-HomeKidsUnder6);
  diff6to14 = ABS(epop6TO14-HomeKids6to14);
  diff15to17 = ABS(epop15TO17-HomeKids15to17);

  IF diffUNDER6 <= 4 THEN flagUNDER6 = "PASS";
  ELSE flagUNDER6 = "FAIL";
  IF diff6to14 <= 4 THEN flag6to14 = "PASS";
  ELSE flag6to14 = "FAIL";
  IF diff15to17 <= 4 THEN flag15to17 = "PASS";
  ELSE flag15to17 = "FAIL";

  diffcTOTAL = diffUNDER6 + diff6to14 + diff15to17;

RUN;
PROC MEANS NOPRINT DATA=CHILDPOPCHECK;
  WHERE flagUNDER6 = "FAIL";
  VAR diffUNDER6;
  OUTPUT OUT = FAIL_UNDER6
  N = cUNDER6_FAILS;
RUN;

```

```

PROC MEANS NOPRINT DATA=CHILDPOPCHECK;
  WHERE flag6to14 = "FAIL";
  VAR diff6to14;
  OUTPUT OUT = FAIL_6to14
         N = c6to14_FAILS;
RUN;
PROC MEANS NOPRINT DATA=CHILDPOPCHECK;
  WHERE flag15to17 = "FAIL";
  VAR diff15to17;
  OUTPUT OUT = FAIL_15to17
         N = c15to17_FAILS;
RUN;
DATA PLAY.childTOTALFAILS;
  MERGE FAIL_UNDER6 FAIL_6to14 FAIL_15to17;
  TOTALFAILS = cUNDER6_FAILS + c6to14_FAILS + c15to17_FAILS;
RUN;

```

```
%MEND PREP_FAM_DATA;
```

```

/*****
/* THIS MACRO RUNS THROUGH AND ASSIGNS ESTIMATED NUMBERS TO EACH */
/* FAMILY TYPE */
/* MULTIPLE ITERATIONS UNTIL THE RESULTING TOTAL NUMBER OF ESTIMATED */
/* FAMILY ADULTS EQUALS THE POPULATION-ESTIMATED NUMBER OF ADULTS */
*****/

```

```
%MACRO ADULT_FAMNUMS;
```

```

/*****
/*          SET GLOBAL VARIABLES          */
*****/

```

```

%LET SEED = 0;          /* FOR RANDOM NUMBER GENERATION - ENSURES
DIFFERENT WITH EACH RUN */
%LET RANGE = 8;        /* SET TO ERROR RANGE IN COUNTS DUE TO PRIVACY
RULES - STATSCAN = 4 EACH WAY, SO RANGE IS 8 */
%LET MIN = 4;          /* ERROR SIZE = 4 */

```

```

DATA _NULL_;
  SET interimFAMILIES;
  CALL SYMPUT ('NUMDAS',_n_);
RUN;

```

```
%DO A_FAMLOOP = 1 %TO &NUMDAS;
```

```

  /* SET STARTING NUMBER OF MISMATCHES FOR OPTIMISATION */
  %LET ATEMPPOPNO MATCH = 999;
  %LET ATEMPPOPNO MATCHC = 999; /* ALLOWS COMPARISON WITH
ATEMPPOPNO MATCH TO EVALUATE ARUNSTOQUIT */
  %LET ARUNSTOQUIT = 100000; /* COUNTER FOR NUMBER OF RUNS TO
ALLOW WITH SAME OR WORSE RESULT BEFORE QUITTING */

```

```
%DO A_FAMOPTIMISER = 1 %TO 100000;
```

```
DATA temp_FAMRECORD;
```

```
SET interimFAMILIES;
```

```
WHERE FAM_RECORDNUM = &A_FAMLOOP;
```

```
CALL SYMPUT ('A_DAUID',GEOGRAPHY);
```

```
CALL SYMPUT ('A_MCKNONE',MARRIEDNOHOMEKIDS);
```

```
ARRAY CensusFamilies (*) MarriedNoHomeKids Married1Kid Married2Kids  
Married3KidsPLUS
```

```
CLNoHomeKids CL1Kid CL2Kids CL3KidsPlus
```

```
FP1Kid FP2Kids FP3KidsPLUS
```

```
MP1Kid MP2Kids MP3KidsPLUS
```

```
HomeKidsUnder6 HomeKids6to14 HomeKids15to17 HomeKids18to24
```

```
HomeKids25PLUS;
```

```
ARRAY cCF (*) MCK0 MCK1 MCK2 MCK3plus
```

```
CLK0 CLK1 CLK2 CLK3plus
```

```
FPK1 FPK2 FPK3plus
```

```
MPK1 MPK2 MPK3plus
```

```
HK6MINUS HK6to14 HK15to17 HK18to24 HK25PLUS;
```

```
DO f = 1 to 19;
```

```
IF &A_FAMOPTIMISER < 10000 THEN DO;
```

```
IF CensusFamilies(f) > 0 THEN cCF(f) = INT(RANUNI(&SEED)* &RANGE +  
(CensusFamilies(f)-&MIN)); /* Attempt convergence as usual with a maximum of 10,000  
attempts */
```

```
ELSE cCF(f) = INT(RANUNI(&SEED)*&MIN);
```

```
END;
```

```
ELSE IF &A_FAMOPTIMISER >= 10000 AND &A_FAMOPTIMISER < 100000
```

```
THEN DO;
```

```
IF CensusFamilies(f) > 0 THEN cCF(f) = INT(RANUNI(&SEED)* &RANGE +  
(CensusFamilies(f)-&MIN)); /* If no convergence after 10,000 runs, then */
```

```
ELSE cCF(f) = 0; /* Set  
records with 0 counts to 0 */
```

```
END;
```

```
ELSE IF &A_FAMOPTIMISER >= 100000 AND &A_FAMOPTIMISER < 250000
```

```
THEN DO;
```

```
IF CensusFamilies(f) > 0 THEN cCF(f) = INT(RANUNI(&SEED)* (&RANGE-4) +  
(CensusFamilies(f)-&MIN)); /* If still no convergence after 100,000 runs, then reduce range by  
4 */
```

```
ELSE cCF(f) = 0; /* And set  
records with 0 counts to 0 */
```

```
END;
```

```
ELSE IF &A_FAMOPTIMISER >= 250000 AND &A_FAMOPTIMISER < 500000
```

```
THEN DO;
```

```
IF CensusFamilies(f) > 0 THEN cCF(f) = INT(RANUNI(&SEED)* (&RANGE-5) +  
(CensusFamilies(f)-&MIN)); /* If still no convergence after 250,000 runs, then reduce range by  
5 */
```

```
ELSE cCF(f) = 0; /* And set  
records with 0 counts to 0 */
```

```
END;
```

```
ELSE IF &A_FAMOPTIMISER >= 500000 AND &A_FAMOPTIMISER < 750000
```

```
THEN DO;
```

```
IF CensusFamilies(f) > 0 THEN cCF(f) = INT(RANUNI(&SEED)* (&RANGE-6) +  
(CensusFamilies(f)-&MIN)); /* If still no convergence after 500,000 runs, then reduce range by  
6*/
```

```
ELSE cCF(f) = 0; /* And set  
records with 0 counts to 0 */
```

```
END;
```

```
ELSE DO;
```

```

        IF CensusFamilies(f) > 0 THEN cCF(f) = INT(RANUNI(&SEED)* (&RANGE-7) +
(CensusFamilies(f)-&MIN)); /* If still no convergence after 750,000 runs, then reduce range by
7 */
        ELSE cCF(f) = 0; /* And set
records with 0 counts to 0 */
        END;

    END;

    min_childHOMEKIDS = (HomeKidsUnder6 + HomeKids6to14 + HomeKids15to17)-12;
    max_childHOMEKIDS = (HomeKidsUnder6 + HomeKids6to14 +
HomeKids15to17)+12;

    childHOMEKIDS = HK6MINUS + HK6to14 + HK15to17;
    adultHOMEKIDS = HK18to24 + HK25PLUS;

    temp_aPOP = ((2*(MCK0 + MCK1 + MCK2 + MCK3plus + CLK0 + CLK1 + CLK2 +
CLK3plus)) + FPK1 + FPK2 + FPK3plus + MPK1 + MPK2 + MPK3plus + adultHOMEKIDS);

        TMC = MCK0 + MCK1 + MCK2 + MCK3plus; /*
Total number of married couples */
        TCL = CLK0 + CLK1 + CLK2 + CLK3plus; /* Total number of
common-law couples */
        TFP = FPK1 + FPK2 + FPK3plus; /* Total number of single
female parent families */
        TMP = MPK1 + MPK2 + MPK3plus; /* Total number of single
male parent families */
        TKH = childHOMEKIDS + adultHOMEKIDS; /* Total number of kids
and adult offspring at home */

        minMADULTS = TMC + TCL + TMP;
        /* NUMBER OF MALE ADULTS BASED ON FAMILY STATUS
(MARRIED, COMMON LAW, OR SINGLE PARENT */
        minFADULTS = TMC + TCL + TMP;
        /* NUMBER OF FEMALE ADULTS BASED ON FAMILY
STATUS (MARRIED, COMMON LAW, OR SINGLE PARENT */

    IF temp_aPOP < aPOP AND iMADULTS >= minMADULTS AND iFADULTS >=
minFADULTS THEN DO;
        SINGLES = aPOP - temp_aPOP;
        AFLAG = 1;
    END;
    ELSE IF temp_aPOP = aPOP AND iMADULTS >= minMADULTS AND iFADULTS >=
minFADULTS THEN DO;
        AFLAG = 1;
    END;
    ELSE DO;
        AFLAG = 0;
    END;

        FAMOPTIMISER = &A_FAMOPTIMISER;

        CALL SYMPUT ('AFLAG',AFLAG);

    new_aPOP = temp_aPOP + SINGLES;
    aPOP_DIFF = abs(new_aPOP-aPOP);

    CALL SYMPUT ('apopdiff',aPOP_DIFF);

/*
CURRENT RUN IS: %PUT CURRENT DA IS: &A_DAUID
&AFLAG &A_FAMOPTIMISER ADULT FLAG IS:
ADULT POP DIFFERENCE IS: &apopdiff;
*/

    RUN;

```

```

%IF &aPOPDIFF < &ATEMPPOPNOMATCH OR &AFLAG = 1 %THEN %DO;
  %LET ATEMPPOPNOMATCH = &aPOPDIFF;
  DATA OPTIMISED_FAM_RECORD;
    SET temp_FAMRECORD;
  RUN;
%END;

%IF &ATEMPPOPNOMATCHC NE &ATEMPPOPNOMATCH %THEN %DO;
  %LET ARUNSTOQUIT = 100000;
  %LET ATEMPPOPNOMATCHC = &ATEMPPOPNOMATCH;
%END;
%ELSE %DO;
  %LET ARUNSTOQUIT = %EVAL(&ARUNSTOQUIT - 1);
%END;

/*
          %PUT CURRENT DA IS:      &A_DAUID      CURRENT
RUN IS:   &A_FAMOPTIMISER      REDUNDANT RUNS TO QUIT:
          &ARUNSTOQUIT      CURRENT OPTIMUM MISMATCH: &ATEMPPOPNOMATCH;
*/

%IF &ARUNSTOQUIT = 0 %THEN %GOTO EXIT_aCHECK; /* IF NO
CHANGE IN MATCHING DIFFERENCE FOR 100,000 RUNS THEN EXIT */
%IF &AFLAG = 1 AND &apopdiff = 0 %THEN %GOTO EXIT_aCHECK; /* IF ADULT
FLAG=1 THEN apopdiff SHOULD BE 0 */
%END;

%EXIT_aCHECK:
DATA interimFAMILIES;
  MERGE interimFAMILIES OPTIMISED_FAM_RECORD;
  BY GEOGRAPHY;
RUN;
%END;

DATA YOW_FAMILIES;
  SET interimFAMILIES;
  /* THIS SIMULATION ASSUMES THAT ALL CHILDREN AGES 0 TO 17 LIVE AT HOME */
  new_cPOP = cPOP;

  IF new_cPOP <= max_childHOMEKIDS AND new_cPOP >= min_childHOMEKIDS
THEN CHILDPASS = 1;
  ELSE CHILDPASS = 0;

  new_totPOP = new_aPOP + new_cPOP;

  KEEP   GEOGRAPHY Population100 cPOP aPOP totPOP SINGLES MCK0 MCK1 MCK2
MCK3plus CLK0 CLK1 CLK2 CLK3plus
        FPK1 FPK2 FPK3plus MPK1 MPK2 MPK3plus HK6MINUS HK6to14 HK15to17
HK18to24 HK25PLUS childHOMEKIDS adultHOMEKIDS
        TMC TCL TFP TMP new_aPOP new_cPOP new_totPOP min_childHOMEKIDS
max_childHOMEKIDS minMADULTS minFADULTS iMADULTS iFADULTS CHILDPASS
AFLAG FAMOPTIMISER;

RUN;

```



```

DATA PLAY.YOW_FAMILIES_v2;
  SET YOW_FAMILIES;
  RUN;

%MEND ADULT_FAMNUMS;

%MACRO HEREGOES;

  %LET tempCHILD_FAILS = 9999999; /* USED TO MINIMISE TOTAL DIFFERENCE OF
CHILD COUNTS */

  %DO child_diffloop = 1 %TO 1000;

    %PUT THIS IS THE OVERALL LOOP RUN NUMBER &child_diffloop;
    %DO popmatchloop = 1 %TO 1000;
      /*%PUT POPULATION MATCHING LOOP NUMBER
&popmatchloop;*/
      %HOUSEKEEPING;
      %ePOP_RUN;
      %HOUSEKEEPING;
      %GENESIS;
      %HOUSEKEEPING;
      %PREP_FAM_DATA;
      %ADULT_FAMNUMS;

      DATA AFLAGCHECK;
        SET PLAY.YOW_FAMILIES_v2;
        KEEP GEOGRAPHY AFLAG;
      RUN;
      PROC SORT DATA = AFLAGCHECK;
        BY DESCENDING AFLAG;
      RUN;
      DATA AFLAGCHECK;
        SET AFLAGCHECK;
        CALL SYMPUT ('LASTAFLAG',AFLAG);
      RUN;
      %IF &LASTAFLAG = 1 %THEN %GOTO EXIT_popmatchloop;
    %END;
  %EXIT_popmatchloop;

DATA PLAY.childTOTALFAILS;
  SET PLAY.childTOTALFAILS;
  CALL SYMPUT ('CHILDFAILS',TOTALFAILS);
  RUN;

%IF &CHILDFAILS < &tempCHILD_FAILS %THEN %DO;
  %LET tempCHILD_FAILS = &CHILDFAILS;
  DATA YOW.YOW_ePOPULATION;
    SET PLAY.YOW_ePOPULATION_v2;
  RUN;
  DATA YOW.YOW_iGENESIS;
    SET PLAY.YOW_iGENESIS_v2;
  RUN;
  DATA YOW.YOW_FAMILIES;
    SET PLAY.YOW_FAMILIES_v2;
  RUN;
  DATA YOW.childTOTALFAILS;
    SET PLAY.childTOTALFAILS;
  RUN;
  /*
  DATA TEMPPRINT;
    SET PLAY.childTOTALFAILS;
    RUN = &child_diffloop;

```

```

                KEEP TOTALFAILS RUN;
            RUN;

            PROC PRINT N NOOBS DATA=TEMPPRINT;
            RUN;
            */
        %END;

        %PUT CURRENT NUMBER OF CHILD FAILS IS &CHILDFAILS;
        %PUT OPTIMISED NUMBER OF CHILD FAILS IS &tempCHILD_FAILS;

        %IF &CHILDFAILS = 0 %THEN %GOTO PARTone_COMPLETE;

        %END;
        %PARTone_COMPLETE:

    %MEND HEREGOES;

    %SET_LIBRARIES;
    %HEREGOES;
    %HOUSEKEEPING;

    /*****
    /*          IMMORTALISE (WRITE TO HARD DISK)          */
    /*****

    DATA FINALS.YOW_ePOPULATION_RAM;
        SET YOW.YOW_ePOPULATION;
    RUN;
    DATA FINALS.YOW_iGENESIS_RAM;
        SET YOW.YOW_iGENESIS;
    RUN;
    DATA FINALS.YOW_FAMILIES_RAM;
        SET YOW.YOW_FAMILIES;
    RUN;
    DATA FINALS.childTOTALFAILS;
        SET YOW.childTOTALFAILS;
    RUN;

    PROC PRINT N NOOBS DATA=FINALS.YOW_FAMILIES_RAM;
        WHERE AFLAG = 0;
    RUN;

```

```

/*****
/*                               EPISIM 2010                               */
/*****
/*                               "ARRANGING MARRIAGES"                       */
/*                               */
/* The following program assigns a female spouse to every male spouse that has been */
/* randomly selected as such from those assigned a relationship status of "couples" */
/* This includes all those who are either married or common law.                */
/*                                     */
/* Previously run code generated this list from the list of simulated males ("YOWMales") */
/* and stored it in a file called "M_Spouses".                                  */
/* Now need to select females from the list of simulated females ("YOWFemales") to */
/* match appropriately to the males in "M_Spouses" based on a series of rules:  */
/*                                     */
/* It is assumed that, for the city of Ottawa:                                  */
/* - Minimum age for spousal selection is 18 years                             */
/* - For 50% of the couples, the male is as old as or older than the female by a */
/*   maximum of 5 years                                                         */
/* - For 24% of the couples, the males is older than the female by at least 6   */
/*   but no more than 12 years                                                  */
/* - For 26% of the couples, the female is older than the male by a maximumm of */
/*   10 years                                                                     */
/* - The likelihood of selection diminishes as the age difference increases     */
/*                                     */
/* The general steps taken are as follows:                                       */
/* - For a given Dissemination Area:                                           */
/*   1. Extract from M_Spouses all males within that DA to create                */
/*     DA_M_Spouses                                                              */
/*   2. Create a list of Eligible_Spouses from YOWFemales for that DA           */
/*   3. For each male in DA_M_Spouses:                                          */
/*     a. Create a list of Eligible_Spouses based on assigned age                */
/*        difference status, assigning the inverse of this as a weight           */
/*     b. Randomly select one female from the list                              */
/*     OPTION: based on weight for probabilistic selection                      */
/*     and assign same family UID to create couple                             */
/*     c. Write out result to a YOWCouples file                                 */
/*   4. Once complete for all males in that DA, check the number of             */
/*     males not assigned a match ("orphaned male spouses")                    */
/*   5. Iterate steps 1 to 4, retaining only optimal solution each time         */
/* - Repeat for next Dissemination Area, until all Dissemination Areas have been */
/*   completed.                                                                  */
/*****
/* REVISIONS: multiple!; 15092010; 16092010; 19092010                        */
/*****

/*****
/*                               SET GLOBAL VARIABLES                       */
/*                               (MODEL PARAMETERS)                          */
/*****

PROC PRINTTO LOG = "V:\RESEARCH\EPISIM\YOW\YOW_spouseLOG";
RUN;

%PUT BEGIN ARRANGED MARRIAGES / COUPLING...;
PROC SQL NOPRINT;
    SELECT name INTO :mymacrovars SEPARATED BY ' '
    FROM dictionary.macros
    WHERE SCOPE = 'GLOBAL';
QUIT;
%SYMDEL &mymacrovars mymacrovars;
PROC DATASETS

```

```

LIBRARY = play
KILL;
QUIT;

%MACRO HOUSEKEEPING;

PROC DATASETS
  LIBRARY = work
  KILL;
QUIT;

%LET syscc = 0;      /* Operating environment condition code */
%LET sysrc = 0;     /* Operating system condition code */
%LET syslibrc = 0;  /* Libname statement condition code */
%LET sysfilrc = 0; /* Filename statement condition code */
%LET syslckrc = 0; /* SAS Shre lock statement condition code */
%LET syslast = ;   /* Contains last created dataset */

OPTIONS NONOTES; /* Requests whether or not notes are output to the log */
OPTIONS obs = max; /* Resets the number of observations to process */

/*****
/*          DEFINE FORMATTING          */
*****/

PROC FORMAT;
  VALUE AGEGROUP 1 = "0-4 years"
                2 = "5-9 years"
                3 = "10-14 years"
                4 = "15-19 years"
                5 = "20-24 years"
                6 = "25-29 years"
                7 = "30-34 years"
                8 = "35-39 years"
                9 = "40-44 years"
                10 = "45-49 years"
                11 = "50-54 years"
                12 = "55-59 years"
                13 = "60-64 years"
                14 = "65-69 years"
                15 = "70-74 years"
                16 = "75-79 years"
                17 = "80-84 years"
                18 = "85-89 years"
                19 = "90-94 years"
                20 = "95-99 years"
                21 = "100+ years";

RUN;

%MEND HOUSEKEEPING;

/*****
/*          PREPARE THE DATASETS          */
/*          ONLY NEED TO RUN THIS MACRO AT THE BEGINNING OF A SESSION          */
*****/

```

**%MACRO** Prep\_FAMDataSets;

/\* ADD A "COUPLES" VARIABLE, RENAME THE SINGLE PARENT VARIABLES AND ADD  
NUMERIC VERSION OF THE GEOGRAPHY CODE \*/

```
DATA SUM_FAM;
  SET YOW.YOW_families;
  /* CALCULATE TOTALS IN EACH "MARITAL STATUS" CATEGORY AS
  GIVEN BELOW */
  COUPLES = TMC + TCL;
  COUPLES = MCK0 + MCK1 + MCK2 + MCK3plus + CLK0 + CLK1 + CLK2 +
  CLK3plus;
  RENAME TFP=FPARENT TMP=MPARENT;
  GEONUM = INPUT(GEOGRAPHY,8.);
  /* NUMERIC VERSION OF THE GEOGRAPHY FIELD (DAUID) */
RUN;
```

/\* ADD SOME FIELDS TO PREP THE iGENESIS DATASET AND BREAK UP INTO  
MALES AND FEMALES \*/

```
DATA iGENESIS;
  SET YOW.YOW_iGENESIS;
  LENGTH FSTATUS $8;
  /* TO HOLD FAMILY STATUS (SINGLE, COUPLE, SINGLE PARENT, CHILD)
  */
  LENGTH UFAMID $20;
  /* TO HOLD A UNIQUE FAMILY IDENTIFIER SO WE CAN MATCH SPOUSES
  AND KIDS */
  COUPLETYPE = .;
  /* TO HOLD COUPLE TYPE (SPOUSAL AGE RELATIONSHIP */
  RUN = 0;
  /* RUN ALLOWS US TO EVENTUALLY KNOW WHAT NUMBER KID ONE IS
  FOR MULTIPLE KID FAMILIES */
  FLAG = 0;
  /* TO ENSURE A RECORD IS ONLY RANDOMLY SELECTED ONCE FOR
  ASSIGNMENT TO A "FAMILY" */

  GEONUM = INPUT(GEOGRAPHY,8.);
  /* NUMERIC VERSION OF THE GEOGRAPHY FIELD (DAUID) */
RUN;
```

```
DATA YOWMales;
  SET iGENESIS;
  WHERE SEX = "M";
RUN;
PROC SORT DATA=YOWMales;
  BY UID;
RUN;
```

```
DATA YOWFemales;
  SET iGENESIS;
  WHERE SEX = "F";
RUN;
PROC SORT DATA=YOWFemales;
  BY UID;
RUN;
```

**%MEND** Prep\_FAMDataSets;

```

/*****
/* MACRO TO ASSIGN MALES TO A FAMILY STRUCTURE TYPE (COUPLE, SINGLE */
/* MALE PARENT...) */
/*****

```

**%MACRO SetMaleSPOUSES;**

```

%LET STRUCT = COUPLES;
/*%PUT CURRENT FAMILY STRUCTURE IS: &STRUCT;*/

DATA SAMPLESIZE;
    SET SUM_FAM;
    WHERE &STRUCT NE 0;
    _NSIZE_ = &STRUCT;
    KEEP GEOGRAPHY &STRUCT _NSIZE_;
    /* _NSIZE_ IS A SAS-RESERVED FIELD TO BE USED LATER IN
    THE RANDOM SAMPLING PROC SURVEYSELECT
    */
RUN;

DATA M_SELECTION;
    SET YOWMales;
    WHERE AGE > 17 AND FLAG = 0;
RUN;

PROC SURVEYSELECT DATA=M_SELECTION NOPRINT
    SAMPSIZE = SAMPLESIZE
    METHOD = SRS
    SEED = 0
    OUT=M_SAMPLEDDATA;
    STRATA GEOGRAPHY;
RUN;

DATA M_SAMPLEDDATA;
    SET M_SAMPLEDDATA;
    DROP SelectionProb SamplingWeight;
    FSTATUS = "COUPLE";
    UFAMID = COMPRESS(GEOGRAPHY||".FAM."||_n_);
    /* UNIQUE ID FOR COUPLE FAMILIES */
    RUN = 1;
    /* SINCE NONE ARE CHILDREN, ALL ASSIGNED RUN AS NUMBER
    1 */
    FLAG = 1;
    /* TO PREVENT RE-SELECTION */
RUN;

DATA YOWMales;
    MERGE YOWMales M_SAMPLEDDATA;
    BY UID;
RUN;

```

**%MEND SetMaleSPOUSES;**

```

/*****/
/* PREP MACRO FOR ASSIGNING AGE RELATIONSHIP BETWEEN SPOUSES (FOR */
/* COUPLES ONLY) */
/*
/* ASSUMPTIONS: */
/* - 50% of couples where the male is up to 5 years older than the female */
/* - 24% of couples where the male is more than 5 years older than the female */
/* - 26% of couples where the female is older than the male */
/*
/* SOURCES: */
/* - BBC NEWS http://news.bbc.co.uk/go/pr/fr/-/2/hi/uk\_news/3312377.stm */
/* - Statistics Canada: Till death do us part? */
/* The risk of first and second marriage dissolution by Warren Clark and */
/* Susan Crompton. Catalogue No/11-008 Summer 2006 */
/*****/

```

**%MACRO PREP\_MATCHMAKING;**

```

DATA M_SPOUSES;
  SET YOWMales;
  WHERE FSTATUS = "COUPLE";
  M_UID = UID;
  M_AGE = AGE;
  KEEP GEOGRAPHY GEONUM M_UID M_AGE FSTATUS UFAMID
  COUPLETYP;
RUN;

DATA YOWCOUPLES;
  SET M_SPOUSES;
RUN;
PROC SORT DATA=YOWCOUPLES;
  BY UFAMID;
RUN;

DATA YOWFemales;
  SET YOWFemales;
  RECORDNUM = _n_;
RUN;

PROC FREQ DATA=M_SPOUSES NOPRINT;
  TABLES GEOGRAPHY*GEONUM / OUT=DA_M_SPOUSE_COUNTS;
RUN;
DATA DA_M_SPOUSE_COUNTS;
  SET DA_M_SPOUSE_COUNTS;
RUN;

```

**%MEND PREP\_MATCHMAKING;**

```

/*****/
/* MATCHING MACRO */
/*****/

```

**%MACRO MatchMaker;**

```

%LET OPTIMUMNUM = 1000;
/* SET THE NUMBER OF ITERATIONS TO ACHIEVE MAXIMUM MATCHING */

DATA SELECTED_DA_M_SPOUSECOUNTS;
  SET DA_M_SPOUSE_COUNTS;
  S_RECORDNUM = _n_;
  CALL SYMPUT ('NUMDAS',_n_);
RUN;

```

```

%DO DAITERATION = 1 %TO &NUMDAs;

/* SET STARTING NUMBER OF MISMATCHES FOR OPTIMISATION */
%LET TEMPNOMATCH = 999;

/* SET CURRENT DA OF FOCUS AND NUMBER OF MALES TO SELECT */
DATA _NULL_;
    SET SELECTED_DA_M_SPOUSECOUNTS;
    WHERE S_RECORDNUM = &DAITERATION;
    CALL SYMPUT ('DAUID',GEOGRAPHY);
    CALL SYMPUT ('MCOUNT',COUNT);
RUN;

/* %PUT CURRENT DA IS &DAUID WITH &MCOUNT MALES; */

/* SELECT MCOUNT NUMBER OF MALES IN CURRENT DA */
DATA SELECTED_M_SPOUSES;
    SET M_SPOUSES;
    WHERE GEOGRAPHY = "&DAUID";
    RECORDNUM = _n_;
RUN;

/* SELECT ALL ELIGIBLE FEMEALS (i.e. AGE 18 YEARS AND OVER) IN
CURRENT DA */
DATA SELECTED_DA_FEMALES;
    SET YOWFemales;
    WHERE GEOGRAPHY = "&DAUID" AND AGE >= 18;
    RECORDNUM = _n_;
RUN;

/* BEGIN SELECTION AND OPTIMISATION ROUTINE */
%DO OPTIMISING = 1 %TO &OPTIMUMNUM;

    /* RESET AGE RELATIONSHIP ASSIGNED TO EACH MALE IN AN
    ATTEMPT TO OPTIMISE MATCHING */
    DATA SELECTED_M_SPOUSES;
        SET SELECTED_M_SPOUSES;
        COUPLETYPE = .;
    RUN;

    %DO COUPLETYPELOOP = 1 %TO 3;

        DATA TEMP_M_SPOUSES;
            SET SELECTED_M_SPOUSES;
            WHERE COUPLETYPE = .;
            IF &COUPLETYPELOOP = 1 THEN DO;
                CALL SYMPUT ('CTYPERATIO',0.5);
                IF M_AGE < 19 THEN DELETE;
                /* ASSUMES NO COUPLED MALES
                AGED 18; ALLOWS COUPLED
                FEMALE TO BE 18 THOUGH */
            END;

            ELSE IF &COUPLETYPELOOP = 2 THEN DO;
                CALL SYMPUT ('CTYPERATIO',0.48);
                IF M_AGE < 24 THEN DELETE;
                /* ALLOWS MINIMUM AGE
                DIFFERENCE TO BE 6 YEARS,
                OLDER MALE */
            END;
            ELSE CALL SYMPUT ('CTYPERATIO',1);
        RUN;
    END;
END;

```



```

PROC SURVEYSELECT DATA=TEMP_M_SPOUSES
  NOPRINT
  RATE=&CTYPERATIO
  METHOD = SRS
  OUT = SELECTED_CTYPE;
RUN;

DATA SELECTED_CTYPE;
  SET SELECTED_CTYPE;
  COUPLETYPE = &COUPLETYPELOOP;
RUN;

PROC SORT DATA=SELECTED_M_SPOUSES;
  BY M_UID;
RUN;
PROC SORT DATA=SELECTED_CTYPE;
  BY M_UID;
RUN;
DATA SELECTED_M_SPOUSES;
  MERGE SELECTED_M_SPOUSES
        SELECTED_CTYPE;
  BY M_UID;
RUN;

%END;

/* SELECTING AN APPROPRIATE SPOUSE FOR EACH MALE */
DATA SELECTED_DA_FEMALES;
  SET SELECTED_DA_FEMALES;
  FLAG = 0;
RUN;

%DO SPOUSEITERATION = 1 %TO &MCOUNT;

  DATA _NULL_;
    SET SELECTED_M_SPOUSES;
    WHERE RECORDNUM = &SPOUSEITERATION;
    CALL SYMPUT ('PASS_UFAMID',UFAMID);
    CALL SYMPUT
      ('PASS_COUPLETYPE',COUPLETYPE);
    CALL SYMPUT ('PASS_MAGE',M_AGE);
    CALL SYMPUT ('PASS_MUID',M_UID);
  RUN;

  DATA ELIGIBLE_SPOUSES;
    SET SELECTED_DA_FEMALES;
    WHERE FLAG = 0;
    IF &PASS_COUPLETYPE = 1 THEN DO;
      IF AGE > &PASS_MAGE OR AGE
        < %EVAL(&PASS_MAGE-5) THEN DELETE;
    END;
    ELSE IF &PASS_COUPLETYPE = 2 THEN DO;
      IF AGE >= %EVAL(&PASS_MAGE-5) OR
        AGE < %EVAL(&PASS_MAGE-12) THEN
        DELETE;
    END;
    ELSE IF &PASS_COUPLETYPE = 3 THEN DO;
      IF AGE <= &PASS_MAGE OR AGE
        > %EVAL(&PASS_MAGE+10) THEN
        DELETE;
    END;
  RUN;

```

```

DATA ELIGIBLE_SPOUSES;
  LENGTH UFAMID $24;
  SET ELIGIBLE_SPOUSES NOBS=N;
  AGEDIFF = ABS(&PASS_MAGE-AGE);
  WEIGHTING = (1/(AGEDIFF+1));
  UFAMID = "&PASS_UFAMID";
RUN;

/* THIS BIT PREVENTS ATTEMPT TO SELECT FROM AN
EMPTY SET, IF NO ELIGIBLE SPOUSES FOUND */
DATA DUMMYSPOUSE;
  SET YOWFemales;
  WHERE RECORDNUM = 1;
  UID="DUMMY"; SEX="DUMMY"; tAGE=.; AGE=.;
  ADULT=.; UFAMID="DUMMY"; RUN=1; FLAG=1;
  GEONUM=.;
RUN;
DATA NOMATCHSPOUSE;
  SET ELIGIBLE_SPOUSES DUMMYSPOUSE;
  DROP RECORDNUM;
RUN;
DATA _NULL_;
  SET NOMATCHSPOUSE NOBS=X;
  CALL SYMPUT('NOMATCHSPOUSE',X);
RUN;

/* IF ELIGIBLE SPOUSE SET IS NOT EMPTY, THEN THE
FOLLOWING BIT WILL RUN */

%IF &NOMATCHSPOUSE > 1 %THEN %DO;
  PROC SURVEYSELECT
  DATA=ELIGIBLE_SPOUSES NOPRINT
    n=1
    METHOD = PPS
    OUT = SELECTED_F_SPOUSE;
    SIZE WEIGHTING;
  RUN;

  DATA SELECTED_F_SPOUSE;
    SET SELECTED_F_SPOUSE;
    FLAG = 1;
    RUN = 1;
    KEEP GEOGRAPHY UID UFAMID RUN
    FLAG;
  RUN;

  DATA SELECTED_DA_FEMALES;
    MERGE SELECTED_DA_FEMALES
    SELECTED_F_SPOUSE;
    BY UID;
  RUN;
%END;
%END;

```

```

DATA SELECTED_F_SPOUSES;
  SET SELECTED_DA_FEMALES;
  WHERE FLAG = 1;
  F_UID = UID;
  F_AGE = AGE;
  KEEP F_UID F_AGE UFAMID;
RUN;

PROC SORT DATA=SELECTED_M_SPOUSES;
  BY UFAMID;
RUN;
PROC SORT DATA=SELECTED_F_SPOUSES;
  BY UFAMID;
RUN;

DATA COUPLESET;
  MERGE SELECTED_M_SPOUSES
  SELECTED_F_SPOUSES;
  BY UFAMID;
  RECORDNUM = _n_;
RUN;

DATA DUMMYDATA;
  SET COUPLESET;
  WHERE RECORDNUM = 1;
  GEOGRAPHY = "DUMDUM"; FSTATUS="DUMMY";
  UFAMID=""; COUPLETYP=.; GEONUM=.; M_UID="";
  M_AGE=.; F_UID=""; F_AGE=.;
RUN;
DATA NOMATCHSET;
  SET COUPLESET DUMMYDATA;
  WHERE F_AGE = .;
  DROP RECORDNUM;
RUN;
DATA _NULL_;
  SET NOMATCHSET NOBS=X;
  CALL SYMPUT('NOMATCH',X);
RUN;

/* OPTIMISATION CHECK AND REPLACE IF APPROPRIATE */

/* %PUT NUMBER OF FORMER OPTIMISED MISMATCHES
WAS %EVAL(&TEMPNOMATCH-1);*/
/* %PUT NUMBER OF CURRENT MISMATCHES
IS %EVAL(&NOMATCH-1);*/
%IF &NOMATCH < &TEMPNOMATCH %THEN %DO;
  %LET TEMPNOMATCH = &NOMATCH;
  DATA OPTIMISED_COUPLES;
    SET COUPLESET;
    DROP RECORDNUM;
  RUN;
%END;

/* %PUT NUMBER OF CURRENT OPTIMISED MISMATCHES
IS %EVAL(&TEMPNOMATCH-1); */
%IF &TEMPNOMATCH = 1 %THEN %GOTO EXIT;
/* %PUT THIS LOOP HAS BEEN RUN &OPTIMISING TIMES;*/
%END;
%EXIT:

```

```

DATA YOWCOUPLES;
    MERGE YOWCOUPLES OPTIMISED_COUPLES;
    BY UFAMID;
RUN;

PROC PRINT N NOOBS;
    WHERE F_AGE NE . AND COUPLETYPE = 1 AND (F_AGE > M_AGE OR
F_AGE < (M_AGE-5));
RUN;
PROC PRINT N NOOBS;
    WHERE F_AGE NE . AND COUPLETYPE = 2 AND (F_AGE > M_AGE OR
F_AGE > (M_AGE-5) OR F_AGE < (M_AGE-12));
RUN;
PROC PRINT N NOOBS;
    WHERE F_AGE NE . AND COUPLETYPE = 3 AND F_AGE <= M_AGE;
RUN;
/*
PROC FREQ DATA=YOWCOUPLES;
    TABLES GEOGRAPHY*COUPLETYPE;
    WHERE F_AGE NE . AND GEOGRAPHY = "&DAUID";
RUN;
*/
PROC FREQ NOPRINT DATA=YOWCOUPLES;
    TABLES UFAMID / OUT = DUPLICATECHECK;
    WHERE F_AGE NE .;
RUN;
PROC PRINT N NOOBS DATA = DUPLICATECHECK;
    WHERE COUNT > 1;
    VAR UFAMID COUNT;
RUN;

DATA YOW.FINALYOWCOUPLES;
    SET YOWCOUPLES;
RUN;

%END;

%MEND MatchMaker;

%PUT BEGIN...;

%HOUSEKEEPING;
%Prep_FAMDataSets;
%SetMaleSPOUSES;
%PREP_MATCHMAKING;
%MatchMaker;

DATA FINALS.FINALYOWCOUPLES;
    SET YOW.FINALYOWCOUPLES;
RUN;

%PUT DONE :);

```

```

%MACRO TESTS;
OPTIONS NOTES;
PROC MEANS DATA=iGENESIS NOPRINT;
  VAR AGE;
  BY GEOGRAPHY;
  WHERE AGE > 17 AND SEX = "M";
  OUTPUT      N = aMALES
             OUT = aMALES;

RUN;
DATA aMALES;
  MERGE aMALES SUM_FAM;
  BY GEOGRAPHY;
  EaMALES = COUPLES + MPARENT;
  KEEP GEOGRAPHY aMALES SINGLES COUPLES MPARENT EaMALES;

RUN;
PROC PRINT N NOOBS;
  WHERE aMALES <= EaMALES;

RUN;

```

```

PROC PRINT N NOOBS DATA=YOWCOUPLES;
  WHERE F_AGE NE . AND COUPLETYPE = 1 AND (F_AGE > M_AGE OR F_AGE <
(M_AGE-5));
RUN;
PROC PRINT N NOOBS DATA=YOWCOUPLES;
  WHERE F_AGE NE . AND COUPLETYPE = 2 AND (F_AGE > M_AGE OR F_AGE >
(M_AGE-5) OR F_AGE < (M_AGE-12));
RUN;
PROC PRINT N NOOBS DATA=YOWCOUPLES;
  WHERE F_AGE NE . AND COUPLETYPE = 3 AND F_AGE <= M_AGE;
RUN;
PROC FREQ DATA=YOWCOUPLES;
  TABLES GEOGRAPHY*COUPLETYPE;
  WHERE F_AGE NE .;

RUN;
PROC FREQ NOPRINT DATA=YOWCOUPLES;
  TABLES UFAMID / OUT = DUPLICATECHECK;
  WHERE F_AGE NE .;

RUN;
PROC PRINT N NOOBS DATA=DUPLICATECHECK;
  WHERE COUNT > 1;
  VAR UFAMID COUNT;

RUN;

```

```

%MEND TESTS;

```

*/\* Statistics Canada's Website at <http://www40.statcan.ca/l01/cst01/agrc41a-eng.htm> gives proportion of total census families that are made up of 5 persons (11.4%), 6 persons (3.7%), and 7 persons and over (1.7%) \*/*

## G. Publications & Reports - Attached as published

# Multidimensional Point Transform for Public Health Practice

P. AbdelMalik; M. N. Kamel Boulos

Faculty of Health and Education, University of Plymouth, Plymouth, UK

## Keywords

Public health, geographic location, privacy, spatial transform, anonymisation

## Summary

**Background:** With increases in spatial information and enabling technologies, location-privacy concerns have been on the rise. A commonly proposed solution in public health involves random perturbation, however consideration for individual dimensions (attributes) has been weak.

**Objectives:** The current study proposes a multidimensional point transform (MPT) that integrates the spatial dimension with other dimensions of interest to comprehensively anonymise data.

**Methods:** The MPT relies on the availability of a base population, a subset patient dataset, and shared dimensions of interest. Perturbation distance and anonymity thresholds are defined, as are allowable dimensional perturbations. A preliminary implementation is presented using sex, age and location as the

three dimensions of interest, with a maximum perturbation distance of 1 kilometre and an anonymity threshold of 20%. A synthesised New York county population is used for testing with 1000 iterations for each of 25, 50, 100, 200 and 400 patient dataset sizes.

**Results:** The MPT consistently yielded a mean perturbation distance of 46 metres with no sex or age perturbation required. Displacement of the spatial mean decreased with patient dataset size and averaged 5.6 metres overall.

**Conclusions:** The MPT presents a flexible, customisable and adaptive algorithm for perturbing datasets for public health, allowing tweaking and optimisation of the trade-offs for different datasets and purposes. It is not, however, a substitute for secure and ethical conduct, and a public health framework for the appropriate disclosure, use and dissemination of data containing personal identifiable information is required. The MPT presents an important component of such a framework.

As public health methods advance with ever-evolving technology to better capture the entire context within which health events occur, requirements for privacy-protective methods also increase. In an attempt to address the privacy issue, algorithms for anonymisation and privacy enhancing techniques (PETs) have been proposed and implemented [13], and calls for public health professionals to challenge policies and lobby legislators have been made [2]. Anonymisation algorithms are often measured as a function of indistinguishable records and re-identification probability. The term *k-anonymisation* refers to the concept where every record becomes indistinguishable from  $k - 1$  other records [14, 15]. While no standards for acceptable anonymity thresholds have been established for public health,  $k$  values of 5 and 20 (representing re-identification probabilities of 20% and 5% respectively) have been suggested and used in the literature [12, 16].

An area that has seen a dramatic increase in concern is location privacy, particularly given the importance of spatial information in public health [17, 18]. With the ubiquitous use of Global Positioning Systems (GPS), online mapping applications that provide high-resolution aerial images, and the increasing use of spatial intelligence in public health, location privacy is becoming increasingly contentious – perhaps more so than with other information technologies [9, 18]. Over a decade ago, Armstrong et al published a paper on various mathematical transformations to mask original point location [19]. Of the methods described, random perturbation was found to perform best overall as measured by retention of pair-wise relations, event-geography relations, clusters, trends and anisotropies [19]. Other public health studies have continued to build on this type of spatial transform [20–23] and a good overview can be found in [18].

## Correspondence to:

Philip AbdelMalik  
Faculty of Health and Education  
University of Plymouth  
Drake Circus  
Plymouth, Devon, PL4 8AA  
UK  
E-mail: philip.abdelmalik@plymouth.ac.uk

Methods Inf Med 2011; 50: ■–■

doi: 10.3414/ME11-01-0001

received: January 11, 2011

accepted: May 10, 2011

prepublished: June 21, 2010

## 1. Background

Privacy, as related to identifiable health information, has been a subject of contention within public health and health research: the literature is littered with comments and complaints [1–3], surveys have sought to assess the perceptions of public health professionals and the general public [4–8], and both privacy advocates and public health professionals appeal to a vaguely painted patch-

work of legislation [1, 9]. The issue does not generally lie with direct identifiers such as name or an identifying number, but rather with attributes that can be used in combination to re-identify individuals. These are referred to as *key attributes* [10] or *quasi-identifiers* [11]. For example, age and sex are commonly used public health quasi-identifiers that have been characterised as having “high utility to an intruder” attempting to re-identify individuals from a dataset [12].

In a classical random perturbation, a circle of radius  $r$  is drawn around the point to be masked such that sufficient population is captured to render the point anonymous, and the point is randomly displaced within the area. This is repeated for each point, resulting in a series of points that are difficult to trace to their original locations due to the stochastic nature of the transform. Not all random perturbations are created equal, however, and advances in their development and implementation have been slow. Ideally, the displacement as measured by the perturbation distance should be minimised, and generally, the more densely populated the area in which a point (case) falls, the less it has to be spatially perturbed to meet a desired anonymity threshold. Adjustments to random perturbation therefore create dynamic radii dependent for each point on its underlying population [21]. This “context-sensitive approach” can be further improved by stratifying on other attributes, such as age and sex, to give a more accurate displacement that minimises information loss [21]. More elaborate revisions of random perturbation have been developed in recent years, including the use of Linear Programming (LP) [22] and a “donut” method of geomasking [23]. However, all of the proposed versions of these transforms modify location-based information almost as an afterthought or secondary anonymisation technique, either assuming that all other identifying information – including important quasi-identifiers such as age and sex – has already been anonymised or stripped, or adjusting the transform accordingly using generalised weighting schemes. Instead, what is needed is a transform that operates discreetly on multiple attributes, in concert with location as part of the overall anonymisation algorithm.

## 2. Objectives

The current methodology refines the random perturbation approach by combining new and previously studied methods to propose a flexible, dynamic and customisable multidimensional point transform (MPT) acting on attribute data. In this context, attributes of interest – such as lo-

cation, age, sex, education, etc. – are referred to as dimensions since they define the scope of the transform. Like previous context-sensitive studies [20, 21, 23], the approach presented is an adaptive geomask. However, unlike others, it allows these other dimensions to be incorporated into the anonymisation algorithm directly based on custodian and user tolerances and requirements.

## 3. Methods

### 3.1 Algorithm: Overview

The proposed algorithm is dependent on the availability of a base population (real or synthesised) matrix,  $A$ , of  $N$  individual records with  $Q$  attributes. The dimensions of interest must be elements of the attribute set, and given the spatial nature of the transform, must include a location attribute – ideally the geographic coordinates of the individual’s relevant address. Given a list of patients,  $B$ , from this base population  $A$ , the goal of the algorithm is to randomly “move” each patient in  $B$  within a maximum perturbation distance  $\Delta$ , while controlling on all dimensions of interest for a defined anonymity threshold,  $k$ . “Move” in this case means selecting an alternate record from  $A$  to represent the patient; in this way, the *locations* are realistic and non-random, but the *selection* is random.

Consider the example where the dimensions of interest are location, age and sex (other dimensions can be added, provided they are elements of both datasets). The algorithm ensures that the anonymity threshold  $k$  is maintained based on these dimensions and sequentially perturbs them as required based on pre-defined conditions and perturbation tolerances. Location perturbation is measured as the distance moved from the original point, and its maximum tolerance is defined by  $\Delta$ ; the age perturbation tolerance allows the dimension to be categorised, for example in 1-year increments, up to a maximum number of categories; and the sex perturbation tolerance is binary, either requiring a perfect match on gender or not.

The acceptable anonymity threshold is defined by  $k$ . For example,  $k = 5$  means that

a given patient is indistinguishable from at least  $k - 1 = 4$  other individuals within the selection area, which in turn translates to a 20% chance of correctly identifying the patient. The maximum perturbation distance  $\Delta$  is the maximum acceptable threshold for spatial displacement. This does not mean that all eligible records for displacement will be up to  $\Delta$  away from the original point, only that this is as far as the algorithm is allowed to go to achieve the desired  $k$ . The actual maximum perturbation radius,  $R$ , will depend on the data and defined  $k$ .

Given the patient dataset  $B$  with  $j = 1$  to  $n$  patients, all patients in  $B$  are removed from  $A$  to give the complement non-patient base population,  $C$ . Removing the patient dataset individuals from the base population records at the onset of the algorithm has two key effects: it prevents selection of one patient in place of another, and it reduces re-identification risk by forcing  $k - 1$  to consist entirely of non-patients. Next, for each record in  $B$ , all records in  $C$  matching  $B_j$  on sex and age are isolated and the distance between each one and  $B_j$  is calculated. If fewer than  $k - 1$  matching records are found within  $\Delta$ , then the sex and age dimensions are perturbed (i.e. grouped or categorised), based on the pre-defined conditions and in parallel in both the case dataset and the population dataset, and the matching is redone. This is repeated until at least  $k - 1$  matching records are found. If the algorithm is unable to reach the desired  $k$ -anonymity, then the record is non-transformable within the current requirements, is flagged as such, and the algorithm proceeds to the next record. Otherwise, the algorithm continues.

Of the matching records, the closest  $k - 1$  records are identified, and a small random distance,  $\delta_r$ , is added to the farthest  $k - 1$  match distance,  $\delta_{\max}$ , defining the perturbation radius  $R$ . The addition of this random distance ensures inclusion of the point used to set the farthest match distance in case any rounding occurs and guarantees a minimum  $k$ -anonymity. It also adds a small stochastic aspect that complicates re-identification of the original patient location, as not only is the selection of the transformed point different with each run, but so also is the underlying pool from which the point is selected. A record within  $R$  of  $B_j$  is then randomly se-



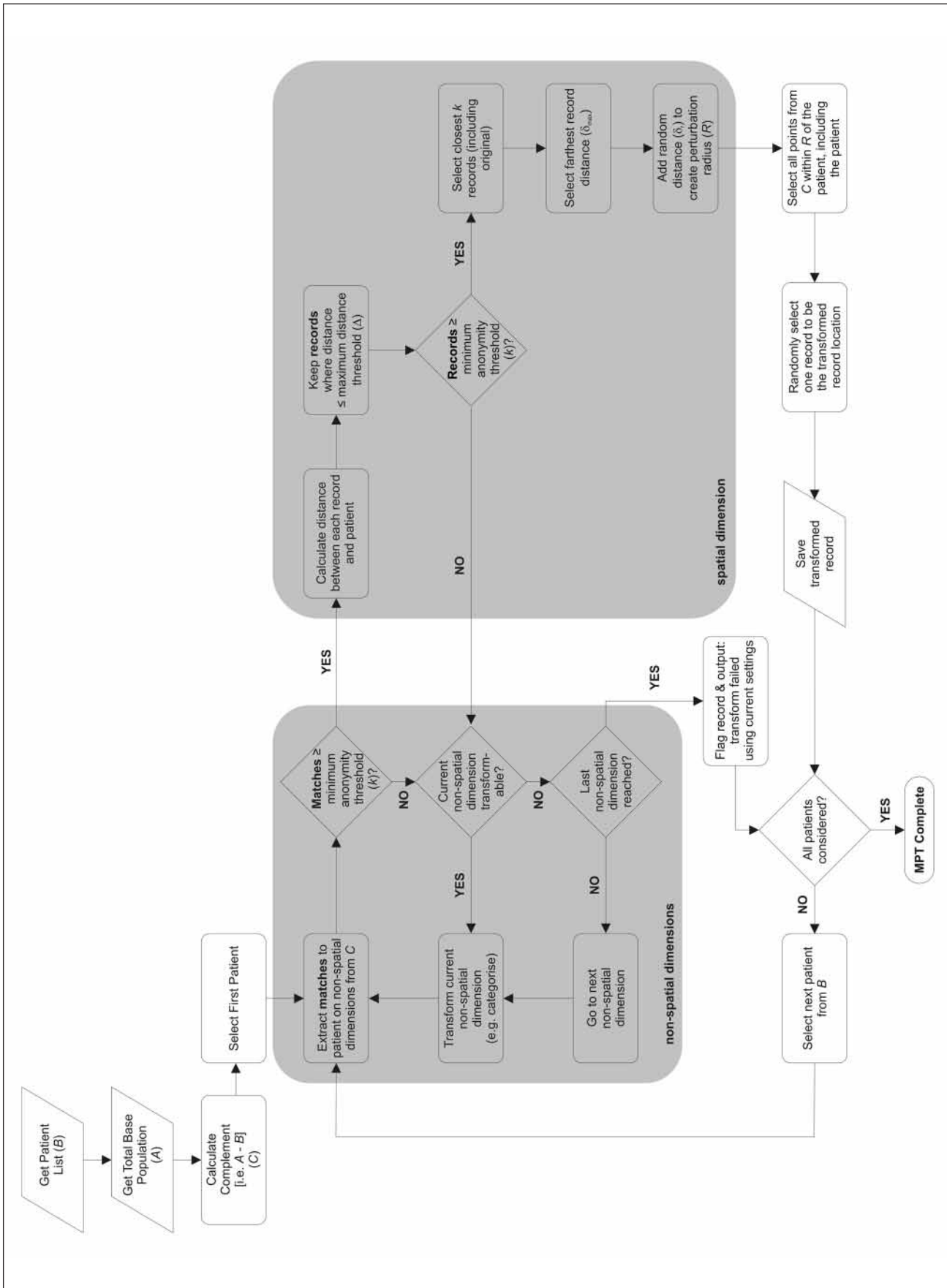


Fig. 1 Multidimensional Point Transform (MPT) flow

lected from *C*, and its location assigned as the perturbed point. This is repeated for the next record until all patient records in *B* have been transformed or flagged. The algorithm flow and components are illustrated in ► Figure 1.

### 3.2 Data

Synthesised population data for New York County were acquired from the MIDAS project [24] by request. The dataset con-

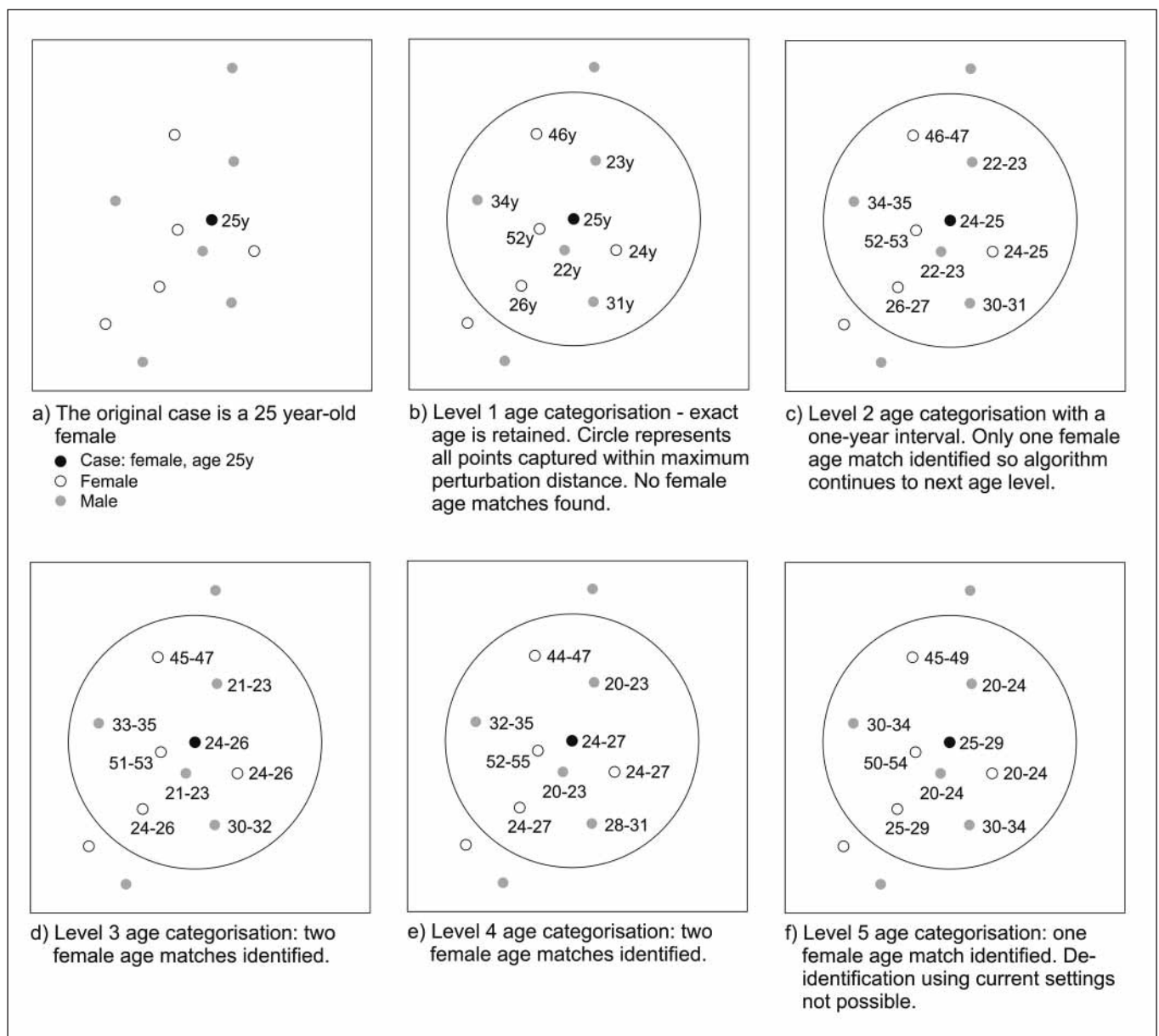
tained synthesised records at the individual level, with the dimensions of interest being age, sex and residential location (latitude and longitude in decimal degrees). For each record, latitude and longitude were converted from decimal degrees to radians prior to algorithm execution for use in extent and distance calculations.

New York county was specifically chosen as the study area to allow for comparison with existing published methods – namely the results of the LP approach [22] – on distances required to achieve specified

*k*-values when additional dimensions are taken into consideration. The two approaches are also similar in that they both seek to minimise perturbation distance and both rely on the presence of underlying spatially-referenced population data.

### 3.3 Algorithm: Preliminary Proof-of-Concept Implementation

Preliminary testing of the algorithm was completed using one thousand iterations



**Fig. 2** Simplified example of age categorisation using one-year intervals with 5 levels and *k* = 5. Note that categorisation always starts at age 0 years (i.e. birth, consistent with census age strata)

for each of 25, 50, 100, 200 and 400 patient dataset simulations, generated by randomly selecting records from the synthesised New York County population.

The controlled dimensions were sex, age and distance. The anonymity threshold  $k$  was set to 5 and the maximum perturbation distance  $\Delta$  to 1 kilometre. An exact match to sex was required (i.e. no perturbation allowed), and 5 levels of age categorisation were permitted (including exact age). Age categories were created by increasing the age range by one year for each successive level: for the first level, age range is 0, so it is the exact age; for level 2, the range is 1 to give age categories 0–1, 2–3, 4–5, etc. A simplified illustration of the implementation of the age categorisation is given in ►Figure 2. Note that level 5 matches the age categories used in the American Community Survey for age and sex stratified population counts [25].

Extent-limiting steps were also added to the algorithm to improve computational performance. At the beginning of each iteration and after creation of the patient dataset, one kilometre was added to the maximum and minimum latitudes and longitudes of the patient dataset and used to constrain the extent of the base population. This method was also used when determining the eligible population for each record.

The small random distance added to create  $R$  was restricted to a range of 1 and 10 metres to minimally impact geographic displacement. Distance was measured using the great circle formula.

Cumulative descriptive statistics (mean and standard deviations, as well as the minimum, maximum and median) were calculated for successive iterations to assess the effects of the transform on the perturbation distance. Analysis of the age dimension sought to identify the proportion of records requiring categorisation on age to achieve the required minimum  $k$ . The effect of adding the small random distance  $\delta_r$  on  $k$  was also described through descriptive statistics, as was the final perturbation radius. The displacement of the spatial mean of each patient dataset was also calculated in terms of perturbation distance.

The algorithm was coded and run in SAS v9.1; the results were also analysed in SAS v9.1 and graphed using Microsoft Of-

**Table 1** Results of the Multidimensional Point Transform (MPT) algorithm with different patient dataset sizes for New York County

	PATIENT DATASET SIZE ( $n$ )				
	25	50	100	200	400
<b>SETTINGS</b>					
Total records	25,000	50,000	100,000	200,000	400,000
k-anonymity setting	5	5	5	5	5
Maximum $\Delta$ setting (kms)	1	1	1	1	1
Unique Individuals	24,569	49,191	96,056	187,101	348,947
<b>MEASURES</b>					
Age-Perturbed records	0	1	1	0	1
<b>Perturbation Distance (kms)</b>					
Mean	0.046	0.046	0.046	0.046	0.046
Standard Deviation	0.035	0.034	0.035	0.034	0.034
Minimum	0	0	0	0	0
Median	0.039	0.039	0.039	0.039	0.039
Maximum	0.975	0.979	0.914	0.980	0.992
<b>Perturbation Radius (kms)</b>					
Mean	0.070	0.070	0.070	0.070	0.070
Standard Deviation	0.042	0.042	0.042	0.042	0.042
Minimum	0.004	0.003	0.002	0.003	0.002
Median	0.060	0.060	0.061	0.061	0.061
Maximum	0.994	0.997	0.983	1.000	1.005
<b>Actual k-anonymity level</b>					
Mean (rounded down)	5	5	5	5	5
Standard Deviation	1	1	1	1	1
Minimum	5	5	5	5	5
Median	6	6	6	6	6
Maximum	17	15	66	17	16
<b>Location Only k-anonymity</b>					
Mean (rounded down)	809	818	818	817	820
Standard Deviation	909	918	976	931	933
Minimum	19	19	15	19	7
Median	587	590	589	593	591
Maximum	38,122	43,782	68,449	66,100	58,965
<b>Spatial Mean Displacement (kms)</b>					
Mean	0.010	0.007	0.005	0.003	0.003
Standard Deviation	0.006	0.004	0.003	0.002	0.001
Minimum	0.0004	0.0002	0.0001	0.0002	0.0001
Median	0.009	0.007	0.005	0.003	0.002
Maximum	0.057	0.027	0.016	0.011	0.007

fice 2007. The code was run in RAM using RAMDisk software [26], as preliminary tests showed this to be approximately ten times faster than using a SATA 7200 RPM hard drive.

## 4. Results

In total, 775,000 records were randomly chosen from the synthesised New York population of 1,482,104 unique individuals and tested with this algorithm, representing 601,790 unique individuals or 41% of the New York county population (►Table 1).

Time taken to complete the algorithm ranged from about 5 minutes per iteration for the 25-patient dataset size, to just under two hours per iteration for the 400-patient dataset size.

The age dimension was seldom transformed, as summarised in ►Table 1. Only one record required an age-transform in each of the 50, 100 and 400-patient datasets, representing 0.0005% of the tested unique individuals.

The mean and median perturbation distances (46 metres and 39 metres respectively), as well as the mean and median perturbation radii (70 metres and 60 to 61 metres respectively), were consistent irrespective of the patient dataset size (►Table 1). Cumulative means of the distance between the original and the transformed points are presented in ►Figure 3 for successive iterations, showing a plateau within 1 metre after less than 200 iterations.

The actual  $k$ -anonymity achieved across all runs averaged five individuals, matching on sex and age within the defined perturbation radius as prescribed by the pre-de-

fined  $k$  requirement. Ignoring age and sex, the average number of individuals within the perturbation radius from which the random selection was made was just over 800 individuals.

The overall spatial mean of the transformed points was within 5.6 metres of the original spatial mean across all runs, and was inversely related to the patient dataset size (►Table 1).

## 5. Discussion

This study describes a multidimensional point transform (MPT) for anonymising data for public health use that includes location perturbation as a core component of the overall anonymisation algorithm. By perturbing the location dimension in concert with other user-defined dimensions,

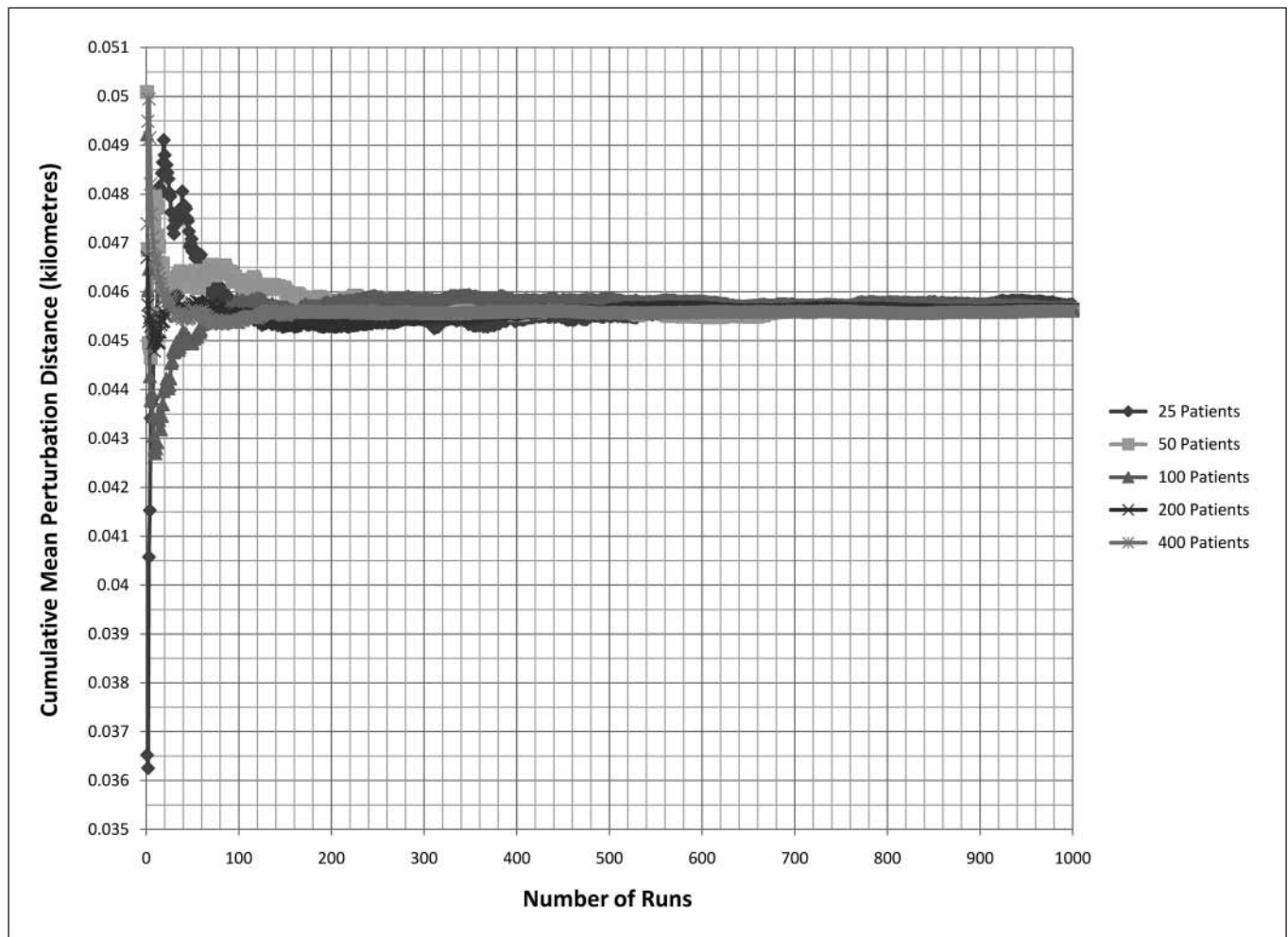


Fig. 3 Mean cumulative perturbation distance for successive runs of the tested patient dataset sizes

the MPT offers a more comprehensive and valid anonymised dataset than existing proposed random perturbation transforms.

Preliminary testing of the algorithm was completed on a synthesised New York County population using three dimensions: sex, age, and location. Since the sex perturbation tolerance was 0, there was no loss of information on the sex dimension; the age dimension was unaltered for 99.9995% of all runs; and the mean spatial displacement of transformed records was 46 metres from original patient points, irrespective of patient dataset size. Fifty percent of points being displaced across all runs were moved less than 40 metres and all transformed records were spatially accurate, representing actual household locations within the population.

Since New York County has a very high population density, it is not surprising that a  $k = 5$  anonymity level was achieved within such a small distance and with no impact on the age dimension. To further illustrate the multi-dimensional aspect of the algorithm, ad hoc analyses were completed using 100 iterations of sample sizes of 25 cases with  $k = 20$  and  $\Delta = 50$  m. With these tightened constraints, the MPT performed precisely as expected: 19 of the total 2500 records were transformed with no age-categorisation required; 148 were transformed with a level 2 age categorisation (i.e. a 1-year interval; recall level 1 is the exact age with no categorisation); 274 were transformed with a level 3 age categorisation; 292 were transformed with a level 4 age categorisation; and 290 were transformed with a level 5 age categorisation. The total successfully transformed records of the 2,500 was thus 1,023 (41%) with a mean perturbation distance of 31 metres. The remaining 1,477 records could not be transformed within 50 metres of the original point at a  $k = 20$  level, and all 1,477 reached level 5 age-categorisation as expected.

Previously described algorithms treat other record attributes separately from location, sometimes with a weighting to account for the generalised underlying population demographics. Armstrong et al assume that all other potentially identifying attributes such as health information, age,

sex, and so on are sufficiently “non-individual specific” [19]. Kwan et al [20] acknowledge the arbitrary nature of their weight factors, and that their results are specific to the unique and particular combination of their underlying population and the case-data. Similarly, Cassa et al. [21] used generalised age-based population density weights – which they refer to as “multipliers” – at the census block group level to implement a probabilistic Gaussian-skewed random perturbation transform.  $K$ -anonymity calculations were also probabilistically based, without taking into consideration other individual dimensions (e.g. age and sex). One could reasonably argue that, given the general importance of and requirement for these dimensions in public health practice, they should be included in the implementation and assessment of anonymisation algorithms.

As with the LP approach of Wieland et al [22], the MPT seeks to minimise perturbation distance and relies on an existing spatial population to do so. However, by accomplishing this in the context of multiple dimensions, it reveals a significant impact on the implications of the transform. Wieland et al. [22] suggest that aggregating to zip code in New York County yields a  $k$  value of about 884 and a corresponding perturbation distance of 519 metres; their LP method perturbation distance using census blocks is only 3.3 metres for the same  $k$ . A similar  $k$ -value acting on non-grouped individual points using the MPT requires a mean perturbation of 46 metres. However the actual  $k$  when taking the other dimensions into consideration drops dramatically to only five.

More recently, Hampton et al. [23] added a minimum perturbation distance to existing context-sensitive approaches [21] to implement a “donut” effect. The authors argue that allowing a case to be perturbed to its original location presents a re-identification risk, since an intruder will know that a few individuals may still be correctly identified, and also prevented cases from being perturbed outside their original administrative boundaries. The authors also suggest that their approach is adaptive because it adjusts for the underlying population density and also for minimum and maximum  $k$ -anonymity, whereas other

random perturbation methods are only “semiadaptive” because they fail to be bound by minimum anonymity constraints. However, the donut method as described is also semiadaptive because it fails to consider the details of the population demographics, such as age and sex. Given the aggregate use of underlying population density, the authors’ suggestion for the incorporation of these dimensions would have to rely on weighting mechanisms similar to those previously described [20, 21]. The donut-algorithm also does not address the possibility of randomly generating a point in a residentially-improbable or impossible location, such as a river or park, and is subject to a re-identification risk associated with multiple iterations [27, 28]. In contrast, the MPT can easily exclude original patient locations, incorporate a “donut-like” effect if desired, and retain points within defined geographies.

## 5.1 Re-Identification Risk

An issue with random perturbation algorithms is that repeated iterations on the same dataset increase the likelihood of re-identification [27, 28]. The MPT reduces this risk by selecting points from a defined base population instead of generating random points from a uniform distribution within a defined area. This avoids inaccurate or unrealistic placement and can skew the point pattern. The stochastic distance added to create  $R$  in the MPT also adds selection variability with each iteration. Therefore, the spatial mean of repeated iterations will depend on the variable spatial distribution of the underlying population and will not necessarily approximate the original location, unless the population is uniformly distributed around the patient.

A possible weakness is the prevention of selection of one case as a transform of another. Given two cases of identical age and sex within  $R$  of one another, it could be discovered that each is excluded from the transformed options for the other in favour of a more distant point, allowing potential re-identification of both original points. Removing the preventative selection criterion can resolve this, though it may also allow re-identification since repeated iter-

ations will result in case location selection twice as many times as others. Yet another potential for re-identification exists if the perturbation tolerances and thresholds used are known, though this would require extensive time and computing power. The multidimensional nature of the transform helps complicate re-identification efforts – the more dimensions are permitted to be perturbed, the more difficult re-identification becomes – while exact dimension matching is mitigated by the anonymity threshold.

The MPT as described does not anonymise records relative to themselves as is the general case with current  $k$ -anonymity techniques. In other words, it anonymises case records to an external dataset – in this case, a population dataset – and not to the case dataset itself. While this does not address potential “prosecutor re-identification” risk [29], the MPT can be configured to anonymise a dataset relative to itself if so desired as noted briefly below under Strengths. This approach, however, was not tested and is not the focus of the current study since in the context of location privacy it only works with generalised or aggregated locations.

It should be noted that meaningful privacy preservation is also a function of prevalence or incidence. For example, given 50 HIV cases in a population of 1,000, random selection of any one individual has a 5% chance of correctly identifying an HIV case. If all information were stripped for the 50 cases except for their location, each individual would be unidentifiable from 999 other individuals (i.e.  $k = 1,000$ ). However, while this has achieved the maximum possible  $k$  for this population, it still remains that correctly identifying an individual as having HIV has an effective  $k$ -value of only 20.

## 5.2 Limitations

The MPT relies on the presence of an underlying base population containing the same dimensions as those required by the data-user, with at least  $k - 1$  non-patients for each dimensionally-matched patient. As the number of dimensions increases, available matches decrease, potentially

necessitating dimensional compromises which can be controlled by increasing the allowable perturbation of the individual dimensions. The MPT allows exploration of the optimum context-specific combinations for appropriate data release and use.

Some issues were encountered that impacted overall performance, including periodic file locks, competing background applications, power outages, and system resources. Although using RAM allowed faster completion, future implementations may be limited by the amount available for allocation and machine specifications. Other performance-enhancing factors may include use of solid state drives, multi-threading and multiple processors, and coding and implementation within an environment other than SAS. Performance will also be a function of the underlying population matrix size.

No amount of masking, de-identification or anonymisation will prevent the misuse of data. Their release must therefore consider other factors such as the user’s trustworthiness, the purpose and scientific or applied merit for which they will be used, implemented security measures and so on.

## 5.3 Strengths

Strengths of the MPT algorithm are its powerful flexibility and customisability, easily allowing criteria to be set on appropriate dimensions relevant to both the study and the target population. For example, in the current implementation, every individual is indiscriminately associated with five age classes. However, some scenarios may require minimum age classes to be set, such that information deemed more sensitive is only released if age is categorised within the appropriate classes. Base population files containing only age classes based on census information can still be used, allowing ages to be classed differently based on the population distribution within the region of interest. The MPT can even be used to transform non-spatial dimensions of the base population for use in future implementations of the transform.

Another advantage of the MPT is its use of a granular base population which can be assigned to increasingly coarser geography. For example, given only postcodes, points can still be approximated using the base population and other dimensions provided in the patient record. Therefore, a 32-year-old female patient in postcode X1X1X1 can be assigned to any point within that postcode matching on age and sex within the allowable perturbations, and the MPT applied. This will incorporate a maximum error approximated by the sum of the maximum distance between residential points within the postcode area and the perturbation radius  $R$ , further confounding potential re-identification.

The MPT does not apply blanket rules to the entire patient dataset; rather, it anonymises each record individually for its own optimum transform. This allows release of the data with the best possible configuration. It also allows for more sophisticated integration of contextual information, facilitating comparison and calculation across datasets. For example, the algorithm could include distance to the nearest school as an added dimension. A maximum spatial displacement of each record from its closest school can be specified in combination with other relevant dimensions, preserving the relative spatial distance to schools. The algorithm is also not independent of the underlying geography. Because it uses pre-established locations for the random selection to meet the required anonymity threshold, knowledge of the existence of non-inhabitable regions or features will not increase re-identification potential (a noted issue with random perturbation techniques [19, 22]). These factors allow the spatial aspect of the transform itself to be bound by multiple, contextually appropriate rules.

If specific dimensions are not known *a priori*, such as education and income, an areal dimension can be added as part of the control to allow retention of the patient within the specified political or administrative boundary. This allows the flexibility to use as little or as much data as are available to achieve optimum results. The advantage to including additional dimensions beyond administrative or political boundaries is the incorporation of actual

contextual variables as opposed to potentially artificially-related areal units.

As mentioned, New York county is extremely population dense making it relatively easy to achieve reasonable anonymity with very little spatial displacement, even when multiple dimensions are considered. As the *ad hoc* analyses show, however, the MPT allows users and custodians to identify this and modify the parameters in order to achieve acceptable results. In this case, for example, the custodian may agree to lower the  $k$ -value if acceptable or pending certain requirements on the part of the user (e.g. use restrictions, security requirements, etc). Conversely or simultaneously, the user may accept additional perturbations (i.e. of sex or age) or increased spatial perturbation. The same decisions would have to be made for a sparsely populated rural area; either way, population density does not impair the MPT. By allowing the user and custodian to have control over the various aspects of the transform, including the appropriate or acceptable anonymity threshold, the MPT provides a “user-sufficient mask” [19].

## 5.4 Using Synthesised Populations

Health data are most valuable and informative in their most granular form, and developing a transform that works on individual point-level data at the address level is highly beneficial. However, such a transform would require knowledge of the underlying population – also at the individual point-address level. Although available through population registries, these data are themselves subject to privacy and confidentiality restraints, and are therefore generally not accessible for public health use. Instead, public health practitioners rely on aggregated census data to infer various population demographics. This is where synthesised populations may play a role. For example, a “synthesised, geospatially explicit” US population based on the year 2000 census has been generated to facilitate agent-based infectious disease modeling for the Modeling of Infectious Disease Agents Study (MIDAS) [30]. This population “correctly and appropriately” describes the age and sex demographics by

household, and accurately reflects the actual US population. Details on the methodology and population characteristics have been published [30].

Since the MPT makes use of a synthesised population, its validity depends on how well the synthesised data mirror reality on the dimensions of interest. Since the population is based on the year 2000 census, it may inadequately reflect population demographics for earlier or more recent studies. However, given the recurring nature of the census, algorithms used to build the synthesised populations can be re-run to generate new and relevant populations with each census year [30]. A synthesised population may also be invaluable in exploring the relationships between perturbation distance and a variety of quasi-identifiers as illustrated through this study. Their use also allows for the creation of realistic, non-circular disease clusters for investigation – an issue that impacts other studies in this field [23].

Synthesised populations for the US and several other countries have been produced for MIDAS and are available by request. These populations were developed for epidemiological modeling, not for de-identification algorithms, further highlighting their general utility in public health. As such, the development of representative synthetic populations would be highly beneficial. Indeed, development of a synthetic 2010 US population is currently underway by MIDAS scientists, as are tools to allow researchers to generate custom populations based on demographic variables of interest [24].

## 5.5 Algorithm Refinement

Further refinement of the MPT could allow the user to set priority levels for the various dimensions. In the current example, the priority is given to age; age is perturbed only if the anonymisation threshold is not met within the prescribed maximum distance. Instead, the algorithm can be modified to prioritise minimum distance moved within a maximum age perturbation (i.e. the algorithm could begin with the maximum age perturbation to minimise distance and work backwards to achieve the

optimum result). This provides maximum flexibility in exploring the optimal transform for a given dataset and context, as minimising changes in one dimension will necessarily impact the effect of other dimensions.

As an example, assume our dimensions of interest are distance, age, sex and race with decreasing priority assignment. In this case, the MPT as illustrated in ►Figure 1 will first search for  $k - 1$  exact matches on age, sex and race. In the absence of meeting this requirement, it will generalise race within the defined generalisation threshold and look for  $k - 1$  exact matches on age, sex and generalised race. Assuming it still fails, it will then generalise sex, and look for  $k - 1$  exact matches on age and generalised sex and race. And so on. Based on the current design, it will only move on to generalising the next dimension once it has reached the maximum designated generalisation of the previous dimensions with failure to identify  $k - 1$  matches in the population, since the loop is intra-dimensional. The loop can also easily be changed to allow several dimensional generalisations within an iteration – i.e. across dimensions. In this case, and using the same example, race would be transformed to its first generalisable level, followed by sex if required, then age; assuming failure, it would then loop back and generalise race to its next level, etc. The intent is to minimise loss on those dimensions deemed by the user to be more important to retain closer to their original value, as opposed to finding an overall perceived “optimal” solution, while minimising spatial disturbance (i.e. distance perturbed).

The MPT settings can also be informed by other research in this area. For example, the maximum number of combinations (MaxCombs) of variables of interest is a good predictor of uniqueness [12, 31] and can be used to determine appropriate “geographic area population size” (GAPS) [31] for privacy preservation. This can be used to inform preliminary decisions on setting  $k$  and  $\Delta$  for the MPT; for example, one can begin by setting  $D$  to the approximate mean radius of the census geography most closely corresponding to a calculated GAPS cutoff. MaxCombs can also be used to inform the dimensional categorisation levels, particu-

larly since it is dependent solely on the number of response categories and not the types of the quasi-identifiers.

It has been shown that  $k$ -anonymity can, in some cases, be “over-protective”, particularly for smaller sampling fractions, resulting in unnecessary information loss [29]. The current methodology helps reduce such information loss by incorporating the relevant dimensions directly into the anonymisation algorithm, allowing the user to set permissible categorisation and priority levels, and performing local recoding (i.e. allowing observations to have different and overlapping response intervals [29]). Appropriate  $k$  values should be a function of the user and the use of the data, as well as the governance structures in place. Some general criteria for setting this threshold have been proposed [12] and should be incorporated into a more comprehensive framework for the disclosure of data.

Preliminary MPT testing was conducted on three dimensions: sex, age and location. Additional dimensions, larger patient datasets and different base populations with varying population densities should also be explored, as should the effects on common spatial statistics used in public health. Since random perturbation techniques generally increase Type II error probability (e.g. cluster dilution) and do not affect Type I error probability [19], further studies on appropriate thresholds and applications of this algorithm are required in various contexts and with different base populations. Additional analyses quantifying the relationship between the anonymity level achieved and the distance displaced for specific contexts and base populations can also serve as part of a framework for assessing appropriate uses. Currently, privacy legislation applies to “identifiable individuals”. However, with the growing literature around anonymisation, one can now ask the question “at what  $k$ -value does an individual cease to become identifiable under the legislation”? Acceptable anonymity thresholds therefore need to be set and standardised, and the legislation needs to be revised to better reflect this in privacy definitions.

Sophisticated software agents [19, 32] could be used to combine the ingredients required (e.g. the base population from a

municipal population registry, the health data from the custodian, and the user requirements) and return an appropriately and optimally transformed dataset (or null result, if no adequate transform is feasible given the data and user specifications). This allows the user to explore analyses that may only become evident after visual exploration of the data’s distribution. A graphical user “front” would be highly beneficial for this purpose, and an image of such an interface is suggested at <http://www.personplacetime.org/tools/MPTinterface.jpg>.

## 6. Conclusions

The multidimensional point transform proposed in this study works concertedly on multiple attributes, *including the spatial attributes*, to give a more complete and appropriate transform that builds location privacy into the anonymisation model from the beginning. Unlike previous studies, this algorithm does not leave other attributes “untouched”, but it does result in a transformed matrix with the same dimensions of the original matrix [19].

The ideal transform preserves the confidential and private nature of individual health records, as well as the geographic integrity of the data, to facilitate public health practice [19]. The optimal approach depends not only on the purpose for the data use and the acceptable disclosure risk [19], but also on the characteristics of the data. Acceptable disclosure risk by the custodian is also a multifaceted consideration based on acceptable anonymity thresholds, trust in the user, adequate security measures, and so on. However, such algorithms cannot substitute for secure and ethical conduct, and a framework for the appropriate disclosure, use and dissemination of data containing personal identifiable information is required [1]. There are also instances in which the release and use of identifiable information in public health are essential [33], and consideration must be made within a developed framework to allow for such cases. The proposed algorithm in this study presents a multidimensional approach that allows one to tweak and optimise the trade-offs for any given dataset

and purpose, presenting a necessary component of the much-needed public health framework.

## Acknowledgements

Many thanks to Professor Ray Jones at the University of Plymouth, Dr. Khaled El Emam at the CHEO Research Institute, Mr. Philip Ng at the Public Health Agency of Canada, Mr. Sami AbdelMalik, Mrs. Janet Honig and Miss Miriam Rawson for providing support, resources, insight, editing and sounding boards.

## Competing Interests

The authors declare that they have no competing interests.

## Authors’ Contributions

PA conceived and drafted the transform and manuscript, and conducted the preliminary implementation and analysis. MNKB critically revised the manuscript and contributed to its content and flow. Both authors read and approved the final manuscript.

## References

1. Boulos MNK, Curtis AJ, AbdelMalik P. Musings on privacy issues in health research involving disaggregate geographic data about individuals. *International Journal of Health Geographics* 2009; 8: 46. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2716332/pdf/1476-072X-8-46.pdf>
2. Wartenberg D, Thompson WD. Privacy versus public health: the impact of current confidentiality rules. *American Journal of Public Health* 2010; 100 (3): 407–412.
3. De Moor GJE, Claerhout B, De Meyer F. Privacy enhancing techniques. *Methods Inf Med* 2003; 42: 148–153.
4. AbdelMalik P, Boulos MNK, Jones R. The perceived impact of location privacy: a web-based survey of public health perspectives and requirements in the UK and Canada. *BMC Public Health* 2008; 8: 156. <http://www.biomedcentral.com/1471-2458/8/156>
5. Saxena N, MacKinnon MP, Watling J, Willison D, Swinton M. Understanding Canadians’ attitudes and expectations: Citizens’ dialogue on privacy and the use of personal information for health research in Canada. Report (Research Report P|09): Canadian Policy Research Networks Inc.; March 2006
6. Robling MR, Hood K, Houston H, Pill R, Fay J, Evans HM. Public attitudes towards the use of primary care patient record data in medical research without consent: a qualitative study. *Journal of Medical Ethics* 2004; 30: 104–109. <http://jme.bmj.com/cgi/content/abstract/30/1/104>



7. Jones C. The utilitarian argument for medical confidentiality: a pilot study of patients' views. *Journal of Medical Ethics* 2003; 29 (6): 348–352. <http://jme.bmj.com/cgi/content/abstract/29/6/348>
8. Barrett G, Cassell JA, Peacock JL, Coleman MP. National survey of British public's views on use of identifiable medical data by the National Cancer Registry. *British Medical Journal* 2006; 332: 1068–1072.
9. Onsrud HJ, Johnson JP, Lopez XR. Protecting personal privacy in using geographic information systems. *Photogrammetric Engineering & Remote Sensing* 1994; 60 (9):1083–1095.
10. Domingo-Ferrer J, Torra V: A critique of *k*-anonymity and some of its enhancements. In: *IEEE* 2008.
11. Dalenius T. Finding a needle in a haystack or identifying anonymous census records. *Journal of Official Statistics* 1986; 2 (3): 329–336.
12. El Emam K, Brown A, AbdelMalik P, Neisa A, Walker M, Bottomley J, Roffey T. A method for managing re-identification risk from small geographic areas in Canada. *BMC Medical Informatics and Decision Making* 2010; 10: 18. <http://www.biomedcentral.com/1472-6947/10/18>
13. Claeihout B, De Moor GJE. Privacy protection for HealthGrid applications. *Methods Inf Med* 2005; 44: 140–143.
14. Sweeney L. *k*-Anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 2002; 10 (5): 557–570.
15. Meyerson A, Williams R. General *k*-anonymization is hard. Report (CMU-CS-03-113). Pittsburgh (PA): School of Computer Science, Carnegie Mellon University; 2003.
16. Curtis AJ, Mills JW, Agustin L, Cockburn M. Confidentiality risks in fine scale aggregations of health data. *Computers, Environment and Urban Systems* 2010. In press. doi: 10.1016/j.compenvurbsys.2010.08.002
17. Boulos MNK. Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. *International Journal of Health Geographics* 2004; 3: 1. <http://www.ij-healthgeographics.com/content/3/1/1>
18. Gutmann MP, Stern PC. Putting people on the map: protecting confidentiality with linked social-spatial data. Washington, D.C.: The National Academies Press; 2007.
19. Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 1999; 18: 497–525. <http://www3.interscience.wiley.com/cgi-bin/fulltext/45002090/PDFSTART>
20. Kwan Mei-Po, Casas I, Schmitz BC. Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks? *Cartographica* 2004; 39 (2): 15–28.
21. Cassa CA, Grannis SJ, Overhage JM, Mandl KD. A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection. *JAMIA* 2006; 13 (2): 160–165.
22. Wieland SS, Cassa CA, Mandl KD, Berger B. Revealing the spatial distribution of a disease while preserving privacy. *Proceedings of the National Academy of Sciences of the United States of America* 2008; 105 (46): 17608–17613.
23. Hampton KH, Fitch MK, Allshouse WB, Doherty IA, Gesink DC, Leone PA, Serre ML, Miller WC. Mapping health data: improved privacy protection with donut method geomasking. *American Journal of Epidemiology* 2010; 172 (9): 1062–1069.
24. Models of Infectious Disease Agent Study. Synthesized data. Last updated: 2009. Accessed: <https://www.epimodels.org/midas/pubsyntdata1.do>
25. 2005 American Community Survey: Age and sex population. Last updated: 2008. Accessed: 11-12-2010. [http://factfinder.census.gov/servlet/STTable?\\_bm=y&-geo\\_id=01000US&-qr\\_name=ACS\\_2005\\_EST\\_G00\\_S0101&-ds\\_name=ACS\\_2005\\_EST\\_G00\\_&-lang=en&-redoLog=false](http://factfinder.census.gov/servlet/STTable?_bm=y&-geo_id=01000US&-qr_name=ACS_2005_EST_G00_S0101&-ds_name=ACS_2005_EST_G00_&-lang=en&-redoLog=false)
26. RAMDisk software. Last updated: 2010. Accessed: <http://memory.dataram.com/products-and-services/software/ramdisk/>
27. Zimmerman DL, Pavlik C. Quantifying the Effects of Mask Metadata Disclosure and Multiple Releases on the Confidentiality of Geographically Masked Health Data. *Geographical Analysis* 2008; 40 (1): 52–76. <http://www3.interscience.wiley.com/journal/119390400/stract?CRETRY=1&SRETRY=0>
28. Cassa CA, Wieland SC, Mandl KD. Re-identification of home addresses from spatial locations anonymized by Gaussian skew. *International Journal of Health Geographics* 2008; 7 (45). <http://www.ij-healthgeographics.com/content/7/1/45>
29. El Emam K, Dankar FK. Protecting Privacy Using *k*-Anonymity. *JAMIA* 2008; 15: 627–637. <http://www.jamia.org/cgi/content/abstract/15/5/627>
30. Wheaton WD, Cajka JC, Chasteen BM, Wagener DK, Cooley PC, Ganapathi L, Roberts DJ, Allpress JL. Synthesized population databases: a US geospatial database for agent-based models. Report (RTI Press publication No. MR-0010-0905). Research Triangle Park, NC: RTI International; May 2009.
31. El Emam K, Brown A, AbdelMalik P. Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk. *JAMIA* 2008; 16 (2): 256–266. <http://www.jamia.org/cgi/content/full/16/2/256>
32. Boulos MNK, Cai Q, Padget JA, Rushton G. Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses. *Journal of Biomedical Informatics* 2006; 39 (2): 160–170. <http://www.sciencedirect.com/science/article/B6WHD-4GR32TM-1/2/be0cb959aa15839693f49f582633e59b>
33. Black N. Secondary use of personal data for health and health services research: why identifiable data are essential. *Journal of Health Services Research and Policy* 2003; 8 (Suppl 1): S1:36 – S1:40. <http://www.ncbi.nlm.nih.gov/pubmed/12869337>

RESEARCH ARTICLE

Open Access

# A method for managing re-identification risk from small geographic areas in Canada

Khaled El Emam<sup>1,2\*</sup>, Ann Brown<sup>1</sup>, Philip AbdelMalik<sup>3</sup>, Angelica Neisa<sup>1</sup>, Mark Walker<sup>4</sup>, Jim Bottomley<sup>5</sup>, Tyson Roffey<sup>5</sup>

## Abstract

**Background:** A common disclosure control practice for health datasets is to identify small geographic areas and either suppress records from these small areas or aggregate them into larger ones. A recent study provided a method for deciding when an area is too small based on the uniqueness criterion. The uniqueness criterion stipulates that an area is no longer too small when the proportion of unique individuals on the relevant variables (the quasi-identifiers) approaches zero. However, using a uniqueness value of zero is quite a stringent threshold, and is only suitable when the risks from data disclosure are quite high. Other uniqueness thresholds that have been proposed for health data are 5% and 20%.

**Methods:** We estimated uniqueness for urban Forward Sortation Areas (FSAs) by using the 2001 long form Canadian census data representing 20% of the population. We then constructed two logistic regression models to predict when the uniqueness is greater than the 5% and 20% thresholds, and validated their predictive accuracy using 10-fold cross-validation. Predictor variables included the population size of the FSA and the maximum number of possible values on the quasi-identifiers (the number of equivalence classes).

**Results:** All model parameters were significant and the models had very high prediction accuracy, with specificity above 0.9, and sensitivity at 0.87 and 0.74 for the 5% and 20% threshold models respectively. The application of the models was illustrated with an analysis of the Ontario newborn registry and an emergency department dataset. At the higher thresholds considerably fewer records compared to the 0% threshold would be considered to be in small areas and therefore undergo disclosure control actions. We have also included concrete guidance for data custodians in deciding which one of the three uniqueness thresholds to use (0%, 5%, 20%), depending on the mitigating controls that the data recipients have in place, the potential invasion of privacy if the data is disclosed, and the motives and capacity of the data recipient to re-identify the data.

**Conclusion:** The models we developed can be used to manage the re-identification risk from small geographic areas. Being able to choose among three possible thresholds, a data custodian can adjust the definition of "small geographic area" to the nature of the data and recipient.

## Background

The disclosure and use of health data for secondary purposes, such as research, public health, marketing, and quality improvement, is increasing [1-6]. In many instances it is impossible or impractical to obtain the consent of the patients *ex post facto* for such purposes. But if the data are de-identified then there is no legislative requirement to obtain consent.

The inclusion of geographic information in health datasets is critical for many analyses [7-15]. However,

the inclusion of geographic details in a dataset also makes it much easier to re-identify patients [16-18]. This is exemplified by a recent Canadian federal court decision which noted that the inclusion of an individual's province of residence in an adverse drug event dataset makes it possible to re-identify individuals [19,20].

Records from individuals living in small geographic areas tend to have a higher probability of being re-identified [21-23]. Some general heuristics for deciding when a geographic area is too small with respect to identifiability have been applied by national statistical agencies [24-29]. For example, the US Health Insurance

\* Correspondence: kelemam@uottawa.ca

<sup>1</sup>Children's Hospital of Eastern Ontario Research Institute, 401 Smyth Road, Ottawa, Ontario K1J 8L1, Canada

Portability and Accountability Act (HIPAA) Privacy Rule defines a small geographic area as one having a population smaller than 20,000.

Common disclosure control actions for managing the re-identification risks from small geographic areas are to: (a) suppress records in the small geographic areas, (b) remove from the disclosed dataset some of the non-geographic variables, (c) reduce the number of response categories in the non-geographic variables (i.e., reduce their precision), or (d) aggregate the small geographic areas into larger ones. None of these options is completely satisfactory in practice. Options (a) and (b) result in the suppression of records or variables respectively. The former leads to the loss of data and hence reduces the statistical power of any analysis, and can also result in bias if the suppressed records are different in some important characteristics from the rest of the data. The latter is often difficult to implement because variables critical to the analysis of the data cannot be removed. Options (c) and (d) reduce the precision of the information in the dataset through generalization. The former generalizes the non-geographic information in the dataset which may make it difficult to detect subtle trends and relationships. The latter can reduce the ability to perform meaningful analysis and can conceal variations that would otherwise be visible at smaller geographical scales [30-35].

Given the detrimental effects of such disclosure control actions, it is important to have accurate and proportionate methods for assessing when a geographic area is too small.

The uniqueness of individuals is often used as a surrogate measure of re-identification risk [36]. An individual is unique if s/he is the only individual with a specific combination of values on their personal characteristics that are included in a dataset. There is a monotonically decreasing relationship between uniqueness and geographic area population size: uniqueness decreases as population size gets larger. A recent study developed a model to decide when a geographic area is too small based on the uniqueness of its population [37]: if uniqueness within a geographic area is approximately zero then the geographic area is not too small.

However, using zero uniqueness as a threshold for disclosure control is quite stringent and can result in excessive record or variable suppression and/or aggregation. Higher uniqueness thresholds have been found acceptable and have been applied in practice. Specifically, previous disclosures of cancer registry data have deemed thresholds of 5% and 20% population uniqueness as acceptable for public release and research use respectively [38-40].

In this paper we extend this line of work by developing models to determine whether a Forward Sortation

Area (FSA - the first three characters of the Canadian postal code) is too small based on the 5% and 20% uniqueness thresholds by analyzing Canadian census data. We also provide data release risk assessment guidelines for deciding which one among the 0%, 5%, and 20% threshold models to use for disclosure control.

## Methods

Our approach was to construct models to determine if the percentage of unique records in a particular FSA was above the 5% and the 20% thresholds. These models characterize each FSA in terms of its population size, and also take into account the characteristics of the non-geographic variables in the dataset that can be used for re-identification.

## Definitions

### Quasi-identifiers

The variables in a dataset that can be used to re-identify individuals are called the *quasi-identifiers* [41]. Examples of common quasi-identifiers are [37,42-44]: dates (such as, birth, death, admission, discharge, visit, and specimen collection), race, ethnicity, languages spoken, aboriginal status, and gender.

### Equivalence Class

An equivalence class is defined as the group of records having a given set of values on the quasi-identifiers. For example, "50 year old male" represents the equivalence class of records with the "50" value on the age quasi-identifier and "Male" on the gender quasi-identifier. The number of records that have these two values on the quasi-identifiers is the size of the "50 year old male" equivalence class.

### Uniqueness

The uniqueness of records in the dataset is based only on the quasi-identifiers. For example, if our quasi-identifiers are age and gender, then say, the only 90 year old female in the FSA "N3E" would be a unique record on these quasi-identifiers within that geographic area. Other sensitive variables that are not considered quasi-identifiers are not taken into account in the computation of uniqueness. If an equivalence class is of size one, then that represents a unique record.

### Focus on the Forward Sortation Area (FSA)

The postal code is the basic geographical unit that we will use in our analysis. The postal code is frequently collected because it is readily available, and consequently, it is used as the geographical location of residence in health datasets [45-50]. The full six character postal code is often more specific than needed for many analyses. Further, in combination with other variables the full postal code would make it easy to re-identify individuals, especially in residential urban areas [43]. While there are many potential ways of aggregating

geographic regions to construct larger areas for analysis [35], the FSA, a higher level in the postal code geographic hierarchy, is the unit that we considered.

### Dataset

The dataset we used comes from the 2001 Canadian census. The census has two forms: the short form and the long form. Approximately a 20% sample of the population completes the long form, and the remainder completes the short form. The long form individual level data is made available to researchers by Statistics Canada through its Research Data Centers (RDCs).

The RDC long form dataset only has geographic information at the level of the census tract. Because our desired analysis is at the FSA geographic unit, we developed a gridding methodology, described in Additional file 1, to assign the FSAs to individual records based on their census tracts. Census tracts are only defined for urban areas and do not cover Prince Edward Island (PEI). Therefore, rural FSAs and PEI were excluded from our analysis.

Table 1 contains the list of quasi-identifiers that were analyzed from the long form census file. These were selected to be representative of commonly used quasi-identifiers in health and health systems research. The table also includes the number of response categories for each quasi-identifier as they were used in our analysis.

### Quasi-identifier Models

A quasi-identifier model consists of two or more quasi-identifiers (*qid*). To manage the scope of the analysis we consider only combinations of up to and including

5 *qids*. A total of 358 quasi-identifier models were analyzed. This results from the following approach of combining the *qids*.

Initially, for the 11 *qids* listed in Table 1, there are some similarities related to ethnicity and therefore they were treated as a group: HLNABDR, ETH1-6, RELIGWI, and DVISMIN. We defined a generic ethnicity variable, and whenever that generic ethnicity variable appears in a model it was replaced by one of the above four variables. Each substitution represented a different model. Thus, this gives 8 distinct *qids*: gender, age, ethnicity (generic), schooling, marital status, total income, aboriginal identity and activity difficulties.

Categorizing the 8 distinct *qids* by their utility by an intruder for re-identification gives the following two types:

- High utility to an intruder: gender, and age
- Possibly used for re-identification/sensitive: ethnicity, schooling, marital status, total income, aboriginal identity and activity difficulties

The different models were defined by the number of *qids* in the model and by having at least one sensitive *qid* included in each model.

For models including both age and gender, there are 42 models for the 8 distinct *qids* as follows:

- 5 *qids*: have age and gender and 20 combinations of 3 of the 6 sensitive *qids*.
- 4 *qids*: have age and gender and 15 combinations of 2 of the 6 sensitive *qids*.
- 3 *qids*: have age and gender and each of the 6 sensitive *qids*.

**Table 1 The list of quasi-identifiers that were analyzed from the census file**

Variable Name in the 2001 Census RDC File	Definition	# Response categories <sup>(*)</sup>
SEXP	Gender	2
BRTHYR	Year of birth (from 1880 to 2001). Age: We defined age categories based on 5 year ranges.	24
HLNABDR	Language: Language spoken most often at home by the individual at the time of the census.	4
ETH1-6	Ethnic Origin: Refers to the six possible answers for the ethnic or cultural group(s) to which the respondent's ancestors belong.	26
ASRR	Aboriginal Identity: Persons identifying with at least one Aboriginal group.	8
RELIGWI	Religious denomination: Specific religious denominations, groups or bodies as well as sects, cults, or other religiously defined communities or systems of belief.	3
TOTYRSR	Total Years of Schooling: Total sum of the years (or grades) of schooling at the elementary, high school, university and college levels. Only available for individuals age 15+.	9
MARST	Marital Status (Legal)	5
TOTINC	Total income: Total money income received from all sources during the calendar year 2000 by persons 15 years of age and over. We defined categories in \$15K ranges.	22
DVISMIN	Visible minority status	4
DISABIL	Activity difficulties/reductions: Combinations of one or more activity difficulties/reduction.	4

<sup>(\*)</sup>The number of response categories excludes non-specific responses such as missing values, not available or "other".

- 2 *qids*: have age and gender only - there is only one model.

Then substituting each of language, religion and visible minority for ethnicity gives an additional 48 models: 30 ( $3 \times 10$ ) models for 5 *qids* (ethnicity appears in 10 of the 20 models), 15 ( $3 \times 5$ ) models for 4 *qids* (ethnicity appears in 5 of the 15 models), and 3 ( $3 \times 1$ ) models for 3 *qids* (ethnicity appears in one of the 6 models).

The subtotal for this group of models containing both age and gender is 90 (42+48).

We repeated the above process for each *one* of age and gender in combination with the sensitive *qids*. That is there are 56 models containing:

- 5 *qids*: have age and 15 combinations of 4 of the 6 sensitive *qids*.
- 4 *qids*: have age and 20 combinations of 3 of the 6 sensitive *qids*.
- 3 *qids*: have age and 15 combinations of 2 of the 6 sensitive *qids*.
- 2 *qids*: have age and each of the 6 sensitive *qids* only.

Similarly to the previous group, by taking into account the ethnicity related variables, there are a sub-total of 134 models for this group.

Lastly, age is replaced with gender for an additional 134 models. Adding up the sub-totals gives a total number of 358 quasi-identifier models.

For each quasi-identifier model, we denote its maximum number of equivalence classes as its *MaxCombs* value. The *MaxCombs* value for any quasi-identifier model can be computed from Table 1. For example, if we consider the four quasi-identifiers: Age, Marital Status, Schooling, Religion, then there are  $24$  (age)  $\times$   $5$  (marital status)  $\times$   $9$  (years of schooling)  $\times$   $3$  (religion) = 3,240 possible values on these variables, which is the *MaxCombs* value. The *MaxCombs* values range from 6 to 718,848 across all quasi-identifier models.

### Estimating Uniqueness

There are a number of different approaches that can be used to estimate uniqueness in the population from the 20% sample.

The first study to examine uniqueness in the general population was conducted in the US by Sweeney [51]. Relying on the generalized Dirichlet drawer principle, she made inferences about uniqueness in specific geographic areas. This principle states that if  $N$  objects are distributed in  $k$  boxes, then there is at least one box containing at least  $\lceil \frac{N}{k} \rceil$  objects (i.e., the largest integer within the brackets). If  $N \leq k$  then there is at least one box with a single object (i.e., a unique).

Sweeney made the conservative assumption that if there is any unique in a particular geographic area, say an FSA, then that FSA is high risk. She then reported the percentage of individuals in high risk geographic areas. For example, if we consider a quasi-identifier model with a *MaxCombs* value of 48 (the  $k$  value), then any FSA with a population smaller than 48, say 15 (the  $N$  value), would likely have a unique individual in it, and therefore all 15 individuals would be considered at a high risk of uniqueness.

However, this approach will tend to overestimate the percentage of uniques because not all individuals in the FSA will be unique. For example, in the case above, on average, 26% of the 15 individuals would be non-unique. Furthermore, the Sweeney method does not help us with estimating if uniqueness is above 5% or 20% for a particular FSA.

An earlier study, which predicted when a geographic area is too small, was based on the zero uniqueness threshold utilizing a public use census file [37]. That study assumed that as sample uniqueness approached zero, the population uniqueness also approached zero. This assumption is not suitable for directly estimating population uniqueness at a 5% or 20% threshold.

Another approach to estimate equivalence class sizes was taken by Golle [52], where he assumed a uniform distribution of dates of birth of individuals living in a geographic area in assigning them to equivalence classes. However, that approach was driven by the author only having access to high level census tabulations, and was limited to a single variable. In our case the uniform distribution assumption cannot be justifiably extended to all of the quasi-identifiers.

For our analysis we used the individual-level Canadian census dataset. Given that the long form census dataset is a 20% sample of the Canadian population, we utilized uniqueness estimators to determine the proportion of unique records for each combination of FSA and quasi-identifier model. The reason we need to estimate population uniqueness is because sample uniqueness does not necessarily equate to population uniqueness, and we are interested in population uniqueness.

One estimator developed by Bethlehem et al. [36,53] over-estimates with small sampling fractions and under-estimates as the sampling fraction increases [54]. We therefore adopted a different estimation approach developed by Zayatz [31,55]. While this approach tends to over-estimate the number of population uniques for small sampling fractions, our 20% sampling fraction would be large enough to alleviate concerns about bias [54].

### Prediction Models

Based on the uniqueness estimate for each quasi-identifier model and FSA, two binary variables were

constructed: the first is 1 if the estimated uniqueness for a particular FSA and quasi-identifier model was above 5% and zero otherwise, and the second was 1 if the estimated uniqueness was above 20% and zero otherwise.

This is illustrated in Table 2 through a series of examples. Here we have seven example FSAs, and for each one a set of quasi-identifiers (quasi-identifier model) is shown. For example, for the “K7N” we have the “age × sex” quasi-identifier model. For each FSA and quasi-identifier model combination we show the uniqueness estimate. Recall that we only have data on 20% of the population, therefore the uniqueness estimate gives us the percentage of individuals in that FSA who are unique on their quasi-identifier values. For instance, in “L6P” 16.7% of the population are unique on their gender, aboriginal status, schooling, and language spoken at home. The last two columns of the table indicate whether the estimated uniqueness is greater than 5% and greater than 20% respectively. Such a table was constructed for all FSAs and for all quasi-identifier models. This table had 342,606 rows.

We developed one binary logistic regression model [56] with the 5% binary variable (denoted by  $I_{05}$ ) as the response variable, and another with the 20% binary variable (denoted by  $I_{20}$ ) as the response variable. The predictor variables in this model characterize the FSA and the quasi-identifiers in the quasi-identifier model.

An FSA can be characterized by its population size, which was obtained from the census data. We denote this variable by  $POP$ . For example, the “K7N” FSA in Table 3 has a  $POP$  value of 6,228, and the “L6P” FSA has a  $POP$  value of 2,247. The  $POP$  variable ranged from 200 to 78,457.

In a previous study it was shown that  $MaxCombs$  was a good predictor of uniqueness [37]. We therefore use it to characterize the quasi-identifier model used. Table 3 includes the  $MaxCombs$  values for each of the quasi-identifier models in our example, as well as the response variables for the logistic regression models. The data in Table 3 are an example of the raw values that we used in building the regression models. An observation is an

**Table 3 Example of what the raw data used to build the models looked like.**

Example of Raw Data Used in Building the Logistic Regression Models				
ID	POP	MaxCombs	$I_{05}$	$I_{20}$
1	6,228	48	0	0
2	14,047	576	0	0
3	100	40	1	0
4	2,247	576	1	0
5	12,916	84,480	1	1
6	7,080	9,360	1	1
7	100	95,040	1	1

(b) The population uniqueness binary value is used in the logistic regression model with the other predictor variables. We used the 2001 Canadian Census population values.

FSA by quasi-identifier model combination (as shown in Table 3). For example, there is one observation for the “K7N” FSA for the quasi-identifier model “age × sex”.

The 5% model was defined as:

$$\text{logit}(\pi_{05}) = b_0 \times POP + b_1 \times MaxCombs + b_2 \times (POP \times MaxCombs)$$

where  $\pi_{05}$  is the probability that an observation is high risk (uniqueness greater than 5%) and the  $b$  parameters were estimated. The logistic regression models were estimated and evaluated using SAS version 9.1. We included an interaction term in the model so that we can adjust the relationship between  $MaxCombs$  and uniqueness according to the population size of the FSA (instead of creating a separate model for each FSA). The 20% model was similarly constructed.

To avoid collinearity with the interaction term in the model, both predictor variables were centered [57]. Collinearity occurs when there are linear dependencies among the predictor variables, and between predictor variables and the intercept [58]. Because both  $POP$  and  $MaxCombs$  have large values, the interaction term in the logistic regression model can create overflow problems during computation. We therefore scaled the predictor variables by 10,000.

**Table 2 Example uniqueness estimates, POP and MaxCombs values for some FSA and quasi-identifier combinations.**

Example of Uniqueness Estimates for FSA and Quasi-identifier Model Combinations					
ID	FSA	Quasi- Identifiers	Uniqueness ( $\hat{U}$ )	$\hat{U} > 5\%$	$\hat{U} > 20\%$
1	K7N	Age, Sex	0%	N	N
2	M2K	Age, Aboriginal, Religion	1.7%	N	N
3	K1A	Sex, Marital Status, Language	14.3%	Y	N
4	L6P	Sex, Aboriginal, Schooling, Language	16.7%	Y	N
5	H3T	Age, Aboriginal, Income, Marital Status, Language	56.0%	Y	Y
6	L1 M	Sex, Disability, Marital Status, Schooling, Ethnicity	67.80%	Y	Y
7	K1A	Age, Disability, Income, Marital Status, Schooling	94.70%	Y	Y

Influential observations were identified and removed [59]. As noted below, models on different subsets of the data were constructed during our evaluation. The percentage of influential observations varied from less than 0.5% to 2.2% across these models.

### Unbalanced Dataset

Our dataset was unbalanced. This means that the proportion of observations with uniqueness less than 20% was quite small, and similarly for the proportion of observations with uniqueness less than 5%. Constructing regression models with an unbalanced dataset can result in poor model fit, inaccuracy in predicting the less prevalent class, and may even impede the convergence of the numeric maximum likelihood estimation algorithms.

There are three approaches for dealing with an unbalanced dataset: (a) a down-sampling or prior correction approach reduces the number of observations so that the two classes in the logistic regression model are equal, (b) the use of weights, and (c) an alternative correction which uses the full dataset and shown to be an improvement over weighting by King and Zeng (KZ) [60]. It has been noted that the weighting approach suffers a loss in efficiency compared to an unweighted approach when the model is exact [61], and the KZ method is shown to be better than using weights [60]. We therefore built models using two approaches and compared their results: (a) re-balancing using down-sampling, and adjusting the parameter estimates accordingly [60,62,63], and (b) the KZ method [60].

### Method for Model Evaluation

We compared both methods for dealing with the unbalanced dataset problem on three values: the area under the curve (AUC) of the Receiver Operating Characteristic curve [64,65], sensitivity, and specificity. The latter two metrics are defined more precisely in Figure 1 (the AUC is based on the definitions of specificity and sensitivity).

The AUC has an intuitive interpretation: it is the estimated probability that a randomly selected observation that is above the uniqueness threshold will have a higher predicted probability from the logistic regression model than a randomly selected observation that is below the uniqueness threshold [66,67]. Sensitivity is defined as the proportion of actually high risk records (above the threshold) which were correctly predicted as such. Specificity is defined as the proportion of actually low risk records (below or equal to the threshold) which were correctly predicted as such. For computing the above metrics, if the predicted probability on the 5% threshold model was greater than 0.5 then the FSA was deemed to have a uniqueness greater than 5%. A similar predicted probability cut-off was used for the 20% threshold model.

We used 10-fold cross-validation to generate the training and test datasets, which is a generally accepted practice to evaluate prediction models in the machine learning literature [68,69]. That is, we divided the dataset used to build the logistic regression model into deciles and used one decile in turn as the test dataset, and the remaining nine deciles to build (train) the model. In the context of ten-fold cross-validation, the down-sampling and KZ methods were performed separately on the nine training deciles each time a model was estimated. All the predictions across the 10-folds were then tabulated in a  $2 \times 2$  confusion matrix and the prediction accuracy was evaluated as illustrated in Figure 1. A confusion matrix shows the cross-tabulation of the number of observations predicted to be above/below the threshold vs. the number of observations that were actually above/below the threshold.

## Results

### Description of Canadian FSAs

Our models pertain to urban FSAs. We therefore provide a descriptive comparison of urban vs. rural FSAs in Canada.

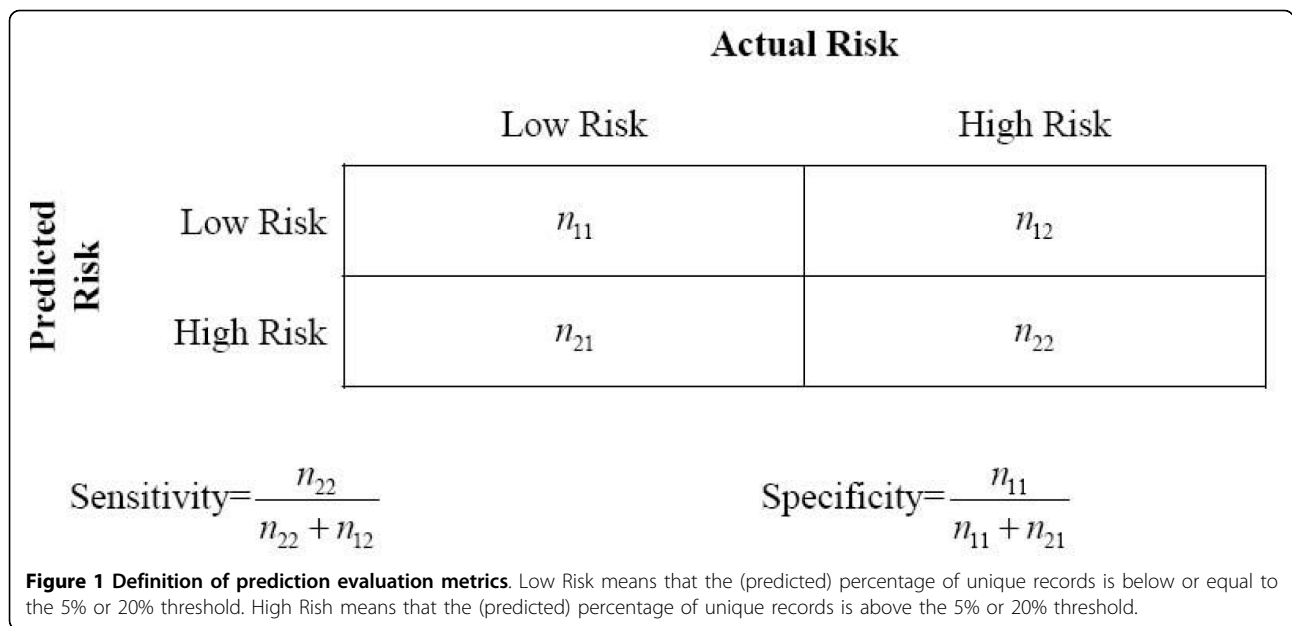
The population distribution for FSAs in the nine Canadian provinces is shown in Figure 2, and overall in Figure 3. Except for New Brunswick, rural FSAs tend to have more people living in them. The majority of the population lives in urban FSAs, except for Newfoundland, and to a lesser extent Saskatchewan, where the population is more evenly split between rural and urban FSAs. Table 4 shows the distribution of FSAs based on whether they are rural or urban. Even though they have smaller populations, the majority of FSAs are urban rather than rural. Figure 4 shows that in terms of physical size rural FSAs tend to have a considerably larger area than urban ones.

### Model Comparison

The two approaches for building the logistic regression models are compared in Table 5 for the two uniqueness thresholds. These results were obtained using 10-fold cross-validation. In terms of the AUC, the differences are very small and for practical purposes their predictive accuracies can be considered equivalent. The table also shows the sensitivity and specificity results using a predicted probability threshold of 0.5, which is consistent with the way that the models would be used in practice. Here we see that both modeling approaches had very similar specificity, but down-sampling had higher sensitivity for both uniqueness thresholds. Therefore, we will use the down-sampling model results in the rest of this paper.

### Model Results

Both models had a significant goodness of fit ( $p < 0.001$ ) [56]. The model parameters are shown in Table 6. All



model parameters are significant, including the interaction term.

### Discussion

#### Using the Models

In this paper we developed models to predict whether the population in a geographic area has uniqueness above the 5% and 20% thresholds using data from the Canadian census. We also demonstrated that the prediction models are sufficiently accurate to meet the risk and utility needs of data custodians and data recipients respectively. The areal unit that we studied was the urban FSA.

The logistic regression models can be used to determine whether or not the FSAs in actual datasets are too small. The *MaxCombs* value is computed based on the quasi-identifiers in the dataset. For each FSA, its population value can be determined from the Statistics Canada population tables. With these two values we can predict the probability that the percentage of uniques is above the 5% or 20% uniqueness thresholds. If the predicted probability is above 0.5, then disclosure control actions are necessary. For example, records in that FSA must be suppressed or combined with another FSA in the dataset. Alternatively, some variables may need to be removed or generalized to reduce the *MaxCombs* value.

Because the predictor variables in the models were centred and scaled, this also has to be done when using the models for actual prediction. Let the *MaxCombs* value for a particular dataset be denoted by  $M$ . We index the FSAs in a dataset by  $j$ . Let the population size for a particular FSA in the dataset be denoted by  $S_j$ .

We have the centered and scaled *MaxCombs* value:

$$M' = \frac{(M - 59861)}{10000} \tag{1}$$

and the centered and scaled population size value:

$$S'_j = \frac{(S_j - 21120)}{10000} \tag{2}$$

Then an FSA is considered to be high risk under the 5% threshold if the following condition is true:

$$\frac{1}{1 + e^{-\left(779.1 + 137.8M' - 37.3S'_j - 6.5M'S'_j\right)}} > 0.5 \tag{3}$$

and an FSA is considered to be high risk under the 20% threshold if the following condition is true:

$$\frac{1}{1 + e^{-\left(63.3 + 11.8M' - 6S'_j - M'S'_j\right)}} > 0.5 \tag{4}$$

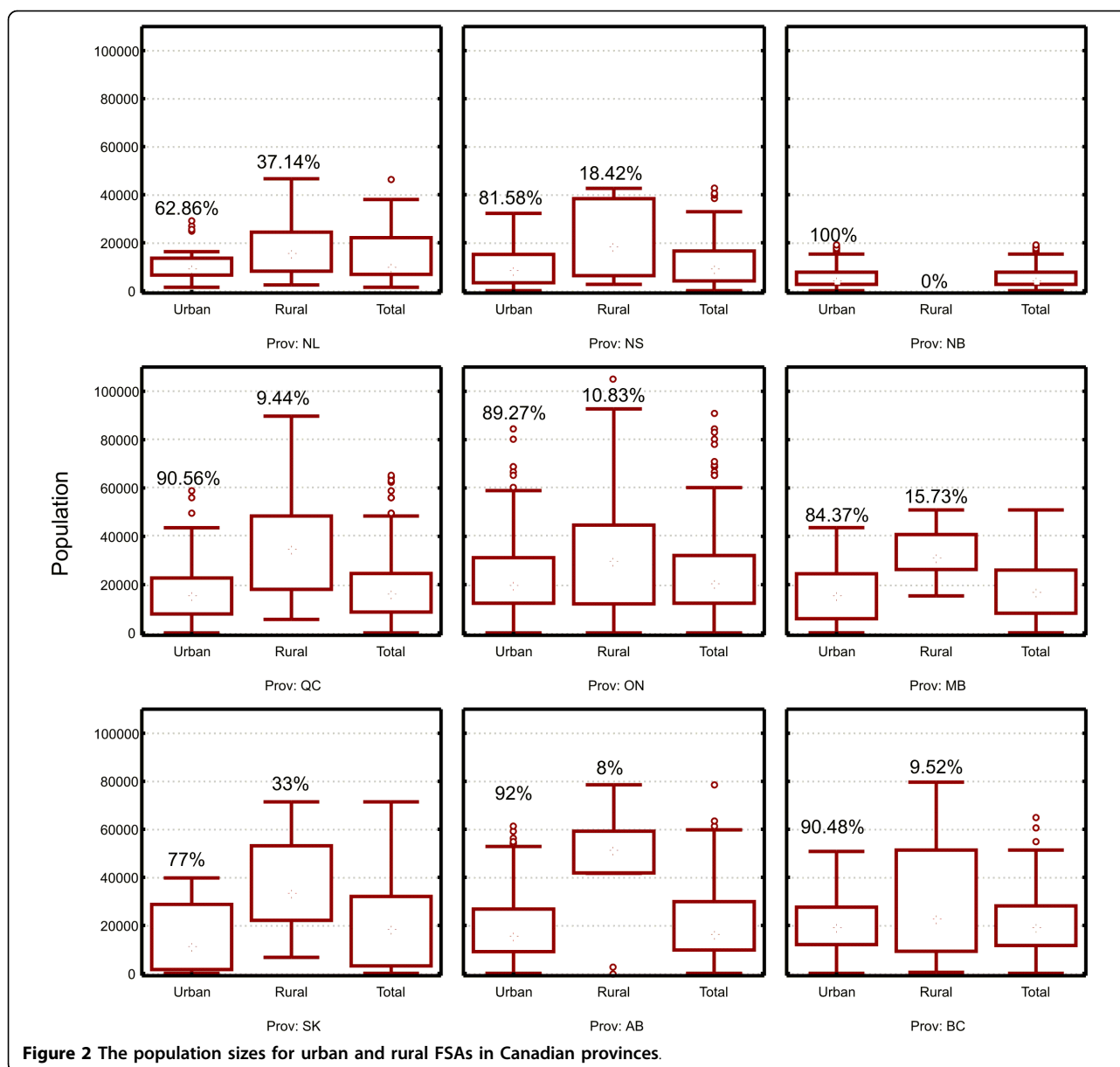
For the FSAs that are flagged through equations (3) or (4) then one should apply disclosure control actions.

#### Generalization of Models

There are two types of generalizations for these models: generalization to other quasi-identifiers and generalizations to other urban areal units apart from the FSA.

Our results indicate that *MaxCombs* is a very good predictor of uniqueness. The value of *MaxCombs* does not care what type of quasi-identifiers we have - it is only affected by the number of response categories in





**Figure 2** The population sizes for urban and rural FSAs in Canadian provinces.

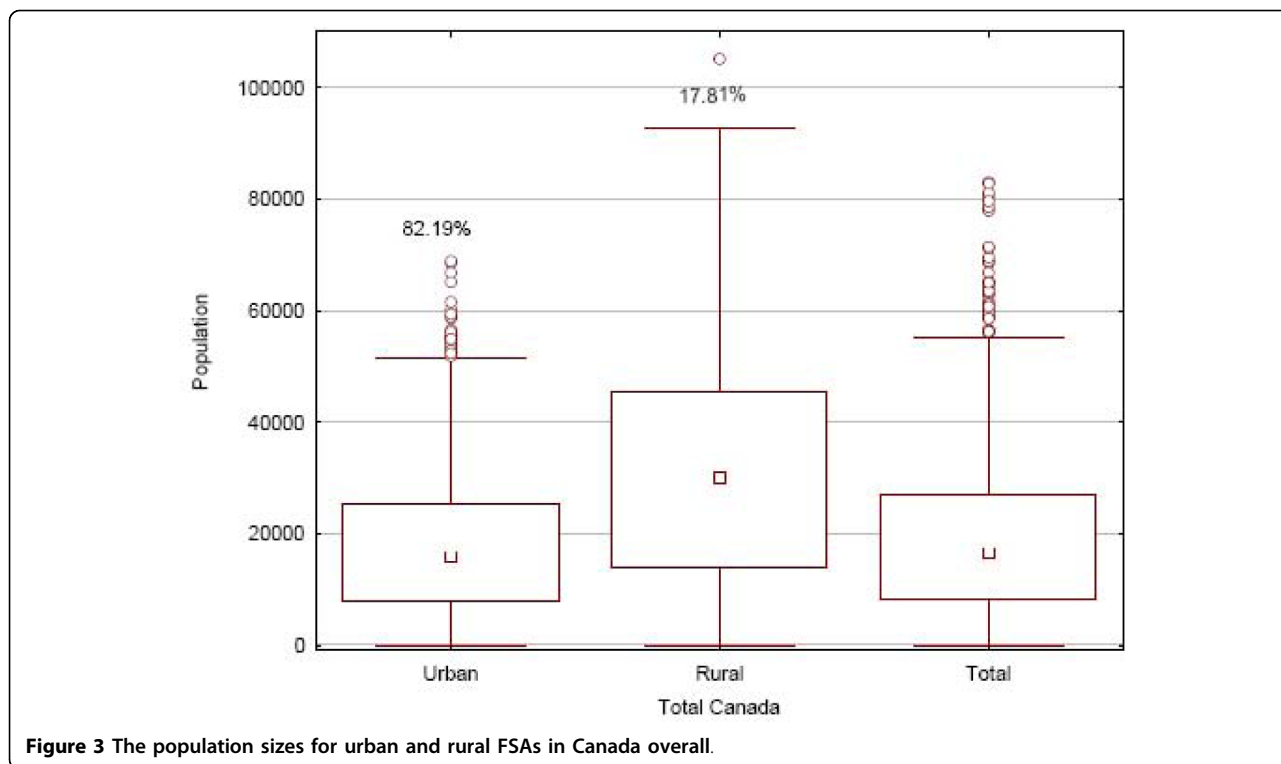
the quasi-identifiers. A previous study has shown that taking into account the distribution of the quasi-identifiers using an entropy metric did not result in any improvement in the prediction of uniqueness [37]. One explanation for this is that we have a ceiling effect: the prediction accuracy is quite high already that the addition of distribution information cannot make a significant improvement. Consequently, a strong case can be made that the models can be used with other demographic quasi-identifiers even if they are not explicitly represented in the census dataset, and if the *MaxCombs* is within the range used in our study.

Another question is whether there is a basis for generalizing the results to other urban areal units, for

example, full postal codes (which are subsets of FSAs) or regions (which are aggregates of FSAs)? Given that the prediction models are quite accurate using only the population size as a characteristic of the area, then there is no a priori reason not to be able to apply the models to other areas as long as their population sizes are within the range used for our models and that they are for urban Canadian areas.

#### Application of Models

We applied the models to evaluate whether the FSA sizes were appropriate on two data sets: the newborn registry of Ontario (Niday) and emergency department data from the children’s hospital in Ottawa. In this



**Table 4** Distribution of FSAs based on whether they are urban or rural.

Prov	Total Rural	Total Urban	Grand Total	% Rural	% Urban
AB	12	138	150	8.00%	92.00%
BC	18	171	189	9.52%	90.48%
MB	10	54	64	15.63%	84.38%
NB		110	110	0.00%	100.00%
NL	13	22	35	37.14%	62.86%
NS	14	62	76	18.42%	81.58%
ON	56	466	522	10.73%	89.27%
QC	39	374	413	9.44%	90.56%
SK	11	37	48	22.92%	77.08%
Grand Total	173	1434	1607	10.77%	89.23%

application we assume that the disclosure control action taken is the suppression of records in small FSAs.

The Niday registry captures information about all births in the province. We used a data extract for all births during 2005-2007 fiscal years. There were 164,272 usable records in the registry during that period. The quasi-identifiers that were considered were: mother's age, baby's month and year of birth, baby's gender, and the primary language spoken at home.

The proportion of records in the Niday registry that would have to be suppressed under each of the three

thresholds was computed. The results of this analysis are shown in Table 7. For example, under the 0% uniqueness threshold, 85% of the dataset would be in FSAs that are deemed too small. These small FSAs would have to be suppressed. As can be seen, there is a pronounced difference between using the 0% threshold and the others, with far less data having to be suppressed for the 5% and 20% thresholds. These results demonstrate that, where the risk profile is acceptably low, using a higher threshold can result in significantly more data being made available.

Using a similar approach, Table 7 also shows the results for the emergency department data for all presentations from 1<sup>st</sup> July 2008 to 1<sup>st</sup> June 2009, which consisted of 107,269 records. This data consists of date of presentation and the age of patient. With the 0% threshold 93% of the records would have to be suppressed, whereas only 54% would be suppressed for the 5% threshold, and none for the 20% threshold.

#### Selection of Threshold

An important decision when using the above models is selecting which of the three uniqueness threshold to use: 0%, 5%, or 20%. The most stringent uniqueness threshold of zero percent would be appropriate for datasets that are released to the public. This threshold would result in the most suppression and aggregation. The most permissive 20% threshold can be used when

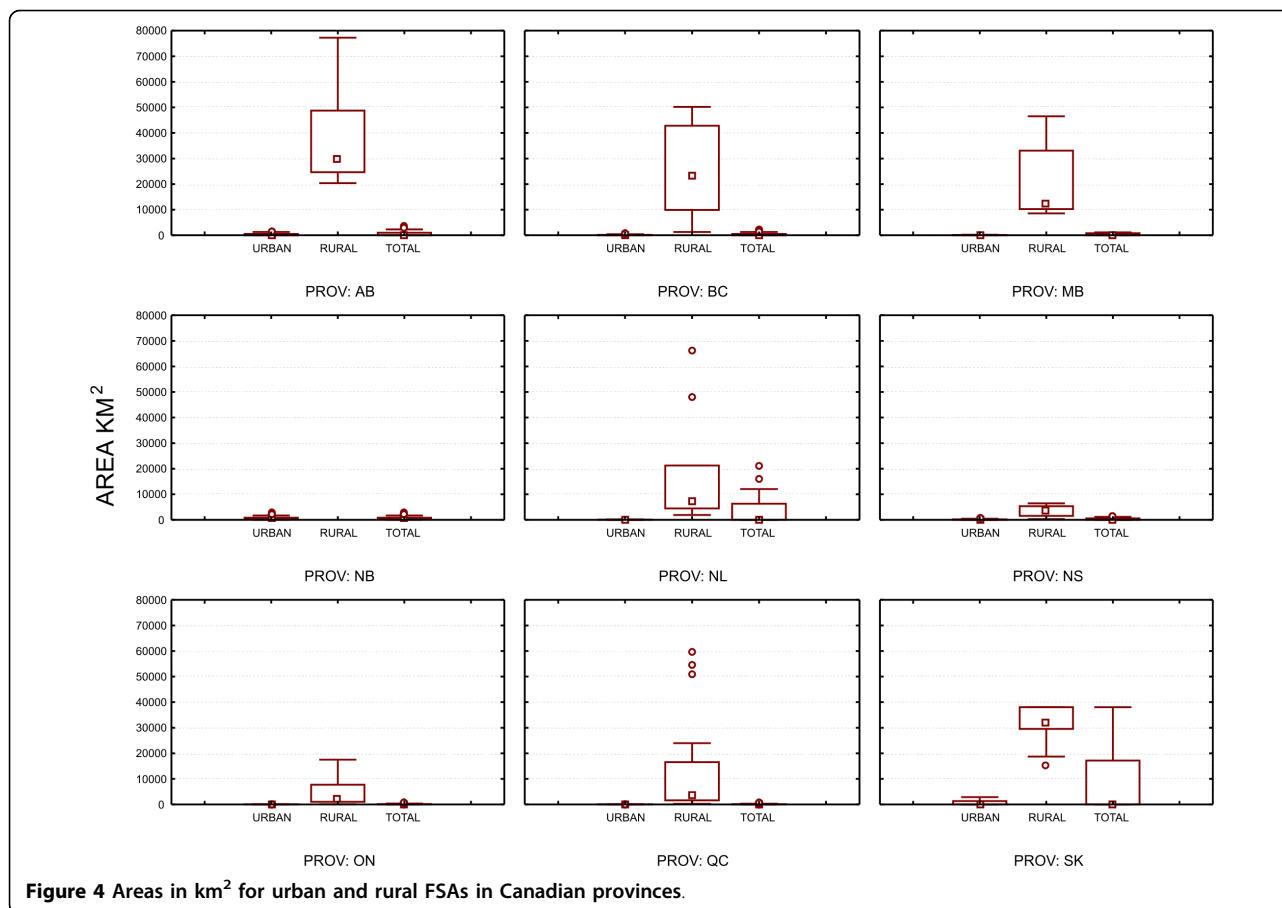


Figure 4 Areas in km<sup>2</sup> for urban and rural FSAs in Canadian provinces.

Table 5 Comparison of unbalanced data modeling methods.

Model Evaluation for the 5% Uniqueness Threshold			
	AUC	Sensitivity	Specificity
Down-Sampling	0.9849	0.87	0.996
KZ	0.9849	0.449	0.992
Model Evaluation for the 20% Uniqueness Threshold			
	AUC**	Sensitivity	Specificity
Down-Sampling	0.947	0.74	0.98
KZ	0.949	0.59	0.949

\*\*We tested the difference between the AUC values, and the difference was statistically significant between the two methods only for 20% uniqueness at an alpha level of 0.05

disclosing data to trusted recipients where the overall risks are quite low. This larger threshold would result in the least suppression and aggregation.

To assist with deciding which of the thresholds is most appropriate under a broad set of conditions, three general criteria have been proposed in the context of secondary use [70-72]:

- Mitigating controls that are in place at the data recipient's organization.

Table 6 Logistic regression model results for the 5% and 20% thresholds using down-sampling.

Logistic Regression Model for 5% Threshold				
	Intercept	POP	MaxCombs	POP × MaxCombs
Coefficient	779.1	-37.35	137.8	-6.5
95% CI	(744, 815.5)	(-60.46, -13.72)	(131.6, 144.2)	(-10.61, -2.36)
p-value	<0.0001	<0.0017	<0.001	0.0019
Logistic Regression Model for 20% Threshold				
	Intercept	POP	MaxCombs	POP × MaxCombs
Coefficient	63.3	-6	11.8	-1
95% CI	(61.85, 64.74)	(-6.83, -5.16)	(11.59, 12.1)	(-1.16, -0.86)
p-value	<0.0001	<0.0001	<0.0001	<0.0001

Mitigating controls evaluate the extent to which the data recipient has good security and privacy practices in place. A recent checklist can be used for evaluating the extent to which mitigating controls have been implemented [73]. The fewer security and privacy practices that the data recipient has in place, the lower the threshold that should be used.

**Table 7 The percentage of Niday and emergency department records that would have to be suppressed because they are high risk for each of the uniqueness thresholds.**

	0% Threshold	5% Threshold	20% Threshold
Niday	85%	77%	0%
Emergency Dept.	93%	54%	0%

- The extent to which a disclosure (inadvertent or otherwise) constitutes an invasion of privacy for the patients.

Additional file 2 contains a set of items that have been developed based on the literature to evaluate the invasion-of-privacy construct [74-79]. This set of items was subsequently reviewed by a panel of 12 Canadian privacy experts for completeness, redundancy, and clarity. The greater the risk of an invasion of privacy, the lower the threshold that should be used.

- The extent to which the data recipient is motivated and capable of re-identifying the data.

Additional file 2 contains a set of items that have been developed based on the literature to evaluate the motives and capacity construct [80-83]. This construct captures the fact that some data recipients can be trusted more than others (e.g., researchers vs. making data available to the general public). The set of items was subsequently reviewed by a panel of 12 Canadian privacy experts for completeness, redundancy, and clarity. The greater the risk that the data recipient is motivated and has the capacity to re-identify the database, the lower the threshold that should be used.

Admittedly, the use of these checklists remains qualitative, but they do provide a starting point for deciding what an appropriate threshold should be.

**Limitations**

The FSAs that were included in our analysis were from urban areas in Canada. As described in Additional file 1, the reason is that the census tract information from the census file that we used is only defined for urban areas. Therefore, FSAs from rural areas were not covered. However, it should be noted that the majority of the Canadian population lives in urban areas.

Our analysis was based on data from the 2001 census. There will be changes in the population over time and therefore the models may not be an accurate reflection of uniqueness the further from 2001 we are. Future studies should replicate this research on subsequent census data (the 2006 census data was not available in the Statistics Canada RDC when we conducted this study).

We used the estimated uniqueness values as the correct values, and validated our prediction model on that basis. However, the uniqueness estimate will not be perfect and such errors will negatively affect the overall accuracy of the 5% and 20% prediction models.

The *MaxCombs* value can only be computed for quasi-identifiers with a finite number of response categories. Continuous variables that are not discretized cannot be sensibly captured using our approach.

**Conclusions**

Disclosure control practices for small geographic areas often result in health datasets that have significantly reduced utility. These practices include the suppression of records from individuals in small geographic areas, the aggregation of small geographic areas into larger ones, suppression of the non-geographic variables, or generalization of the non-geographic variables. Previous work has used a rather stringent definition of a small geographic area: when it has no unique individuals on the potentially identifying variables (quasi-identifiers). However, less stringent thresholds have been used in the past for the disclosure of health datasets: 5% uniqueness and 20% uniqueness.

In this paper we develop models to determine whether urban FSAs in Canada are too small by the 5% and 20% criteria by analyzing 2001 census data. We have also provided a set of concrete guidelines to help custodians decide which one these thresholds to use. Within this framework, a data custodian can manage the amount of geographic suppression or aggregation in proportion to the risks of disclosing a particular dataset.

**Additional file 1: Mapping census geography to postal geography using a gridding methodology.** Describes the methodology we used to assign a postal code to each record in the census file.

**Additional file 2: Evaluating dimensions of risk.** Presents the validated checklists for evaluating the "invasion of privacy" and "motives and capacity" dimensions of disclosure risk.

**Acknowledgements**

This work was funded by the GeoConnections program of Natural Resources Canada, the Public Health Agency of Canada, the Ontario Centers of Excellence, and the Natural Sciences and Engineering Research Council of Canada. We wish to thank David Paton (Canadian Institute for Health Information) for his feedback on an earlier version of this paper. We also wish to thank our panel of privacy experts for reviewing the items we used to evaluate risk described in Additional file 2. This study was approved by the research ethics board of the Children's Hospital of Eastern Ontario Research Institute.

**Author details**

<sup>1</sup>Children's Hospital of Eastern Ontario Research Institute, 401 Smyth Road, Ottawa, Ontario K1J 8L1, Canada. <sup>2</sup>Pediatrics, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada. <sup>3</sup>GIS Infrastructure, Office of Public Health Practice, Public Health Agency of Canada, Ottawa, Ontario K1A 0K9, Canada. <sup>4</sup>Ottawa Hospital Research Institute, Ottawa, Ontario, Canada.

<sup>5</sup>Children's Hospital of Eastern Ontario, 401 Smyth Road, Ottawa, Ontario K1J 8L1, Canada.

#### Authors' contributions

KEE designed the study, directed the data analysis, and contributed to writing the paper. AB performed the Statistics Canada RDC statistical analysis and contributed to writing the paper. PA performed the geospatial data analysis and contributed to writing the paper. AN performed the model building analysis work. MW contributed to the application of the results. JB contributed to the application of the results. TR contributed to the application of the results. All of the authors have read and approved the final manuscript.

#### Competing interests

The authors declare that they have no competing interests.

Received: 28 May 2009 Accepted: 2 April 2010 Published: 2 April 2010

#### References

1. Safran C, Bloomrosen M, Hammond E, Labkoff S, S K-F, Tang P, Detmer D: **Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper.** *Journal of the American Medical Informatics Association* 2007, **14**:1-9.
2. Roy D, Fournier F: **Secondary use of personal information held on national electronic health record systems.** Centre for Bioethics, Clinical Research Institute of Montreal (study commissioned by the Office of the Privacy Commissioner of Canada) 2007.
3. Kosseim P, Brady M: **Policy by procrastination: Secondary use of electronic health records for health research purposes.** *McGill Journal of Law and Health* 2008, **2**:5-45.
4. Black C, McGrail K, Fooks C, Baranek P, Maslove L: **Data, Data, Everywhere – Improving access to population health and health services research data in Canada.** Centre for Health Services and Policy Research and Canadian Policy Research Networks 2005.
5. Willison D, Gibson E, McGrail K: **A roadmap to research uses of electronic health information.** *CIHR Health Information Summit: 20-21 October 2008, Toronto* .
6. PWC Healthcare: *Transforming healthcare through secondary use of health data* Dallas: PriceWaterhouseCoopers 2009.
7. Boulos M: **Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom.** *International Journal of Health Geographics* 2004, **3**(1).
8. O'Dwyer LA, Burton DL: **Potential meets reality: GIS and public health research in Australia.** *Australian and New Zealand Journal of Public Health* 1998, **22**(7):819-823.
9. Ricketts TC: **Geographic information systems and public health.** *Annual Review of Public Health* 2003, **24**:1-6.
10. Cromley EK: **GIS and Disease.** *Annual Review of Public Health* 2003, **24**:7-24.
11. Brindley P, Maheswaran R: **My favourite software: geographic information systems.** *Journal of Public Health Medicine* 2002, **24**(2):149.
12. Richards TB, Croner CM, Rushton G, Brown CK, Fowler L: **Geographic information systems and public health: mapping the future.** *Public Health Reports* 1999, **114**:359-373.
13. Ricketts T: **Geographic information systems and public health.** *Annual Review of Public Health* 2003, **24**:1-6.
14. McLafferty S: **GIS and health care.** *Annual Review of Public Health* 2003, **24**:25-42.
15. Cromley E: **GIS and disease.** *Annual Review of Public Health* 2003, **24**:7-24.
16. Muggge R: **Issues in protecting confidentiality in national health statistics.** *Proceedings of the Social Statistics Section, American Statistical Association* 1983, 592-594.
17. Mackie C, Bradburn N: *Improving access to and confidentiality of research data: Report of a workshop* Washington: The National Academies Press 2000.
18. Croner C: **Public health, GIS, and the Internet.** *Annual Review of Public Health* 2003, **24**:57-82.
19. Gibson Justice: *Mike Gordon and The Minister of Health and Privacy Commissioner of Canada* Federal Court of Canada 2008.
20. El Emam K, Kosseim P: **Privacy Interests in Prescription Records, Part 2: Patient Privacy.** *IEEE Security and Privacy* 2009, **7**(2):75-78.
21. Hawala S: **Enhancing the "100,000" rule: On the variation of percent of uniques in a microdata sample and the geographic area size identified on the file.** *Proceedings of the Annual Meeting of the American Statistical Association: 5-9 August 2001, St. Louis*.
22. Greenberg B, Voshell L: **Relating risk of disclosure for microdata and geographic area size.** *Proceedings of the Section on Survey Research Methods, American Statistical Association* 1990, 450-455.
23. Greenberg B, Voshell L: *The geographic component of disclosure risk for microdata.* *Statistical Research Division Report Series* Washington: Bureau of the Census 1990.
24. Zayatz L, Massell P, Steel P: **Disclosure limitation practices and research at the US Census Bureau.** *Netherlands Official Statistics* 1999, **14**(Spring):26-29.
25. Zayatz L: *Disclosure avoidance practices and research at the US Census Bureau: An update.* *Statistical Research Division Report Series* Washington: US Census Bureau 2005.
26. Hawala S: **Microdata disclosure protection research and experiences at the US census bureau.** *Presented at the Workshop on Microdata: 21-22 August 2003, Stockholm* .
27. Marsh C, Dale A, Skinner C: **Safe data versus safe settings: Access to microdata from the British census.** *International Statistical Review* 1994, **62**(1):35-53.
28. Statistics Canada: *Canadian Community Health Survey (CCHS) Cycle 3.1 (2005) Public Use Microdata File (PUMF) User Guide* 2006.
29. Willenborg L, de Waal T: *Statistical Disclosure Control in Practice* New York: Springer-Verlag 1996.
30. Fefferman N, O'Neil E, Naumova E: **Confidentiality and confidence: Is data aggregation a means to achieve both?** *Journal of Public Health Policy* 2005, **26**(4):430-449.
31. Willenborg L, Mokken R, Pannekoek J: **Microdata and disclosure risks.** *Proceedings of the Annual Research Conference of US Bureau of the Census* 1990, 167-180.
32. Olson K, Grannis S, Mandl K: **Privacy protection versus cluster detection in spatial epidemiology.** *American Journal of Public Health* 2006, **96**(11):2002-2008.
33. Marceau D: **The scale issue in social and natural sciences.** *Canadian Journal of Remote Sensing* 1999, **25**(4):347-356.
34. Bivand R: *A review of spatial statistical techniques for location studies* Bergen: Norwegian School of Economics and Business Administration 1998.
35. Ratcliffe J: **The Modifiable Areal Unit Problem.** [http://www.jratcliffe.net/research/maup.htm].
36. Bethlehem J, Keller W, Pannekoek J: **Disclosure control of microdata.** *Journal of the American Statistical Association* 1990, **85**(409):38-45.
37. El Emam K, Brown A, Abdelmalik P: **Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk.** *Journal of the American Medical Informatics Association* 2009, **16**(2):256-266.
38. Howe H, Lake A, Shen T: **Method to assess identifiability in electronic data files.** *American Journal of Epidemiology* 2007, **165**(5):597-601.
39. Howe H, Lake A, Lehnher M, Roney D: **Unique record identification on public use files as tested on the 1994-1998 CINA analytic file.** *North American Association of Central Cancer Registries* 2002.
40. El Emam K: **Heuristics for de-identifying health data.** *IEEE Security and Privacy* 2008, 72-75.
41. Dalenius T: **Finding a needle in a haystack or identifying anonymous census records.** *Journal of Official Statistics* 1986, **2**(3):329-336.
42. El Emam K, Jabbouri S, Sams S, Drouet Y, Power M: **Evaluating common de-identification heuristics for personal health information.** *Journal of Medical Internet Research* 2006, **8**(4):e28.
43. El Emam K, Jonker E, Sams S, Neri E, Neisa A, Gao T, Chowdhury S: *Pan-Canadian De-Identification Guidelines for Personal Health Information* Ottawa: Prepared for the Office of the Privacy Commissioner of Canada 2007.
44. The International Organization for Standardization: *ISO/TS 25237: Health Informatics - Pseudonymization* Geneva: The International Organization for Standardization 2008.
45. Bow C, Waters N, Faris P, Seidel J, Galbraith P, Knudtson M, Ghali W: **Accuracy of city postal code coordinates as a proxy for location of residence.** *International Journal of Health Geographics* 2004, **3**(5).
46. Ng E, Wilkins R, Perras A: **How far is it to the nearest hospital? Calculating distances using the Statistics Canada Postal Code Conversion file.** *Health Reports* 1993, **5**:179-183.
47. Mackillop W, Zhang-Salmons J, Groome P, Pazat L, Holowaty E: **Socioeconomic status and cancer survival in Ontario.** *Journal of Clinical Oncology* 1997, **15**:1680-1689.

48. Spasoff A, Gilkes D: **Up-to-date denominators: Evaluation of taxation family for public health planning.** *Canadian Journal of Public Health* 1994, **85**:413-417.
49. Demissie K, Hanley J, Menzies D, Joseph L, Ernst P: **Agreement in measuring socio-economic status: Area-based versus individual measures.** *Chronic Diseases in Canada* 2000, **21**:1-7.
50. Guernsey J, Dewar R, Weerasinghe S, Kirkland S, Veugelers P: **Incidence of cancer in sydney and Cape breton County, Nova Scotia 1979-1997.** *Canadian Journal of Public Health* 2000, **91**:285-292.
51. Sweeney L: **Uniqueness of Simple Demographics in the US Population.** Carnegie Mellon University, Laboratory for International Data Privacy 2000.
52. Golle P: **Revisiting the uniqueness of simple demographics in the US population.** *Workshop on Privacy in the Electronic Society* 2006.
53. Skinner C, Holmes D: **Estimating the re-identification risk per record in microdata.** *Journal of Official Statistics* 1998, **14**(4):361-372.
54. Chen G, Keller-McNulty S: **Estimation of identification disclosure risk in microdata.** *Journal of Official Statistics* 1998, **14**(1):79-95.
55. Zayatz L: **Estimation of the percent of unique population elements on a microdata file using the sample** Washington: US Bureau of the Census 1991.
56. Hosmer D, Lemeshow S: *Applied Logistic Regression* New York: John Wiley & Sons 1989.
57. Jaccard J: *Interaction Effects in Logistic Regression* London: Sage Publications 2001.
58. Simon S, Lesage J: **The Impact of Collinearity Involving the Intercept Term on the Numerical Accuracy of Regression.** *Computer Science in Economics and Management* 1988, **1**:137-152.
59. Pergibon D: **Logistic Regression Diagnostics.** *The Annals of Statistics* 1981, **9**(4):705-724.
60. King G, Zeng L: **Logistic regression in rare events data.** *Political Analysis* 2001, **9**(2):137-163.
61. Scott A, Wild C: **Fitting logistic models under case-control or choice based sampling.** *Journal of the Royal Statistical Society* 1986, **48**(2):170-182.
62. Lowe W: **Rare events research.** *Encyclopedia of Social Measurement* Cambridge: Academic Press Kempf-Leonard K 2005, 293-297.
63. Ruiz-Gazen A, Villa N: **Storms prediction: Logistic regression vs. random forests for unbalanced data.** *Case Studies in Business, Industry and Government Statistics* 2007, **1**(2):91-101.
64. Metz C: **Basic Principles of ROC Analysis.** *Seminars in Nuclear Medicine* 1978, **VIII**(4):283-298.
65. DeLong E, DeLong D, Clarke-Pearson D: **Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach.** *Biometrics* 1988, **44**:837-845.
66. Hanley J, McNeil B: **The Meaning and Use of the Area Under a Receiver Operating Characteristic Curve.** *Diagnostic Radiology* 1982, **143**(1):29-36.
67. Spiegelhalter D: **Probabilistic Prediction in Patient Management in Clinical Trials.** *Statistics in Medicine* 1986, **5**:421-433.
68. Cherkassky V, Muller F: *Learning from data: concepts, theory, and methods* New York: Wiley 1998.
69. Alpaydin E: *Introduction to machine learning* Cambridge: MIT Press 2004.
70. El Emam K: *De-identifying health data for secondary use: A framework* Ottawa: CHEO Research Institute 2008.
71. Jabine T: **Statistical disclosure limitation practices of United States statistical agencies.** *Journal of Official Statistics* 1993, **9**(2):127-454.
72. Jabine T: **Procedures for restricted data access.** *Journal of Official Statistics* 1993, **9**(2):537-589.
73. El Emam K, Dankar F, Vaillancourt R, Roffey T, Lysyk M: **Evaluating patient re-identification risk from hospital prescription records.** *Canadian Journal of Hospital Pharmacy* 2009, **62**(4):307-319.
74. Treasury Board of Canada Secretariat: **Privacy impact assessment guidelines: A framework to manage privacy risks.** Government of Canada 2002.
75. Treasury Board of Canada Secretariat: **Guidance document: Taking privacy into account before making contracting decisions.** Government of Canada 2006.
76. Canadian Institutes of Health Research: *CIHR best practices for protecting privacy in health research* Ottawa: Public Works and Government Services Canada 2005.
77. Canadian Institutes of Health Research: *Secondary use of personal information in health research: Case studies* Ottawa: Public Works and Government Services Canada 2002.
78. Office of the Privacy Commissioner of Canada: **Key Steps for Organizations in Responding to Privacy Breaches.** 2007 [[http://www.priv.gc.ca/information/guide/2007/gl\\_070801\\_02\\_e.pdf](http://www.priv.gc.ca/information/guide/2007/gl_070801_02_e.pdf)].
79. Office of the Saskatchewan Information and Privacy Commissioner: **Privacy breach guidelines.** [[http://www.oipc.sk.ca/Resources/Privacy%20Breach%20Guidelines1%20\(3\).pdf](http://www.oipc.sk.ca/Resources/Privacy%20Breach%20Guidelines1%20(3).pdf)].
80. Elliot M, Dale A: **Scenarios of attack: the data intruder's perspective on statistical disclosure risk.** *Netherlands Official Statistics* 1999, **14**(Spring):6-10.
81. Sweeney L: **Guaranteeing anonymity when sharing medical data: The Datafly system.** *Proceedings of the American Medical Informatics Association Symposium, 25-29 October 1997; Nashville. JAMIA 1997, Symposium Suppl:* 51-55.
82. Willenborg L, de Waal T: *Elements of Statistical Disclosure Control* New York: Springer-Verlag 2001.
83. Pong R, Pitblado J: **Don't take geography for granted ! Some methodological issues in measuring geographic distribution of physicians.** *Canadian Journal of Rural Medicine* 2001, **6**(2):103-112.

#### Pre-publication history

The pre-publication history for this paper can be accessed here:  
[<http://www.biomedcentral.com/1472-6947/10/18/prepub>]

doi:10.1186/1472-6947-10-18

**Cite this article as:** El Emam et al.: A method for managing re-identification risk from small geographic areas in Canada. *BMC Medical Informatics and Decision Making* 2010 **10**:18.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

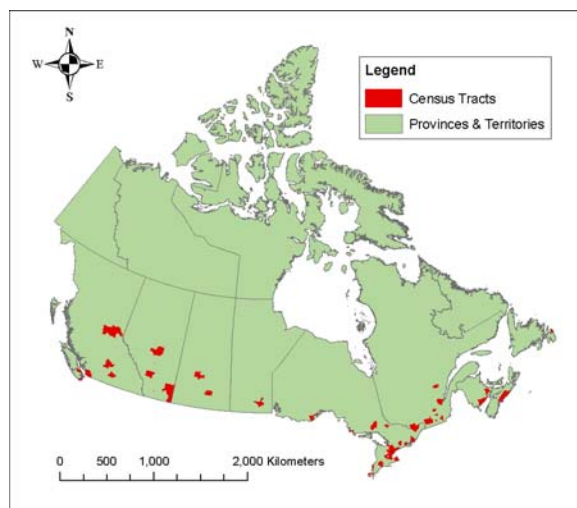
Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



## Additional File 1: Mapping Census Geography to Postal Geography Using a Gridding Methodology

### **Background**

The smallest geographic unit provided in the census microdata file available through Statistics Canada's Research Data Centre (RDC) is the census tract (CT). CTs are only defined for census metropolitan areas and census agglomerations with urban core populations of at least 50,000 individuals. They are defined by Statistics Canada as "...small, relatively stable geographic areas that usually have a population of 2,500 to 8,000." [1]. The 2001 census contained a total of 4,798 CTs distributed over 9 provinces (no CTs are defined for the Territories or the province of PEI; see Figure 1).



**Figure 1:** Distribution of 2001 census tracts across Canada.

In order to compute re-identification risk by Forward Sortation Area (FSA) in our current study, we needed to devise a method to estimate conversion between census and postal geography. A gridding methodology similar in nature to the Gridded Population of the World Project (GPW) [2] at the Center for International Earth Science Information Network at Columbia University [3] was utilized, allowing assignment of geography based on areal weighting using a population grid for Canada.

### **Methods**

Population-based weights were assigned to CT-FSA unions based on a created population grid for all of Canada. The grid cell size was one kilometre by one kilometre, and assigned populations were based on the 2001 census profile at the Dissemination Area level (DA). This is the smallest geography at which census profile information is released by Statistics Canada [4]. Similar to the PCCF+ [5-8], these population weights were then used to randomly assign census tracts to their associated FSAs. Details of the steps taken to create the population grid are described below.

Twenty six (26) complete grids of dimensions 1554 by 546 Kilometres were created using a script in ESRI's ArcMap 9.2 [9], as specified in Table 1. This created 848,484 one kilometre square cells per grid, for a total of 22,909,068 cells covering the Canadian landmass.

Once the grids were created, the next task was to assign an estimated population to each cell. This was done using the Statistics Canada DA file [10]. First, all DA polygons identified as water were removed. A new DA shape file containing only land DAs was created. DA boundaries were then dissolved so that DAs with disparate polygons were captured within one record. Areas and perimeters were summed for each polygon to give the total DA area and perimeter. This reduced the number of records from 62,015 to 52,924, which matches the number of DAs as reported by Statistics Canada. Total population, as well as sex and age-stratified populations were extracted for all DAs across Canada, using four separate profile files (Western Canada and the Territories, Ontario, Quebec, and Atlantic Canada). Next, the 2001 DA population file was joined with the 2001 DA boundary file, to create a 2001 Canada DA boundary file containing total and sex and age stratified populations.

A "Select by attributes" function where population was not zero (0) was completed on the above file to create a new boundary file containing only DA polygons with reported populations. This further reduced the number of records to 49,153, creating a boundary file for non-water, populated DAs only. A "Select by location" function was completed on all 26 grids, for any cells that intersected the boundary file from the previous function. The resultant grids had a combined total cell count of 2,367,457.

A model was created using the ArcGIS model builder, and run for each of the 26 grids, to create grid section intersects with the 2001 DAs, FSAs and CTs. The model also calculated proportional grid subsection areas and the corresponding population, based on underlying DA population and an assumption that the population was distributed proportionally to area within each of the geographic areas.

A summary was done by each CT-FSA combination, to create unique CT-FSA records with the corresponding sum of the calculated grid-section populations. These summed populations were then divided by the total sum of the gridded-CT population to give the proportion of the population in each CT that lay within the corresponding FSA. In essence, this creates a population-based weight for each CT-FSA combination, allowing us to randomly assign any given record within a CT to its most likely (population-weighted) FSA.

A simplified hypothetical example of the end result is given in Table 2 and Figure 2. In this example, 64.07% of the population in CT16003 is found in FSA K2S, and 35.93% in FSA K2T. For CT 16004, 49.35% of its population is in K2R, 19.48% in K2S and 31.17% in K2T. This reduces the table to five rows, with a population-based weight for each unique CT-FSA combination. If, for example, there were then 28 records from the microdata file falling in CT 16003, 18 (~65.86%) would be allocated to K2S, and 10 (~34.14%) to K2T.



<b>Grid Section</b>	<b>x</b>	<b>y</b>	<b>rows</b>	<b>columns</b>	<b># Cells</b>	<b># Cells (DA-clipped)</b>	<b># Cells (populated DA-clipped)</b>
00	-2341699	310266	1554	546	848,484	147,282	95,225
01	-1795699	310266	1554	546	848,484	323,759	292,052
02	-1249699	310266	1554	546	848,484	400,335	352,048
03	-703699	310266	1554	546	848,484	421,104	252,417
04	-157699	310266	1554	546	848,484	442,583	112,863
05	388301	310266	1554	546	848,484	444,187	47,006
06	934301	310266	1554	546	848,484	588,000	220,587
07	1480301	310266	1554	546	848,484	514,762	202,006
08	2026301	310266	1554	546	848,484	222,848	139,035
09	2572301	310266	1554	546	848,484	79,825	30,635
10	-2341699	1864266	1554	546	848,484	490,304	181,644
11	-1795699	1864266	1554	546	848,484	843,129	253,796
12	-1249699	1864266	1554	546	848,484	753,391	84,386
13	-703699	1864266	1554	546	848,484	749,156	802
14	-157699	1864266	1554	546	848,484	563,822	1,239
15	388301	1864266	1554	546	848,484	192,569	1,005
16	934301	1864266	1554	546	848,484	587,718	1,420
17	1480301	1864266	1554	546	848,484	342,289	683
18	2026301	1864266	1554	546	848,484	220,305	48,694
19	2572301	1864266	1554	546	848,484	55,829	25,720
20	-2341699	3418266	1554	546	848,484	21,506	0
21	-1795699	3418266	1554	546	848,484	168,942	531
22	-1249699	3418266	1554	546	848,484	135,498	686
23	-703699	3418266	1554	546	848,484	229,560	0
24	-157699	3418266	1554	546	848,484	424,214	1,101
25	388301	3418266	1554	546	848,484	258,726	210
26	934301	3418266	1554	546	848,484	26,188	160
<b>TOTAL</b>					<b>22,909,068</b>	<b>9,647,831</b>	<b>2,345,951</b>

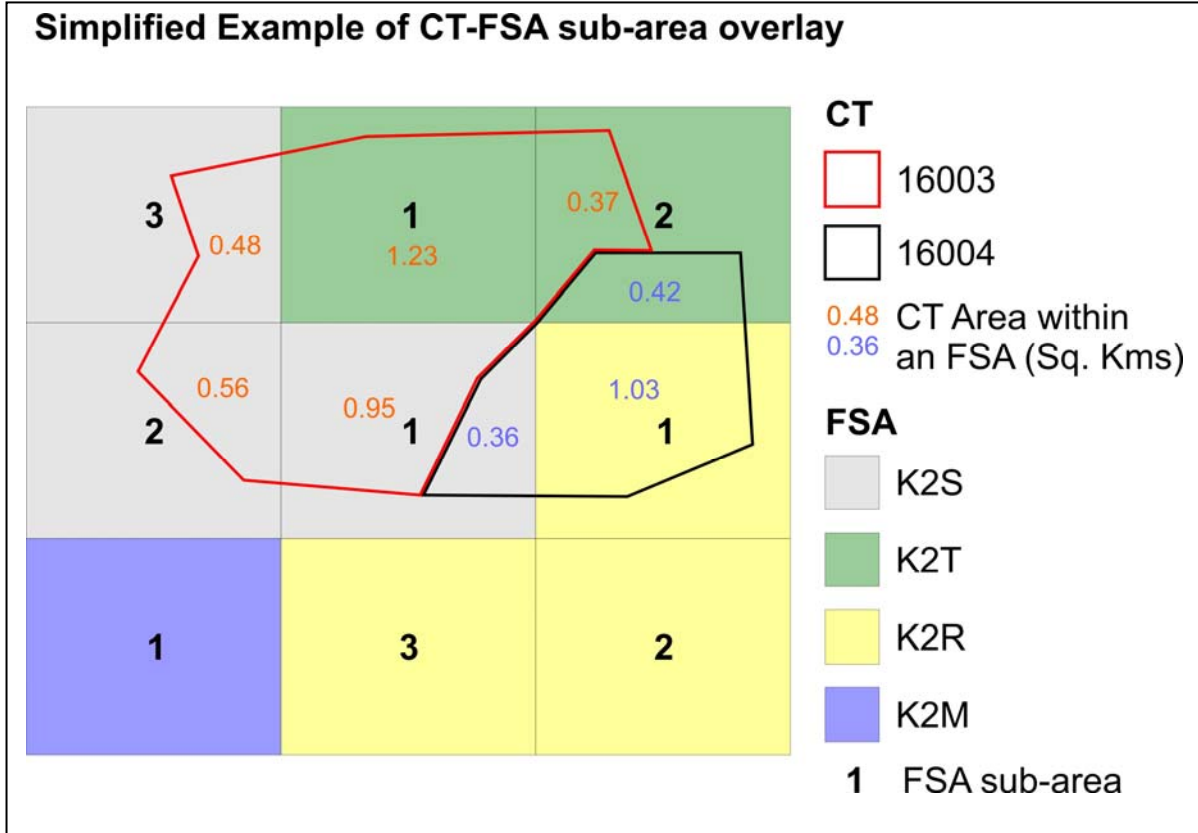
**Table 1:** Canadian grid development table.

CT	FSAsa	FSAsa Pop Density (per Sq. Km.)	CT Area in FSA (Sq. Km.)	Pop	CT Pop	Weight
16003	K2S-1	50	0.95	48	128	0.3750
16003	K2S-2	25	0.56	14	128	0.1094
16003	K2S-3	42	0.48	20	128	0.1563
16003	K2T-1	20	1.23	25	128	0.1953
16003	K2T-2	56	0.37	21	128	0.1641
16004	K2R-1	37	1.03	38	77	0.4935
16004	K2S-1	42	0.36	15	77	0.1948
16004	K2T-2	56	0.42	24	77	0.3117

FSAsa = FSA sub-area

Pop = Population

**Table 2:** Simplified hypothetical example of the weighted association between CTs and FSAs.



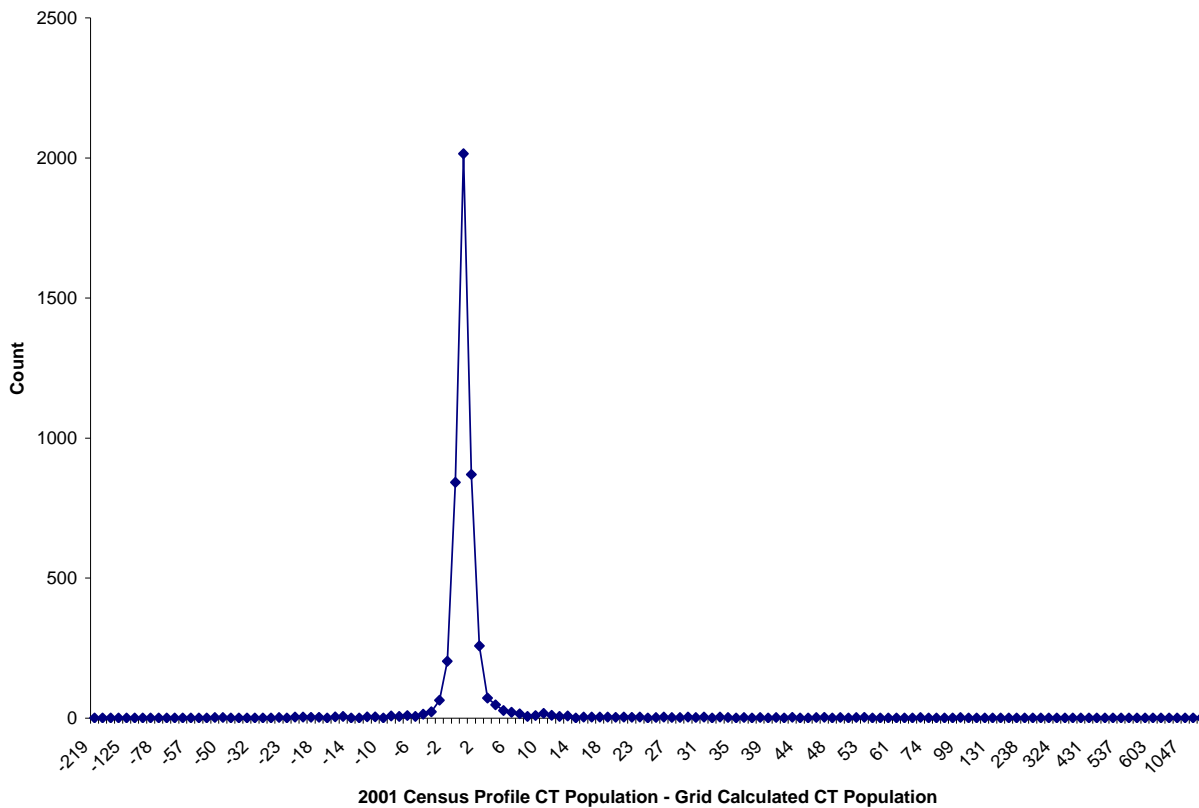
**Figure 2:** Example CT-FSA sub-area overlay to illustrate the hypothetical example.

### Results

The CT population assignments based on the gridding methodology proved to be very similar to the 2001 Statistics Canada Census Tract population profile (Table 3). The mean difference between the populations was 3.45 individuals, with a standard deviation of 48.96 individuals (median was 0). A graphical representation of the distribution of the population differences, by census tracts, is given in Figure 3.

	2001 Statistics Canada Population Profile Census Tract	Canada Population Grid Project Census Tract
Total n	4757	4757
Mean population	4413.99	4410.54
Standard Deviation	1911.77	1911.33
Minimum population	40	0
Median population	4290	4287
Maximum population	20635	20636

**Table 3:** Census tract population comparison between created population grid and 2001 census profile.



**Figure 3:** Distribution of Census Tract Population Difference between Grid-Calculated Population and 2001 Census Profile.

Provincial analyses also showed a high concordance between the CT populations using the gridding methodology as compared to the 2001 Statistics Canada Census Tract population profile (Table 4). The greatest differences were in New Brunswick (mean difference = 6.97 individuals, standard deviation = 75.26 individuals) and Alberta (mean difference = 6.75 individuals, standard deviation = 81.67 individuals).

	NL	NS	NB	QC	ON	MB	SK	AB	BC
N	45	85	70	1246	2001	164	101	449	596
Mean	3.71	2.6	6.97	1.55	3.68	2.93	-1.18	6.75	4.79
Std Dev	12.01	19.37	75.26	26.19	51.38	27.14	37.86	81.67	51.38
Median	0	0	0	0	0	0	0	0	0

**Table 4:** Provincial differences between Profile and grid CT populations.

## Conclusions

The population grid created in this study provides a means for linking census geography to postal geography in Canada. While creating population grids in and of itself is not a novel idea, the created grid in this project allows the mapping of census geography to postal geography, based on population weights. The procedure assumes a uniform population distribution within the geography being used. However, since CTs only occur in highly populated urban areas, this was felt to be an appropriate assumption. A similar assumption would not hold in rural or less densely populated areas, and this technique would therefore not be appropriate. However, it could be utilized, and further refined, by incorporating additional information, such as ecumene areas, satellite imagery for residential and inhabited areas, address data, etc.

## References

1. Statistics Canada. *Cartographic boundary files: 2001 census*. 2002.
2. Yetman G, Deichmann U, Balk D. *Creating a global grid of human population*. Available from: <http://gis.esri.com/library/userconf/proc00/professional/papers/PAP552/p552.htm>. Archived at: <http://www.webcitation.org/5jxHZXuVB>.
3. *Center for International Earth Science Information Network (CIESIN)*. Available from: <http://beta.sedac.ciesin.columbia.edu>.
4. Statistics Canada. *Profile of all levels of geography in Canada, 2001 census*. 2003.
5. Wilkins R. *Use of postal codes and addresses in the analysis of health data*. Health Reports, 1993; 5(2):157-177.
6. Wilkins R. *More about PCCF+ (for the hard core)*. 2005; Public Health Agency of Canada.
7. Statistics Canada. *Postal Code Conversion File (PCCF), Reference Guide*. 2006.
8. Mechanda K, Puderer H. *How postal codes map to geographic areas*. 2007; Statistics Canada.
9. Nicholas R. *ESRI Support Center: Create a grid polygon shapefile (FISHNET)*. 2003; Available from: <http://arcscripts.esri.com/details.asp?dbid=12807>.

10. Statistics Canada. *Dissemination Areas Cartographic Boundary Files (Geography Products: Spatial Information Products, 2001 Census)*. 2002.

## Additional file 2: Evaluating Dimensions of Risk

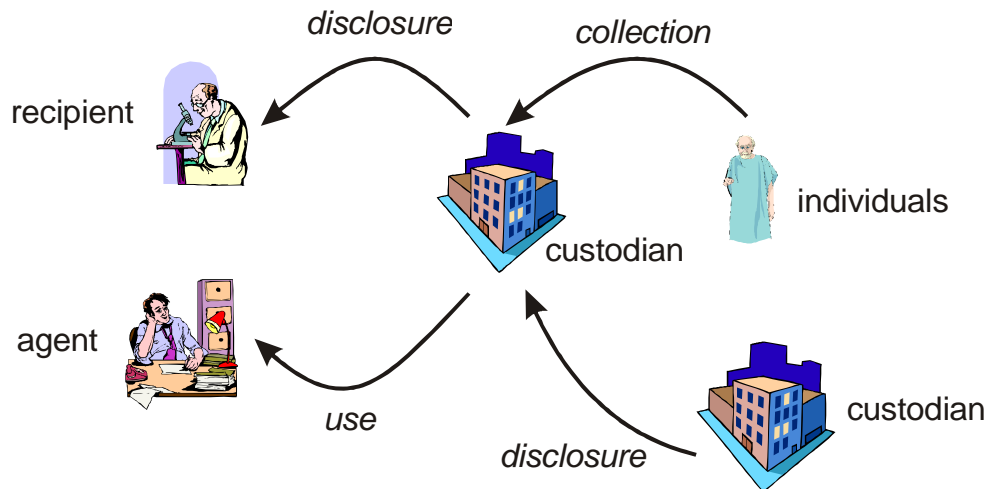
The purpose of this appendix is to provide a set of items that can be used by a custodian to evaluate the risk when health information is disclosed or used for secondary purposes. The specific dimensions we look at are “invasion-of-privacy” and “motives and capacity”. Some background for the context is first provided, followed by a detailed description of each item.

### **Background**

As illustrated in Figure 1, personal information is collected from individuals. This collection can be direct or indirect through reporters. For example, in the case of an adverse drug event, a hospital or a physician may report the adverse event rather than the patient herself.

This information remains with the custodian. An example of a custodian is a hospital or a disease registry. The custodian may have collected the information for a primary purpose, such as providing a service to a customer or providing care to a patient, or explicitly for a secondary purpose, such as a prospective diabetes registry.

A custodian may disclose personal information to another custodian. For example, a hospital may disclose personal health information to a public health agency. In such a case, the information is not coming directly from a patient but indirectly through one (or possibly more) custodians.



**Figure 1:** Basic data flow during a disclosure or use of personal information for secondary purposes.

An agent of the custodian may *use* the information for a secondary purpose. An agent is broadly defined as a person who acts on behalf of the custodian in respect of the personal information for the purposes of the custodian. For example, an analyst employed by a hospital to produce reports on resource utilization would be an agent. There is generally no legislative requirement to de-identify information that an agent uses and no requirement to obtain additional consent from the individuals/patients for such uses.

The custodian may also get a request to *disclose* the information to a recipient (or another custodian, but we will subsume that under recipient) for some secondary purpose. The recipient can be an individual (e.g., a researcher), or an organization (e.g., a pharmaceutical company or a public health agency). The recipient can also be internal or external to the custodian. For example, a researcher may be based within a hospital or can be an external researcher at a university or a government department requesting the information from the hospital.

Some disclosures are mandatory and some are discretionary to the custodian. An example of a mandatory disclosure is reporting communicable diseases or reporting gunshot wounds in some jurisdictions. In these situations the disclosure of personal information to a particular recipient is required.

Otherwise, there are different types of recipients and purposes where disclosures of personal information are discretionary. However, the conditions for discretionary disclosure do vary. There are a set of permitted disclosures in privacy legislation where personal information may be disclosed without consent, for example, disclosures for research and disclosures for planning and improving the health system.

Other discretionary disclosures that are not explicitly permitted in legislation require that either consent be obtained from the individuals/patients or that the information be de-identified. For example, the disclosure of personal health information (PHI) to a pharmaceutical company requires that consent be obtained or that the information is deemed to be de-identified.

Therefore, to summarize, there are four scenarios to consider:

- A. It is mandatory to provide personal information to a recipient (usually external to the custodian), and no consent is required.
- B. Personal information is used by an agent without consent.
- C. It is permitted by legislation to provide personal information to a recipient without consent (either internal or external to the custodian) under the discretion of the custodian.
- D. The custodian *must* de-identify the information *or* obtain consent before disclosing the data to the recipient.

The need for de-identification of the information under each of the above scenarios will vary. This is discussed further below.



## ***The Need for De-identification***

In three out of the above four scenarios where data is used or disclosed by a custodian, a strong case can be made for de-identification. Below we consider each in turn.

### **Scenario A: Mandatory Disclosures**

Disclosures under this scenario are outside our scope since they do not require any de-identification.

### **Scenario B: Uses by an Agent**

While agents are permitted to access personal information, if that is not necessary to perform their functions then it may be better to de-identify that information to minimize the consequences of a breach. The reason would be to mitigate risks due to data breaches, whose frequency has been increasing rapidly.

For example, consider a hospital network that has developed a system to provide its patients web access to its electronic health records. The hospital has sub-contracted the work to perform quality control for this software to a testing company across town. The testing company needs realistic patient data to test the software, for example, to make sure that the software can handle large volumes of patient records, that it displays the correct information to each patient, and so on. The testing company would be considered an agent of the hospital, so it can obtain identifiable patient records without consent, and use these records for testing. Giving the testing company PHI potentially exposes the hospital to additional risk if there is an inadvertent disclosure of this data (e.g., a breach at the testing company's site). It is always preferable from a risk management perspective to minimize the number of people who have access to PHI, and making that information available to the whole test team should be avoided if possible. Therefore in cases where there is a legitimate use of the PHI, one should still consider using de-identification techniques even if this is not a legal or regulatory requirement.

### **Scenario C: Permitted Disclosures**

In some cases, even though the disclosure of identifiable health information is permitted by legislation, the custodian may consider de-identification anyway. This, of course, makes sense only if the purpose can be satisfied without having identifiable information. In practice, achieving many purposes does not require identifiable information. A good example of that is in the context of research.

A Research Ethics Board (REB) determines whether custodians can disclose personal information to researchers, and whether that information needs to be de-identified. REBs have total discretion to make that decision.

In practice, most REBs will require that either consent from the patients be sought if the information needs to be identifiable or they will require that the disclosed information is adequately de-identified [2]. However, because of the discretionary nature of this type of disclosure, they may allow identifiable information to be disclosed without consent.

For example, consider the situation where a researcher is collecting clinical information from electronic health records (EHRs) and wants to link it with data in a provincial administrative database. The linking will not work if the EHR data is de-identified. In that case the REB may allow identifiable information to be disclosed for the purpose of linking without requiring the consent of the patients.

### **Scenario D: De-identification vs. Consent**

In this scenario the custodian does not have the option to disclose identifiable information without consent. However, there will be situations where obtaining consent is not possible or practical. For example, in a health research context, making contact with a patient to obtain consent may reveal the individual's condition to others against their wishes, the size of the population represented in the data may be too large to obtain consent from everyone, many patients may have relocated or died, there may be a lack of existing or continuing relationship with the patients to go back and obtain consent, there may be a risk of inflicting psychological, social or other harm by contacting individuals and/or their families in delicate circumstances, it may be difficult to contact individuals through advertisements and other public notices, and undue hardship may be caused by the additional financial, material, human, organizational or other resources required to obtain consent. In those instances, the disclosure of personal information would not be permissible and de-identification provides the only practical option for disclosure (assuming that the purpose can be achieved with the de-identified information). There is no legislative requirement to obtain consent for de-identified information.

Even if obtaining consent was possible and practical, it may have a severe adverse consequence on the information's quality because individuals who consent tend to be different on many characteristics than those who do not consent (e.g., on age, gender, socioeconomic status, whether they live in rural or urban areas, religiosity, disease severity, and level of education) [3]. These differences can result in biased findings when the information is analyzed or used. In such circumstances a strong case can be made for not seeking consent and de-identifying the information instead (again, assuming that the de-identified information will achieve the purpose of the disclosure).

Consider an example where a hospital is disclosing prescription data to a commercial data broker. It is not practical to obtain consent from the patients for this disclosure. Just the cost of administering the additional consent forms for admitted patients would be difficult to justify, and it would be difficult to do so retroactively for historical data. Therefore, the hospital would have to de-identify the prescription data before disclosure.

### ***Items***

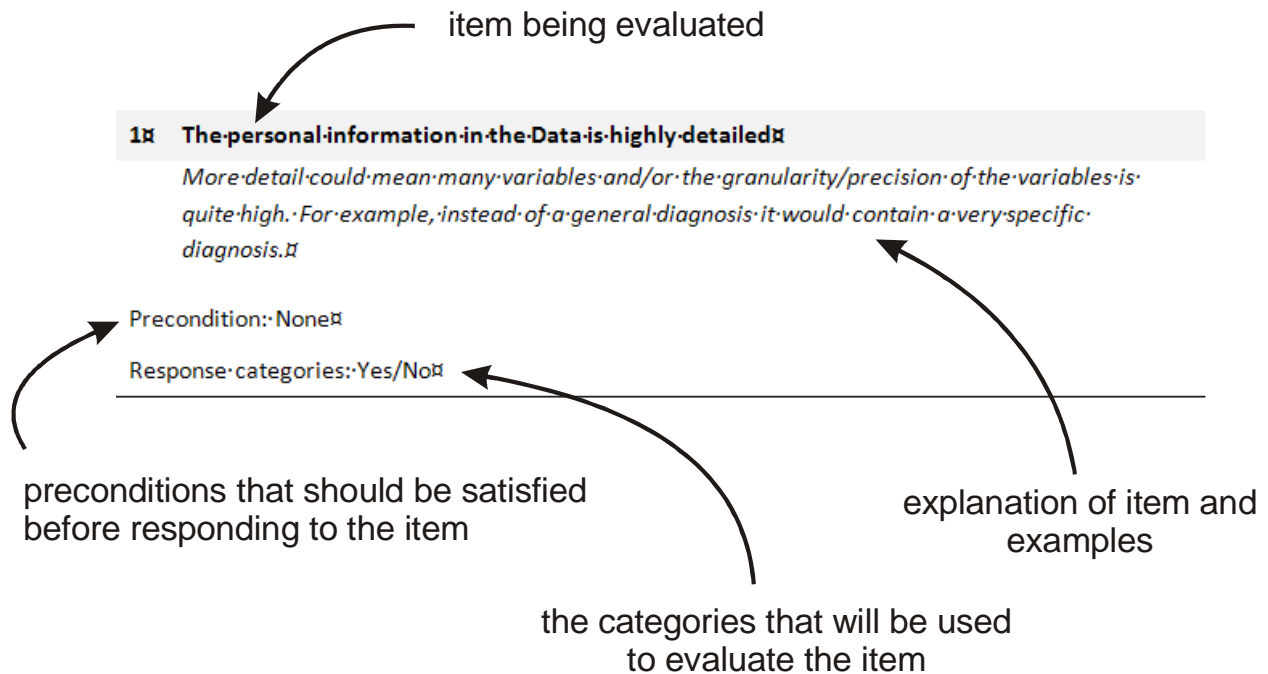
The following items along the two dimensions of risk, "invasion-of-privacy" and "motives and capacity", can be used to decide the risk profile of the disclosure or use.

#### **Subjectivity**

The items that are defined below can be subjective in some cases. The custodian would need to define their own standards for interpreting them to reduce the subjectivity. For example, the item on the level of detail in the health information that is used/disclosed requires that the custodian define that given the nature of their data and apply it consistently across all disclosures/uses.

#### **Layout**

The following is an explanation of the layout of each item that is used to evaluate invasion-of-privacy and motives and capacity:



## Assessing Invasion-of-Privacy

The objective of this section is to define a way to measure the invasion-of-privacy construct. There are degrees of invasion-of-privacy, and the items in this section determine that degree. By measuring the extent of the potential invasion-of-privacy, it will be possible for the custodian to decide how much de-identification needs to be done. For example, if data on a stigmatized disease is disclosed to a recipient, then that would score higher on invasion-of-privacy than disclosing data on common allergies. In both cases there would be an invasion-of-privacy, but in terms of degree the latter would be greater and therefore the data requires more de-identification.

Invasion-of-privacy is a hypothetical construct. In our definition of invasion-of-privacy we make two important assumptions:

1. An invasion-of-privacy can only occur if the data that is disclosed/used is identifiable. Therefore, all of the items below are based on the assumption that all of the data is identifiable by the recipient/agent. The custodian may have disclosed identifiable data, the disclosed data was de-identified and the recipient was able to re-identify it somehow, or the agent is using identifiable information provided by the custodian. When we talk about data in the context of this construct, then, we are referring to personal information or personal health information.
2. The disclosure/use will not entail going back to the patients and seeking their consent.

With the above assumptions, an invasion-of-privacy can occur under three conditions:

1. If the custodian inappropriately discloses the data to the recipient or there is an inappropriate use of the data.
2. If the recipient or agent inappropriately processes the data (e.g., in terms of the analysis performed on it).
3. There is a data breach at the recipient or agent site (whether it is deliberate or accidental).

The items below are intended to assess the different dimensions of invasion-of-privacy if any of the three conditions above are satisfied.

The custodian is expected to be able to respond to/assess all of the items below. In some cases the custodian may have to exercise their best judgment in order to respond.

It is assumed that it would be possible to make general assessments about all of the patients covered by the data, even if this is an approximation. For example, some patients may care if they have been consulted if their data is disclosed/used for secondary purposes, while others may not. However, if a nontrivial proportion of the patients would have cared then the particular item would be rated closer to the affirmative.

## ***Definitions***

<b>Data</b>	This is identifiable or potentially identifiable information. The Data can be identifiable if it explicitly contains identity information, such as names and phone numbers. The Data is potentially identifiable if it is relatively easy for the recipient or agent to assign identity to the Data. For example, if the identity information was replaced by pseudonyms and the recipient/agent is able to reverse engineer the pseudonyms because s/he has the pseudonym-to-identity mappings or can get them. Alternatively, the recipient/agent may have the power to compel the release of identity information. For example, if the Data has an IP address and the recipient is a law enforcement agency, then the agency may be able to compel the ISP to reveal the name and address associated with the IP address at the specified date and time.
<b>Purpose</b>	This is the purpose for which the recipient/agent has requested the Data.

## ***Dimensions***

The invasion-of-privacy construct has four dimensions:

- The sensitivity of the Data: the greater the sensitivity of the data the greater the invasion-of-privacy.
- The potential injury to patients from an inappropriate disclosure/use/breach/processing: the greater the potential for injury the greater the invasion of privacy.

- The appropriateness of consent for disclosing/using the Data: the less appropriate the consent the greater the invasion-of-privacy.

These are detailed further below.

## A. Sensitivity of the Data

### 1 The personal information in the Data is highly detailed

*More detail could mean many variables and/or the granularity/precision of the variables in the Data is quite high. For example, instead of a general diagnosis it would contain a very specific diagnosis. For instance, a high level diagnosis would be “disorders of the thyroid gland”, whereas a more detailed diagnosis would be “nontoxic nodular goiter”, and “absence of teeth” can be generalized to “diseases of oral cavity, salivary glands, and jaws”.*

Precondition: None

Response categories: Yes/No (Yes = more invasion-of-privacy; No = less invasion-of-privacy)

---

### 2 The information in the Data is of a highly sensitive and personal nature

*This could mean, for example, information about: sexual attitudes, practices, and orientation; use of alcohol, drugs, or other addictive substances; illegal activities; suicide; sexual abuse; sexual harassment; mental health; certain types of genetic information; and HIV status.*

*Information about a stigmatized disease/condition or that can adversely affect a patient’s business dealings, insurance, or employment would also be considered sensitive.*

Precondition: None

Response categories: Yes/No (Yes = more invasion-of-privacy; No = less invasion-of-privacy)

---

### 3 The information in the Data comes from a highly sensitive context

*For example, in most cases data about individuals participating in a youth employment program are less sensitive than a similar list containing names and addresses of Hepatitis C and HIV compensation victims. But the sensitivity may also be dependent on the specifics. For example, a list of customers on a newspaper carrier’s route may not be sensitive, unless the newspaper or publication being distributed indicates sexual orientation, for instance.*

Precondition: None

Response categories: Yes/No (Yes = more invasion-of-privacy; No = less invasion-of-privacy)

---

## B. Potential Injury to Patients

### **1 Many people would be affected if there was a Data breach or the Data was processed inappropriately by the recipient/agent**

*This item pertains to the number of patients covered by the data. More patients would be injured if there was, say, a breach of data on 10000 patients than a breach of data on 10 patients. In both cases it is an undesired outcome, but the former is more severe.*

*The new US HITECH Stimulus Package stipulates that any breach involving 500 or more than individuals must be reported to the Department of Health and Human Services. This can be used as a guide for what is considered as a large number of people.*

*If an inappropriate disclosure would affect a defined community (e.g., a minority group living in a particular area) then the number of people affected would be larger than the patients covered by the Data.*

Precondition: None

Response categories: Yes/No (Yes = more invasion-of-privacy; No = less invasion-of-privacy)

---

### **2 If there was a Data breach or the Data was processed inappropriately by the recipient/agent that may cause direct and quantifiable damages and measurable injury to the patients**

*Damages and injury would include physical injury such as due to stalking or harassment; emotional or psychological harm; social harm such as stigmatization, humiliation, damage to reputation or relationships; financial harm, such as (medical) identity theft and financial fraud; and if the data can be used in making a decision that is detrimental to the patient, for example, a business, employment or insurance decision. The damages and injury can occur to the patient(s) themselves, their family unit, or to a defined group/community (e.g., neighborhood, minority groups, band leaders, Aboriginal people, people with disabilities).*

Precondition: None

Response categories: Yes/No (Yes = more invasion-of-privacy; No = less invasion-of-privacy)

---

**3 If the recipient/agent is located in a different jurisdiction, there is a possibility, for practical purposes, that the data sharing agreement will be difficult to enforce**

*It is assumed that there is some form of data sharing agreement between the custodian and the recipient/agent. For example, if the agent is an employee of the custodian, then there would be obligations in employment contracts. If the agent is a different company then there would be a contract between the custodian and that company. If the recipient is a researcher in a different institution, then a data sharing agreement would be signed by the recipient.*

*This particular item becomes relevant under the circumstances where the recipient/agent is in a different jurisdiction than the custodian, for example, in the US the PATRIOT Act compels custodians to disclose data in secret. In that case a law in a different jurisdiction effectively overrides the provisions in the data sharing agreement.*

*In some jurisdictions enforcing contracts in courts is difficult or exceedingly slow that for practical purposes the data sharing agreement cannot be enforced in that jurisdiction.*

Precondition: None

Response categories: Yes/No (Yes = more invasion-of-privacy; No = less invasion-of-privacy)

---

**C. Appropriateness of Consent**

**1 There is a provision in the relevant legislation permitting the disclosure/use of the Data without the consent of the patients**

*In some cases there will be legislative authority to disclose the Data without consent. For example, when the Data is being disclosed to a medical officer of health at a public health authority. But if the recipient was a commercial data broker then there is no exception allowing the disclosure without consent. In Ontario, custodians can disclose Data to Prescribed Entities without the patients' consent.*

*In the case of research, a Research Ethics Board (REB) is permitted in most jurisdictions to disclose the Data without consent. If the REB elects not to do so the response to this question would still be Yes.*

*Uses of Data by agents without consent are permitted in Canadian jurisdictions. Therefore, all subsequent items in this section pertain to disclosures only.*

Precondition: None

Response categories: Yes/No (Yes = less invasion-of-privacy; No = no change in invasion-of-privacy)

---

**2 The Data was unsolicited or given freely or voluntarily by the patients with little expectation of it being maintained in total confidence**

*This would pertain, for example, to patients posting their Data on a public web site as part of a discussion group. It is not always obvious that when patients post their Data on the web there is an expectation of privacy, but in some cases they may not understand the privacy settings or policies of the web site, or the organization running the web site may change their policy after the Data was collected in unexpected ways. Therefore, the response to this question must take into account the specific context and history of the location where the patients posted their information.*

Precondition: If item C(1) is endorsed, then this item would not apply

Response categories: Yes/No (Yes = less invasion-of-privacy; No = no change in invasion-of-privacy)

---

**3 The patients have provided express consent that their Data can be disclosed for this secondary Purpose when it was originally collected or at some point since then**

*This item refers to obtaining explicit consent from the patients (opt-in or opt-out). The consent may have been for the recipient's specific project (for example, in the case of patients consenting for the data that was collected during the provision of care to also be used for a specific research analysis), or may have been broad to encompass a class of projects that include the recipient's Purpose for processing the Data (for example, the patients consented for their data to be used for research on cardiovascular diseases, without knowing in advance what the possible research questions may be).*

Precondition: If items C(1) or C(2) are endorsed, then this item would not apply

Response categories: Yes/No (Yes = less invasion-of-privacy; No = more invasion-of-privacy)

---



**4 The custodian has consulted well-defined groups or communities regarding the disclosure of the Data and had a positive response**

*This item would be endorsed Yes if these well defined groups or communities did not raise objections to the particular disclosure/use. If they did consult and the outcome was negative, then the item is scored No.*

*Well defined groups or communities include neighborhood members, minority groups, band leaders, Aboriginal people, people with disabilities, consumer associations, community representatives, privacy oversight bodies, and patient advisory councils.*

*The assumption with this item is that a nontrivial proportion of patients care what their group/community thinks about the disclosure and that they be consulted.*

Precondition: If items C(1), C(2), or C(3) are endorsed, then this item would not apply

Response categories: Yes/No (Yes = less invasion-of-privacy; No = more invasion-of-privacy)

---

**5 A strategy for informing/notifying the public about potential disclosures for the recipient's secondary Purpose was in place when the data was collected or since then**

*The custodian may have given notice of potential disclosures for secondary purposes, for example, through well located posters at their site. The notice does not need to explicitly mention the particular recipient's Purpose, but should describe potential purposes that include the recipient's Purpose.*

*This is an example of obtaining implicit consent when there are no legislative exceptions and express consent was not obtained.*

Precondition: If items C(1), C(2), or C(3) are endorsed, then this item would not apply

Response categories: Yes/No (Yes = less invasion-of-privacy; No = more invasion-of-privacy)

---

## 6 Obtaining consent from the individuals at this point is inappropriate or impractical

*For example, making contact to obtain consent may reveal the individual's condition to others against their wishes, the size of the population is too large to obtain consent from everyone, many patients have relocated or died, there is a lack of existing or continuing relationship with the patients, the consent procedure itself may introduce bias, there is a risk of inflicting psychological, social or other harm by contacting individuals and/or their families in delicate circumstances, it would be difficult to contact individuals through advertisements and other public notices, and undue hardship that would be caused by the additional financial, material, human, organizational or other resources required to obtain consent.*

*This assessment may be contextual. For example, obtaining consent may be difficult for a researcher with limited funds, but if a large organization is requesting the data and they are expected to generate a large amount of revenue from processing the data, then the custodian may be able to convince the recipient that it is worth their while to invest in obtaining consent.*

Precondition: If items C(1), C(2), C(3), C(4) or C(5) are endorsed, then this item would not apply

Response categories: Yes/No (Yes = less invasion-of-privacy; No = more invasion-of-privacy)

---

## **Assessing Motives and Capacity**

The objective of this sub-section is to define a way to measure the motives and capacity construct. This construct assumes that the custodian is disclosing/using data that has gone through some kind of de-identification. Therefore, we are concerned with the motives and capacity of the recipient/agent to re-identify this data.

This construct has two dimensions: "motives" and "capacity". Since "motives" pertain to individuals, the motives dimension can be considered in terms of the staff, collaborators, or employees of the recipient/agent entity. The motive to re-identify the data implies an intentional re-identification. The capacity dimension evaluates whether the recipient/agent is able to re-identify the data, irrespective of whether the re-identification is intentional or not.

The custodian is expected to be able to respond to/assess all of the items below. In some cases the custodian may have to exercise their best judgment in order to respond as some of the items are subjective.

## Definitions

<b>Data</b>	The data is assumed to have gone through some kind of de-identification before it is disclosed/used. The amount of de-identification will vary depending on the specifics of the disclosure/use.
<b>Purpose</b>	This is the purpose for which the recipient/agent has requested the Data.

### A. Motives to Re-identify the Data

#### 1 The recipient/agent has directly or indirectly worked/collaborated with the custodian in the past without incident

*This item assumes that this collaboration has not resulted in any incidents where the recipient/agent processed the data in an inappropriate way or attempted to re-identify the data (i.e., it was perceived to be a successful collaboration). If the custodian has worked with the recipient/agent before then there is an empirical trust that has been built up, suggesting that the recipient is trustworthy.*

Precondition: None

Response categories: Yes/No (Yes = fewer motive to re-identify; No = greater motive to re-identify)

---

#### 2 The recipient/agent can potentially gain financially from re-identifying the Data

*The first consideration is whether the recipient/agent is in financial distress. Although, this may be difficult to assess in practice.*

*Consider if the recipient/agent or his/her family/employees/collaborators may receive financial benefits from processing identifiable data. For example, a pharmaceutical company may want to contact the patients directly for marketing purposes or to recruit them in a study.*

*Another consideration is if the Data, once re-identified, can be useful for committing financial fraud or identity theft (e.g., the database has dates of birth and mother's maiden name).*

Precondition: None

Response categories: Yes/No (Yes = greater motive to re-identify; No = fewer motive to re-identify)

---

**3 There is possibly a non-financial reason for the recipient/agent to try to re-identify the Data**

*For example, there may a reason that the recipient/agent may want to embarrass the custodian by demonstrating that re-identification is possible, or say a reporter wanting to do a story about a specific identifiable person in the Data or a famous person known to be in the Data. Also, a disgruntled employee may wish to adversely affect the custodian's reputation by re-identifying a patient and making that public.*

Precondition: None

Response categories: Yes/No (Yes = greater motive to re-identify; No = fewer motive to re-identify)

---

**B. Capacity to Re-identify the Data**

**1 The recipient/agent has the technical expertise to attempt to re-identify the Data**

*Re-identification requires some basic database and statistical expertise. However, in real data sets there are missing data and data errors which would also have to be accounted for in terms of expertise. Of course an incorrect re-identification can also be problematic, but we are only concerned with a correct re-identification here.*

Precondition: None

Response categories: Yes/No (Yes = greater capacity to re-identify; No = less capacity to re-identify)

---

**2 The recipient/agent has the financial resources to attempt to re-identify the Data**

*Some types of re-identification require funds to get data sets to link with. Also, gathering background information about a patient in the Data who is a target of re-identification can be costly.*

Precondition: None

Response categories: Yes/No (Yes = greater capacity to re-identify; No = less capacity to re-identify)

---

**3 The recipient/agent has access to other private databases that can be linked with the Data to re-identify patients**

*Such private databases would only be useful if they contain the identity information about the patients. Then linkage with the de-identified database could reveal the identity of one or more patients in the Data.*

*Some data that can be used for linking and re-identification could be publicly available. In such a case we would consider item B(2) on the financial resources of the recipient/agent. This item pertains to private databases.*

*A recipient may obtain such private databases from previous disclosures by the custodian. For example, the custodian may have disclosed a particular dataset to a researcher last year, and this year the same researcher wants another dataset that can be linked to the earlier one. The recipient may also obtain a private database by colluding with someone else. For example, the researcher may arrange to link a new administrative dataset from the custodian with another researcher who has obtained a different clinical dataset from the same custodian (and the custodian would not approve for the two datasets to be linked).*

*An agent can also have access to data useful for linking. For example, in hospitals many staff have access to administrative data but not to clinical data. An employee can get a de-identified clinical data set and link it with the readily available administrative data set to re-identify patients in the clinical dataset.*

Precondition: None

Response categories: Yes/No (Yes = greater capacity to re-identify; No = less capacity to re-identify)

---

Editorial

Open Access

## Musings on privacy issues in health research involving disaggregate geographic data about individuals

Maged N Kamel Boulos\*<sup>1</sup>, Andrew J Curtis<sup>2</sup> and Philip AbdelMalik<sup>1</sup>

Address: <sup>1</sup>Faculty of Health and Social Work, University of Plymouth, Drake Circus, Plymouth, Devon, PL4 8AA, UK and <sup>2</sup>GIS Research Laboratory, Department of Geography, University of Southern California, Kaprielian Hall (KAP), Room 416, 3620 South Vermont Avenue, Los Angeles, CA 90089-0255, USA

Email: Maged N Kamel Boulos\* - mnkamelboulos@plymouth.ac.uk; Andrew J Curtis - ajcurtis@usc.edu; Philip AbdelMalik - philip.abdelmalik@plymouth.ac.uk

\* Corresponding author

Published: 20 July 2009

Received: 5 July 2009

Accepted: 20 July 2009

*International Journal of Health Geographics* 2009, **8**:46 doi:10.1186/1476-072X-8-46

This article is available from: <http://www.ij-healthgeographics.com/content/8/1/46>

© 2009 Boulos et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

This paper offers a state-of-the-art overview of the intertwined privacy, confidentiality, and security issues that are commonly encountered in health research involving disaggregate geographic data about individuals. Key definitions are provided, along with some examples of actual and potential security and confidentiality breaches and related incidents that captured mainstream media and public interest in recent months and years. The paper then goes on to present a brief survey of the research literature on location privacy/confidentiality concerns and on privacy-preserving solutions in conventional health research and beyond, touching on the emerging privacy issues associated with online consumer geoinformatics and location-based services. The 'missing ring' (in many treatments of the topic) of data security is also discussed. Personal information and privacy legislations in two countries, Canada and the UK, are covered, as well as some examples of recent research projects and events about the subject. Select highlights from a June 2009 URISA (Urban and Regional Information Systems Association) workshop entitled 'Protecting Privacy and Confidentiality of Geographic Data in Health Research' are then presented. The paper concludes by briefly charting the complexity of the domain and the many challenges associated with it, and proposing a novel, 'one stop shop' case-based reasoning framework to streamline the provision of clear and individualised guidance for the design and approval of new research projects (involving geographical identifiers about individuals), including crisp recommendations on which specific privacy-preserving solutions and approaches would be suitable in each case.

### Introduction

#### **Definitions—the security-confidentiality-privacy triad**

In micro-scale geographical analyses involving data about specific individuals, data security, confidentiality and privacy form an intertwined triad. A recent US CDC (Centers for Disease Control and Prevention) foundation course on public health law [1] defines privacy as the "individual's right to control the acquisition, use and disclosure of

their identifiable health information". The same course goes on to define confidentiality as the "privacy interests that arise from specific relationships (e.g., doctor/patient, researcher/subject) and corresponding legal and ethical duties", and then describes security as the "technological or administrative safeguards or tools to protect identifiable health information from unwarranted access, use, or disclosure". To explain the relationships between the

three terms, the course quotes a key sentence from Ware [2]: "If the security safeguards in an automated system fail or are compromised, a breach of confidentiality can occur and the privacy of data subjects invaded".

#### **Mainstream media and public interest in the subject**

Actual or potential breaches (technological or legal) of data security and confidentiality and the subsequent actual or potential invasions of individuals' privacy are quite commonly reported in mainstream media. For example, in March 2009, the Joseph Rowntree Reform Trust published its 'Database State' report on the legality, safety and effectiveness of the British government's major database systems [3,4]. Of 46 databases assessed in this report, only six were found to have a proper legal basis for any privacy intrusions and were deemed proportionate and necessary in a democratic society. The report authors concluded that two NHS (National Health Service) systems, the Detailed Care Record (DCR) and the Secondary Uses Service (SUS) [5], were almost certainly illegal and that a number of others including the Summary Care Record (SCR) would be legal only with patient consent, but, with the current absence of an effective opt-out, it too was almost certainly illegal.

We also read the story of an anonymous Canadian girl whose death was associated with a prescribed acne drug. She was eventually identified by the media who compared the de-identified prescription data set against obituaries. The comparison helped in narrowing down the search to four possible girls, then by contacting all families the right one was found [6].

High-profile security breaches (e.g., data loss or theft of ill-protected confidential data) are also not uncommon. For example, it was reported in May 2009 that a laptop containing non-encrypted data (names, addresses, dates of birth, employers, national insurance numbers, salary information, and bank details) of 109,000 UK pensioners has been stolen [7]. The data were merely password-protected, and possibly without any appropriate safeguards for data self-destruction in case of brute-force password attacks. It is very easy to find out passwords in a short time using common hardware, e.g., NVIDIA CUDA GPUs (Compute Unified Device Architecture Graphics Processing Units), and readily available software [8], or even to completely bypass the passwords and directly access the underlying non-encrypted data.

In public health worldwide, any public identification of an individual's health status and address, regardless of contagion level or risk, is usually prohibited. But individual privacy rights must also be balanced with legitimate public concerns and interests. The publicly-accessible, online mapping of SARS (Severe Acute Respiratory Syn-

drome) in Hong Kong a few years ago using disaggregate case data at individual infected building level in near real time was one of the noticeable exceptions to the well-established public health confidentiality rule [9,10].

#### **Research literature: location privacy concerns and solutions**

The biomedical and public health literature on geographic information systems (GIS) and spatio-temporal analyses features a large number of research papers mentioning or addressing location privacy, e.g., [11-28]. A must-read paper (not specifically health-related) dating back to 1994 [29] shows how chronic privacy issues are in GIS research. Some research papers identified privacy as a potential or actual issue of concern (e.g., in reproductive health research [18]; in birth defects surveillance and research [19]; in research relevant to policy on diet, physical activity, and weight [20]; in environmental health research [21]; and in health and social care planning [22]), while others went one step further by suggesting some comprehensive solutions (e.g., [23-26]), workarounds, or frameworks and principles of practice (e.g., [29]) to mitigate or resolve these privacy concerns.

A number of confidentiality-preserving statistical and epidemiological data processing methods (data aggregation and transformations) have been proposed that can be applied to original location data to preserve individuals' privacy while maintaining some acceptable level of data usefulness for geographical analyses. But the use of precise addresses will continue to be needed in many cases to improve data analysis results or make them possible at all. The famous John Snow's map of the 1854 Cholera outbreak in London only solved the problem because the unique locations of individual cases were known [15,30]. There will always be this implicit trade-off between privacy concerns (e.g., easiness of re-identification) and the types and accuracy of the results of geographical health analyses that are possible with a given data set (original, unaltered vs. transformed or aggregated data) [25,27,28]. And that is where software agents can offer a potential solution that preserves the full fidelity of the original data [25].

Moving beyond conventional GIS research and geographical analyses, mobile phones and other electronic gadgets are rapidly gaining location awareness and wireless Web connectivity, thus promising new spatial technology applications and services (e.g., [31-33]), which will yield vast amounts of spatial information and online maps that can even reveal users' whereabouts in real time. These novel spatial tools and services are certainly opening many new useful possibilities, but are not without their challenging security and privacy concerns [34,35].

### **The 'missing ring': data security**

Data security is relatively under-mentioned in discussions about confidentiality-preserving solutions for location data, despite its key importance in the aforementioned security-confidentiality-privacy triad. Consider the following scenario: a health GIS researcher has legitimate and IRB (institutional review board)-approved access to patient data containing precise geographical identifiers for analysis and reporting purposes, with full patient consent. The reporting is done in ways that do not identify individual patients when posting publicly-accessible/online results and maps. If the reporting must be made at some level of detail or granularity that can potentially identify individual patients, the results are only shared within approved, small teams of users with legitimate access rights and 'need to know'. The whole scenario seems fine as far as the protection of individuals' privacy is concerned. IRB approval has been sought, adequate reporting methods and policies are in place to prevent the disclosure of any confidential data to non-authorised parties, and we even have the patients' explicit consent to conduct the study. However, without appropriate additional security safeguards, there will always be many unmitigated risks of data theft or loss and of unwanted data disclosure to non-authorised or non-authenticated parties, all of which can compromise the privacy of the data subjects. (Ideally, IRBs should be scrutinising the security component as well before granting approvals.)

A carefully blended, purpose-built combination of overlapping security measures is always the solution, depending on the type, sensitivity, value, and risks/costs assessment of the data to be protected. Various types of advanced cryptography, multimodal biometrics, and other methods can be combined, as necessary. Data access can also be controlled or restricted in such a way that two or more persons must be physically present each time and authenticated (e.g., via biometrics) to unlock the data. Security measures cover and include, among other things, ensuring physical building security, using computer security cable locks, using computers with a built-in TPM (Trusted Platform Module) chip for cryptographic functions, performing full disk encryption with TPM (e.g., using BitLocker [36]), implementing brute-force password attack protection (data are automatically erased after a pre-set number of failed access attempts), using hardware/software firewalls and other forms of network security, implementing adequate access policies and authentication [37] (at computer BIOS-Basic Input/Output System level, Operating System-level, and application level), considering Multilevel Security (MLS), using biometrics (e.g., fingerprint readers and facial recognition), using advanced secure USB flash drives with military grade hardware encryption (e.g., [38,39]) instead of ordinary flash drives, keeping detailed data inventories and

electronic audit trails of all accesses and transactions, blanking of computer display and machine locking or auto-log-off if a machine is left unattended, and the secure decommissioning and discarding of old equipment and data storage media, e.g., using software utilities like SDelete [40] to prevent the kind of issues described in [41]. Also equally important are staff training and the development of a 'security culture' in the organisation, e.g., guided by ISO/IEC 27002 2005 (formerly ISO/IEC 17799 2005), an ISO (International Organization for Standardization) standard for information security and a code of practice for information security management.

### **Personal information and privacy legislations**

A discussion on location privacy solutions for health research would be incomplete without reflecting on some of the underlying reasons that necessitate their development. The very notion of privacy is itself a complex fabric of interwoven philosophical and psychosocial threads. Perhaps this is why the associated bureaucratic and legal landscape is as complex as it is – and often blamed for the issue. A large majority of public health professionals consider privacy to be an obstacle to public health; when asked for the underlying reasons, survey respondents in Canada and the UK most commonly identified bureaucracy and legislation [42].

There is no universal legislation to guide and govern the activities of public health professionals, particularly where issues of privacy are concerned. Instead, nations have their own constraining or enabling privacy and data protection laws, with some being such a maze of cross-referenced "legalese" that familiarising oneself with them – let alone gaining a thorough understanding of them – becomes a daunting task. 'Additional file 1' provides a brief compilation and comparison of relevant personal information and privacy legislation in Canada and UK, with particular focus on location and public health as seen and understood by an epidemiologist.

### **Some recent research projects and events about the subject**

The issues of location privacy were also the subject of GeoPKDD (Geographic Privacy-Aware Knowledge Discovery and Delivery [43]), a three-year EU-funded project that was recently completed in November 2008. GeoPKDD's main research question was 'how to discover useful knowledge about human movement behaviour from mobility data (e.g., location data from mobile phones), while preserving the privacy of the people under observation?' The project attempted to develop new privacy-preserving methods for extracting knowledge from large amounts of raw data about individuals referenced in space and time. GeoPKDD organised the 'First Interdisciplinary Workshop on Mobility, Data Mining and Privacy: Preserv-



ing anonymity in geographically referenced data' on 14 February 2008 in Rome, Italy [44].

Another research activity worth mentioning in our context is the proposal by Helen Chen at Agfa HealthCare in Canada and her colleagues at the World Wide Web Consortium–W3C Semantic Web for Health Care and Life Sciences Interest Group (HCLSIG) to explore Semantic Web solutions for patient data security, confidentiality, consent and privacy (in general, i.e., they are not focusing on location privacy, but their proposal is still broadly relevant to our topic). Previously sufficient de-identification techniques can be rendered inadequate because it is now possible to re-identify an identity via inference on the Web. Semantic Web technology is making headway to even more powerful data links, connections and inferences of this type. However, in the healthcare domain, this very success of the technology is putting individuals' privacy at much greater risks. Chen's idea is to develop novel privacy-preserving solutions by harnessing the very same Semantic Web technology that can exacerbate these privacy risks [45].

From 5–8 June 2009, the Urban and Regional Information Systems Association (URISA [46]), a non-profit American association of professionals using GIS and other information technologies to solve challenges in state/provincial and local government agencies and departments, organised its Second GIS in Public Health Conference in Providence, Rhode Island, USA. One of the pre-conference workshops held on the 5th of June 2009 focused on issues related to 'Protecting Privacy and Confidentiality of Geographic Data in Health Research'. Select highlights from this workshop are presented in the remaining part of this article.

### Select highlights from a recent URISA workshop on location privacy in health research

At the 2009 URISA GIS in Public Health Conference, a workshop organised by Ellen Cromley and Andrew Curtis focused on the issue of location privacy in health research. Among the topics covered by panellists and attendees were methods of spatial data protection, the need to "educate" IRBs, challenges facing data owners and custodians wishing to visualise and disseminate data, how published maps continue to violate confidentiality, some general cartographic guidelines and "fixes", and new methods of spatial data masking. In addition, the participants spent considerable time discussing the ethical and legal challenges researchers now face as HIPAA (US Health Insurance Portability and Accountability Act) regulations change, placing more responsibility on the data user (researcher). Although the majority of attendees to the meeting were data owners or custodians, this article is written mainly from the perspective of the data user, espe-

cially a social science/geographic information science researcher. As researchers, our usual role is to spatially analyse data, collect new spatial data with health implications, and visualise results in multiple forums, especially academic journals.

In 2006 Curtis *et al.* published a paper in *International Journal of Health Geographics* highlighting the potential for point level data to be reengineered from published maps through a process of digital scanning and georeferencing, even with only limited geographical features [11]. By heads-up digitising these points, coordinates could be used to direct field teams to actual homes. This conceptual approach had previously been impossible to replicate with real data, but by using this case from Hurricane Katrina, the map of mortality locations, and search and rescue markings that actually identified where bodies were found, validation was possible. Concurrent to this article, other reengineering approaches appeared using simulated data and a more systematic approach to identifying homes from a low resolution map [12]. Both papers revealed that published maps, even of low resolution and with limited geographical information, could still be reengineered back to an exact address, or so close to the 'real world' location that even without resorting to use other quasi identifiers, the spatial confidentiality of those being mapped was violated.

As researchers specialising in geographic information, we need to be proactive in setting guidelines for the display of confidential data, in policing our own actions, and in educating those sitting in positions of data power, especially our IRBs. Critics of the presentation usually focus on the data source—"this is a newspaper map so there is no confidentiality violation". However, there have been at least two maps appearing in journals that have also published the same Katrina mortality point locations. But irrespective of this, the real message is, *any point level map can be reengineered back in the same way*. As academics where does our ethical path lie with these secondary sources obtained from the media? We may not legally be violating confidentiality, but does that give us the right to use non-official sources, apply our geospatial skills and create sensitive layers in other outlets?

Now in mid-2009, what has changed? Are maps still being published in academic journals that violate spatial confidentiality? And where are we on the issue of cartographic guidelines? Unfortunately it is still too easy to find similar map violations appearing after 2006. One can find examples of maps with point level mortality locations, pregnancies, at-risk pregnancy programme participants, and people suffering from different respiratory ailments – indeed we challenge the reader to see how many point level health maps they can find. Of particular concern are

those sub-disciplines which have just discovered the value of GIS—we cannot expect that confidentiality violation through cartographic design is uppermost in the minds of those effusing over the wonders of buffering.

What else has changed? Paradoxically, the attention currently being paid to geocoding accuracy – which is important from a health research perspective, and which has received considerable attention in *International Journal of Health Geographics* – also has a detrimental side in terms of making published source maps both more accurate and precise. This means the chance of successful reengineering in terms of being closer to the actual address has increased. In effect, this previously unintentional form of masking has been reduced. Secondly, smoothing approaches, such as density surfaces, are being used to preserve confidentiality in maps (and stated as such by the authors). On one hand this is good news in terms of researchers' understanding that there is a confidentiality issue, but on the other hand this quick-fix is problematic due to a reliance on techniques that do not achieve this goal. The combination of window/kernel/filter size, the underlying grid cell resolution, and especially if there is no option for a minimum denominator, may result in "bulls-eyes" for areas of the map with relatively few residential alternatives, otherwise known as the 'geographic area population size' [47]. It should also be remembered that less dense geographies are not necessarily rural; many urban areas also contain physical features (inlets, lakes, even hills) can remove alternative possible locations. By referring to high resolution aerial photography (now found easily in applications such as Google Earth [35]), it is relatively easy to identify the cause of the intensity. On this subject, geospatial Internet applications in general have made the reengineering process even easier for those with and without a working knowledge of GIS.

From a data users' perspective, we are still limited by data being released at an aggregation that is limited for research, the standard for HIPPA being a zip code with 20,000 individuals. A group of Canadian researchers showed that this is an archaic approach and that minimum denominators should vary when taking into account the underlying geography and the number of quasi identifiers [47]. Similar papers written for researchers in other countries, possibly even providing a series of size guidelines for different urban areas, would be invaluable. It would also help the job of IRBs.

And on the subject of IRBs, from our experience there is still a disconnection in terms of understanding exactly where the risks lie in geospatial output and confidentiality. This is understandable given the confusion even amongst geospatial researchers. What would benefit all concerned would be a well-respected body in the field of

public health to commission a "guidelines" paper. This could become *the reference* in terms that researchers, IRBs, and even research subjects could understand and cite, along with other existing key texts, such as [48]. These reference guidelines should include clear visual examples of what is not acceptable, including the pitfalls of common "fixes" such as smoothing. They could also provide guidance for appropriate aggregation denominator sizes. This is important as researchers seek IRB approval in the use of mobile geospatial devices for collecting health and built environment data. We cannot expect IRBs to understand where such cartographic risks lie. Finally, language should be included that would help IRBs and be required in any letter gaining subject permission. In other words, "*if we ask for an address (or a street intersection... or a zip code...) this is the only way we will display it on a map*". This simple approach would mean that IRB, researcher and subject would all have the same understanding of what will happen with these data. (Ellen Cromley has vested considerable time on spatially appropriate language for informed consent as a guideline for IRBs. She disseminated examples of this language at the URISA workshop.)

This 'best practice guide' should be circulated to all journals who publish maps, clearly stating the risks involved in accepting point level maps. At least this would enlighten editors [49] and hopefully force them to ask submitting authors about '*what steps have been taken to preserve confidentiality?*'

Until we have such a universally accepted document, self-policing is the main option, and with this in mind, we have a few issues a researcher should ponder before publishing any map. Most importantly, is a point-level (or smoothed, or small aggregation) map necessary? As a geographer this last statement certainly hurts, but unless a map is really needed to help frame a paper's content, or improve the understanding of the reader regarding a spatial process, and especially if it is not even specifically referred to in the text, then it is better to err on the side of caution.

We fully realise that some point-level maps will still need to be published; it is often easier to explain a spatial process through a graphic, but if this is the case then is the underlying geography needed? If we are overlaying points against output from a spatial analysis, do we need political boundaries or street networks? If geographical references are necessary in the map, then data masking is essential.

There is some good news though, as we have noticed more researchers referencing steps taken to preserve confidentiality during recent presentations.

**Emerging issues**

There are three emerging spatial confidentiality topics of concern. The first involves Google Street View [50], an excellent research tool that allows us to "see" areas that are described or mapped in publications. The implication this has for reengineering is the ability to see potential candidates within an area. If we again think of the "bull's eye" effect within a smoothed surface, if this area has been driven by the Google Street View team (and thankfully at this point areas of sparse geography also tend to be the least covered), we could literally view each option within the central pixel until a house match is found. Even with multiple alternatives, it might be possible to spatially prioritise the potential buildings based on characteristics of the health conditions, or other information gleaned from the paper. For example, is the disease more typically associated with a multi-family unit than a single residence?

The second area of concern involves the use of biometric sensors synched to a GPS (Global Positioning System) unit. This field of research offers great potential in terms of linking health outcomes to the fine-scale built environment. However, a fear expressed at the URISA workshop was that output from these devices, usually shown as a series of dots on an aerial photograph, will begin to accompany research papers. Sure enough, within one day of the workshop a new issue of a GIS journal published this exact output. The underlying aerial photograph makes reengineering from the image extremely easy, and the point concentrations from the GPS unit correspond to areas of highest activity, including the home. This is not a good situation, especially when the participants are part of a vulnerable population, such as children.

Finally, we are worried about the current trend by social scientists of including spatial data in their research, especially those who use mixed methods. A mixed method approach combines both qualitative and quantitative data. For example, spatial video data of the recovering neighbourhoods of New Orleans, LA, USA, are currently being collected. These data are extracted from the video as three-dimensional surfaces that can be mapped or analysed for recovery or abandonment. At the same time, videos of the narrative of the neighbourhood participants add further commentary to the surfaces, such as why a building has not been returned to. Many of these comments contain sensitive information such as the health of an owner. If we map this information, others could easily disseminate it through online consumer geoinformatics services like Google Earth and Google Map, and even link it using suitable geo-mashups [51] to other readily available online information about the individuals concerned (e.g., on social networking sites), thus revealing a more detailed picture about them. Do our subjects really know all what could be exposed through such mapping? (But

one should also consider the difference between what is *technically possible* and what is *practically likely to happen*, i.e., will there really be someone with the motive, will and ability to do these privacy threatening Web inferencing and mapping exercises in each and every case? (A risks-costs-benefits assessment might help in such situations.) Although these situations may not fall foul of any HIPAA standard, nor probably concern an IRB, we are now at a point where changing geospatial technologies must stimulate debate that goes beyond the normal community participatory ethical standards used by researchers [35].

Because of the widespread adoption of GIS-light Internet applications, and cheap and easy-to-use mobile mapping devices (for example, ones which can tag pictures with coordinates), health related spatial confidentiality is now no longer the concern of only geographic information scientists, or even GIS users, but also of a far broader range of academics and other people.

**Conclusion**

Although the general public's concerns about privacy in research have sometimes been exaggerated by the scientific community [52] (and by a few vocal privacy advocates in the media, who do not adequately represent the position of the wider masses), we believe there are still many cases where these concerns are real and legitimate, and where data subjects need to be protected (e.g., from identity theft). A 'one-size-fits-all' privacy-preserving solution is unlikely to be successful or to be able to capture and properly address the complex requirements, which might also vary from country to country, of the very many (i) user roles, with different access privileges and 'needs to know' in relation to various input and output data types; (ii) intra and extra-mural data sharing arrangements, especially when data need to be moved across heterogeneous organisations; (iii) governing legislations and policies; (iv) possible forms of data inputs that can be released for research and the associated conditions; (v) health study types and goals, data analysis methods and the data requirements in each case; (vi) possible study outputs/results reporting and publication forms (closed or public); (vii) situation-specific security risks; and (viii) risks-costs-benefits assessment, among other aspects and requirements that are involved in this area of research and need to be considered on a case-by-case basis.

Different privacy-preserving solutions can be applied concurrently or singly on various elements of this complex chain, e.g., on input data prior to release to researchers (e.g., aggregation or transformations) and/or on the research outputs (e.g., access restriction or masking), depending on the specific situation at hand; so a comprehensive, context-aware approach is needed to assist

researchers in choosing and applying the right solution(s) in each case.

Kamel Boulos (unpublished research notes, 2008–2009) proposed the development of a case-based reasoning software framework (*cf.* case law) that covers, and continuously "learns from", the growing body of possible and emerging health research scenarios and applications involving precise geographical identifiers about individuals. The goal of such a 'one stop shop' framework would be to streamline the provision of clear and individualised guidance for the design and approval of new research projects, including crisp recommendations on which specific privacy-preserving solution(s) and approach(es) would be suitable in each case. This would spare researchers and IRBs the need to 'reinvent the wheel' with each new study, saving them precious time and efforts spent investigating the same issues every time, and preventing avoidable errors and omissions along the way. This decision framework should ideally have an easy-to-use, wizard-based visual frontend, guiding users throughout the whole process of describing and diagnosing their needs, and proposing (with appropriate explanations/justifications) suitable solutions to address them.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

MNKB conceived and drafted the manuscript, and conducted a mini-survey of the literature on the subject. AJC provided material on URISA's workshop held on 5 June 2009 about location privacy issues in health research. PA contributed material on 'personal information and privacy legislations', including 'Additional file 1'. All authors read and approved the final manuscript.

### Additional material

#### Additional file 1

A brief compilation and comparison of relevant personal information and privacy legislation in Canada and the UK, with particular focus on location and public health

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1476-072X-8-46-S1.pdf>]

### References

1. US CDC Public Health Law 101 Foundational Course for Public Health Practitioners – Unit 6: Privacy and Confidentiality [<http://www2a.cdc.gov/php/phi101/docs/PHL101-Unit%206%20-%2016Jan09-Secure.ppt>]
2. Ware W: **Lessons for the future: Privacy dimensions of medical record keeping.** In *Proceedings of the Conference on Health Records: Social Needs and Personal Privacy, Sponsored by the Department of Health and Human Services Task Force on Privacy, Office of the Assistant Secretary for Planning and Evaluation and the Agency for HealthCare Privacy and Research: 11–12 February 1993 (Document No. PB94-168192)* Washington, DC: US Government Printing Office; 1993:43-51.
3. Anderson R: **The devil is in the detail-A case in Finland on the privacy of medical records puts two major NHS systems in legal peril (Smart Healthcare – 1 April 2009).** [<http://www.smarthealthcare.com/anderson-database-01apr09>].
4. Anderson R, Brown I, Dowty T, Inglesant P, Heath W, Sasse A: *Database State 2009* [<http://www.jrrt.org.uk/uploads/Data%20base%20State.pdf>]. York, England: The Joseph Rowntree Reform Trust Ltd
5. **Secondary Uses Service (SUS) – NHS Connecting for Health** [<http://www.connectingforhealth.nhs.uk/systemsandservices/sus>]
6. Malheiros M: **Medical data secondary use issues (Privacy Value Networks – 10 June 2009).** [<http://www.pvnets.org/2009/06/medical-data-secondary-use-issues/>].
7. **Pension details of 109,000 stolen (BBC News – 28 May 2009)** [<http://news.bbc.co.uk/1/hi/business/8072524.stm>]
8. **ElcomSoft Distributed Password Recovery Software: High-performance distributed password recovery with NVIDIA GPU acceleration** [<http://www.elcomsoft.com/edpr.html>]
9. Kamel Boulos MN: **Descriptive review of geographic mapping of severe acute respiratory syndrome (SARS) on the Internet.** *Int J Health Geogr* 2004, **3**:2.
10. Woo RB: **Epidemics, Privacy Rights and Public Concerns: The Hong Kong SARS Experience.** *Workshop: Globalisation and New Epidemics: Ethics, Security and Policy Making, Organised by European Commission – Science and Society: 22–23 May 2006; Brussels, Belgium* [[http://www.pcpd.org.hk/english/files/infocentre/speech\\_20060522.pdf](http://www.pcpd.org.hk/english/files/infocentre/speech_20060522.pdf)].
11. Curtis AJ, Mills JW, Leitner M: **Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina.** *Int J Health Geogr* 2006, **5**:44.
12. Brownstein JS, Cassa CA, Kohane IS, Mandl KD: **An unsupervised classification method for inferring original case locations from low-resolution disease maps.** *Int J Health Geogr* 2006, **5**:56.
13. Cassa CA, Wieland SC, Mandl KD: **Re-identification of home addresses from spatial locations anonymized by Gaussian skew.** *Int J Health Geogr* 2008, **7**:45.
14. Van Wey LK, Rindfuss RR, Gutmann MP, Entwisle B, Balk DL: **Confidentiality and spatially explicit data: concerns and challenges.** *Proc Natl Acad Sci USA* 2005, **102(43)**:15337-15342.
15. Gutmann M, Witkowski K, Colyer C, O'Rourke JM, McNally J: **Providing Spatial Data for Secondary Analysis: Issues and Current Practices relating to Confidentiality.** *Popul Res Policy Rev* 2008, **27(6)**:639-665.
16. Werneck GL: **Georeferenced data in epidemiologic research.** *Cien Saude Colet* 2008, **13(6)**:1753-66.
17. Cassa CA: **Privacy and identifiability in clinical research, personalized medicine, and public health surveillance.** *PhD thesis* 2008 [<http://hdl.handle.net/1721.1/45624>]. Harvard University–MIT Division of Health Sciences and Technology
18. Sherman JE, Fetters TL: **Confidentiality concerns with mapping survey data in reproductive health research.** *Stud Fam Plann* 2007, **38(4)**:309-21.
19. Siffel C, Strickland MJ, Gardner BR, Kirby RS, Correa A: **Role of geographic information systems in birth defects surveillance and research.** *Birth Defects Res A Clin Mol Teratol* 2006, **76(11)**:825-33.
20. Matthews SA, Moudon AV, Daniel M: **Work group II: Using Geographic Information Systems for enhancing research relevant to policy on diet, physical activity, and weight.** *Am J Prev Med* 2009, **36(4 Suppl)**:S171-6.
21. Smolders R, Casteleyn L, Joas R, Schoeters G: **Human biomonitoring and the INSPIRE directive: spatial data as link for environment and health research.** *J Toxicol Environ Health B Crit Rev* 2008, **11(8)**:646-59.
22. Foley R: **Assessing the applicability of GIS in a health and social care setting: planning services for informal carers in East Sussex, England.** *Soc Sci Med* 2002, **55(1)**:79-96.
23. Armstrong MP, Rushton G, Zimmerman DL: **Geographically masking health data to preserve confidentiality.** *Stat Med* 1999, **18(5)**:497-525.
24. Cassa CA, Grannis SJ, Overhage JM, Mandl KD: **A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection.** *J Am Med Inform Assoc* 2006, **13(2)**:160-5.

25. Kamel Boulos MN, Cai Q, Padget JA, Rushton G: **Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses.** *J Biomed Inform* 2006, **39(2)**:160-70.
26. Wieland SC, Cassa CA, Mandl KD, Berger B: **Revealing the spatial distribution of a disease while preserving privacy.** *Proc Natl Acad Sci USA* 2008, **105(46)**:17608-13.
27. Olson KL, Grannis SJ, Mandl KD: **Privacy protection versus cluster detection in spatial epidemiology.** *Am J Public Health* 2006, **96(11)**:2002-8.
28. Ozonoff A, Jeffery C, Manjourides J, White LF, Pagano M: **Effect of spatial resolution on cluster detection: a simulation study.** *Int J Health Geogr* 2007, **6**:52.
29. Snow J: *On the Mode of Communication of Cholera* 2nd edition. :1855 [<http://www.ph.ucla.edu/EPI/snow/snowbook.html>]. London, England: John Churchill
30. Onsrud HJ, Johnson JP, Lopez X: **Protecting Personal Privacy in Using Geographic Information Systems.** *Photogrammetric Engineering and Remote Sensing* 1994, **60(9)**1083-1095 [<http://www.spatial.maine.edu/~onsrud/tempe/onsrud.html>].
31. **Google Latitude** [<http://www.google.com/latitude/>]
32. **Microsoft Vine** [<http://www.vine.net/>]
33. **Yahoo! Fire Eagle** [<http://fireeagle.yahoo.net/>]
34. Mokbel MF: **Privacy in Location-Based Services: State-of-the-Art and Research Directions.** In *Proceedings of the 8th International Conference on Mobile Data Management (MDM'07): 7-11 May 2007; Mannheim, Germany* IEEE; 2007:228-228. DOI: 10.1109/MDM.2007.45
35. Kamel Boulos MN: **Chapter 49: Principles and techniques of interactive Web cartography and Internet GIS.** *Manual of Geographic Information Systems* 2009:935-974 [[http://www.asprs.org/gis\\_manual/index.html](http://www.asprs.org/gis_manual/index.html)]. Bethesda, Maryland: ASPRS—American Society for Photogrammetry and Remote Sensing ISBN: 1-57083-086-X
36. **Microsoft Windows BitLocker Drive Encryption** [[http://technet.microsoft.com/en-us/library/cc766200\(VS.10\).aspx](http://technet.microsoft.com/en-us/library/cc766200(VS.10).aspx)]
37. **Securing Sensitive Information with Identity and Access Assurance (RSA/Courion White Paper)** [[http://www.rsa.com/solutions/IA/wp/10292\\_RSA-Courion\\_WP\\_0609.pdf](http://www.rsa.com/solutions/IA/wp/10292_RSA-Courion_WP_0609.pdf)]
38. **IronKey: Secure USB Flash Drive with Internet Protection Services** [<https://www.ironkey.com/>]
39. **Integral Crypto Drive** [<http://www.integralmemory.com/crypto/?gclid=CMWV528i5ipsCFU0B4wod7GSUoA>]
40. **SDelete** [<http://technet.microsoft.com/en-us/sysinternals/bb897443.aspx>]
41. El Emam K, Neri E, Jonker E: **An Evaluation of Personal Health Information Remnants in Second-Hand Personal Computer Disk Drives.** *J Med Internet Res* 2007, **9(3e24)** [<http://www.jmir.org/2007/3/e24>].
42. AbdelMalik P, Kamel Boulos MN, Jones R: **The perceived impact of location privacy: a web-based survey of public health perspectives and requirements in the UK and Canada.** *BMC Public Health* 2008, **8**:156.
43. **GeoPKDD – Geographic Privacy-Aware Knowledge Discovery and Delivery** [<http://www.geopkdd.eu/>]
44. **First Interdisciplinary Workshop on Mobility, Data Mining and Privacy: Preserving anonymity in geographically referenced data: 14 February 2008; Rome, Italy** [<http://wiki.kdu.big.org/mobileDMprivacyWorkshop/>]
45. **HCLS Patient Data Security and Privacy** [<http://esw.w3.org/topic/HCLS/SecurityPrivacy>]
46. **URISA – The Association for GIS Professionals** [<http://www.urisa.org/>]
47. El Emam K, Brown A, AbdelMalik P: **Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk.** *J Am Med Inform Assoc* 2009, **16**:256-266.
48. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data. National Research Council: *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data* Washington, DC: The National Academies Press; 2007.
49. Brownstein JS, Cassa CA, Mandl KD: **No place to hide—reverse identification of patients from published maps.** *N Engl J Med* 2006, **355(16)**:1741-2.
50. **Google Street View** [<http://maps.google.com/help/maps/streetview/>]
51. Kamel Boulos MN, Scotch M, Cheung KH, Burden D: **Web GIS in practice VI: a demo playlist of geo-mashups for public health neogeographers.** *Int J Health Geogr* 2008, **7**:38.
52. Tondel M, Axelson O: **Concerns about privacy in research may be exaggerated.** *BMJ* 1999, **319(7211)**:706-7.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)



# **A brief compilation and comparison of relevant personal information and privacy legislation in Canada and the UK, with particular focus on location and public health**

*Philip AbdelMalik - philip.abdelmalik@plymouth.ac.uk*

A discussion on location privacy solutions for health research would be incomplete without reflecting on some of the underlying reasons that necessitate their development. The very notion of privacy is itself a complex fabric of interwoven philosophical and psychosocial threads. Perhaps this is why the associated bureaucratic and legal landscape is as complex as it is – and often blamed for the issue. A large majority of public health professionals consider privacy to be an obstacle to public health; when asked for the underlying reasons, survey respondents in Canada and the UK most commonly identified bureaucracy and legislation [1].

There is no universal legislation to guide and govern the activities of public health professionals, particularly where issues of privacy are concerned. Instead, nations have their own constraining or enabling privacy and data protection laws, with some being such a maze of cross-referenced “legalese” that familiarising oneself with them – let alone gaining a thorough understanding of them – becomes a daunting task. What ensues is a brief compilation and comparison of relevant personal information and privacy legislation in Canada and the United Kingdom (UK), with particular focus on location and public health as seen and understood by an epidemiologist.

## **Overview**

The Canadian privacy-legislation landscape is additionally muddled by its political system: ten provinces and three territories, each with its own legislation and jurisdiction over its own health system. Overarching is the federal government, providing guidelines, support, oversight and funding. Although the words “privacy” and “personal information” do not occur anywhere in Canada’s Constitution (Charter of Rights and Freedoms) [2], Section 7, granting the right to life, liberty and security, and Section 8, guaranteeing protection from unreasonable search and seizure, have been determined by the courts to capture the right to privacy [3,4]. These cases have expanded on the Charter sections to include privacy as related to protection from government or other intrusion, autonomy, and dignity.

Federally, Canada has two privacy laws. The *Privacy Act* [5] governs roughly 160 federal public bodies, whereas the *Personal Information and Protection of Electronic Documents Act* (PIPEDA) [6] governs private sector organisations regulated federally and provincially. Provinces with privacy legislation similar to PIPEDA are exempt from its provincial aspect. At the time of writing, British Columbia, Alberta and Québec have such legislation, and Ontario has health-specific legislation that exempts it from the corresponding section.

All provinces and territories have legislation similar to the *Privacy Act*, whereas only three provinces have private-sector legislation similar to PIPEDA. In addition, four provinces have specific health *information* legislation: Alberta, Manitoba, Ontario and Saskatchewan.

The UK has three legal jurisdictions: England and Wales, Scotland and Northern Ireland. However, it itself is also part of a larger community - the European Union (EU). European Union legislation is generally intended to “direct” that of its member states, and takes precedence in cases where there is no concurrence; the UK is obligated to align itself with EU law (referred to as Community law) [7] or else give way in a court of law to the latter [8]. Let us therefore begin with the EU.

The concepts of privacy and personal information are captured in core EU legislative documents as fundamental rights. The *European Convention for the Protection of Human Rights and Fundamental Freedoms* (ECHR), building on the 1948 *Universal Declaration of Human Rights* [9], includes a “Right to respect for private and family life” in Article 8 [10]. The *Charter of Fundamental Rights of the European Union*, proclaimed in 2000, builds on the ECHR [11]. Updated in 2007, the Charter includes two particularly relevant articles. Article 7 reiterates the ECHR’s position on the respect for private and family

life, whereas Article 8 explicitly limits the processing of personal data to specified purposes, requiring either individual consent or legislated “permission”.

Recognising the importance of data sharing and the threats and benefits of developing technologies, the EU introduced a number of legislative pieces to harmonise, regulate and facilitate the flow of personal information. In 1995, *Directive 95/46/EC* was adopted for the protection of personal data [12] - the core directive at the heart of data protection in EU member states. It does not, however, apply, to personal information used solely for personal reasons, household activities, public security, national defence or criminal law enforcement, and falls short when dealing with issues around communication. Two years later, the EU adopted *Directive 97/66/EC* for protecting privacy and confidentiality in telecommunications [13]. As technology and the web became increasingly ubiquitous, this directive quickly became limited in scope. It was therefore replaced in 2002 by *Directive 2002/58/EC* [14] covering electronic communications more broadly, and updated again in 2006 by *Directive 2006/24/EC* [15]. In addition, *Data Protection Regulation (EC) 45/2001* [16] ensures the protection of personal information in EU institutions and bodies, such as the European Parliament, for example, and accountability to a governing body, the European Data Protection Supervisor.

In the UK, the *Data Protection Act* was first enacted on July 12, 1984, thereby preceding the *Directive on Data Protection* adopted by the European Union (EU) by more than a decade. Upon adoption of the EU directive, however, the Act was amended in 1998. Though simpler than Canadian legislation in the sense that it applies to both public and private entities, it is none-the-less a complex document. In 2003, Lord Phillips of the Supreme Court of Judicature, Court of Appeal (Civil Division) in the UK referred to it as “...a cumbersome and inelegant piece of legislation” [17]. Other UK health-related Acts have been amended to reference the *Data Protection Act 1998*, including the *Access to Health Records Act 1990*, the *Access to Medical Reports Act 1988* and the *Access to Personal Files and Medical Reports (Northern Ireland)*. The UK also has a *Health and Social Care Act 2008* [18], which replaced its 2001 predecessor and legislated the creation of a Care Quality Commission for the protection and promotion of the health, safety and welfare of the public. The Act makes it an offence to recklessly disclose confidential personal information obtained by the Commission that “relates to and identifies an individual.” (S. 76)

Scotland has a *Freedom of Information Act 2002*, but a search on the UK Office of Public Sector Information website [19] yielded no specific data protection legislation for either Scotland or Northern Ireland. Scotland also has a *Public Health Act* enacted just last year, in 2008 [20], which obligates Scottish Ministers, health boards and local authorities to protect public health. It allows for the disclosure of information to facilitate its directives despite any other legal prohibition or restriction, except, interestingly, the *Data Protection Act 1998* (S. 117(6)). Northern Ireland’s *Health and Social Care (Reform) Act 2009* [21] has a similar clause (S. 13(8)).

Both Canada and the UK have a tapestry of legislative documents in place to protect the privacy of personal information “...as something worth protecting as an aspect of human autonomy and dignity.” [22] But what, exactly, constitutes personal information?

## Definitions

There is no consistent definition for “personal information” in Canadian legislation. Where a definition is included, it ranges from “information about an identifiable individual” in Alberta’s *Personal Information Protection Act* [23] to very well-defined and explicit components in Manitoba’s *Freedom of Information and Protection of Privacy Act* [24]. Of the 30 acts and regulations reviewed, four include health information in their definition of personal information, three include location information, 14 include both and nine include neither (Table 1).

This definition of personal information as pertaining to an “identifiable individual” appears quite often in legislation, including in *Directive 95/46/EC*. However, the *Directive* goes one step further to clarify: “...an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity” [12]. Health information is defined as a “special” category of personal information (S. III, Article 8 (1)), but there is no specific mention of location information in the *Directive*.

In the UK, the *Data Protection Act 1998* defines "personal data" vaguely as any information that, in isolation or in concert with other data available to the data controller, can identify a living individual. The *Act* also includes health in the definition of "sensitive personal data", but does not capture location information specifically. As mentioned previously, the *Health and Social Care Act 2008* also identifies confidential personal information as that which "relates to and identifies and individual", but does not specifically identify location as part of that definition.

As recent as April 2009, the Supreme Court of Canada stated that "Privacy analysis is laden with value judgements that are made from the independent perspective of the reasonable and informed person who is concerned about the long-term consequences of government action for the protection of privacy" [25]. As described, the definition of "personal information" in most cases casts a wide net, capturing anything and everything that can subjectively be argued as identifying. This has obvious implications on the use of disaggregate geographic data in health research. Or does it? The answer depends on the applications and exceptions made in the legislation.

### **Application and exceptions**

Legislation in Canada, the EU and the UK specifically limits the processing of personal information. What constitutes "processing", however, is not consistently defined across legislation. The broadest definition to capture what this means is found in EU *Directive 95/46/EC*: "any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organisation, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction". Generally, any such processing of personal information is prohibited *in the absence of the individual's informed consent*, unless it is first stripped of all identifying information (thereby ceasing to be personal information according to the legal definition).

In public health research, however, it is often impossible or impractical to pursue informed consent. Despite being incredibly information and data-rich, health researchers in both Canada and the UK have often expressed frustration over their inability to use existing data due to privacy concerns [1]. Is the prohibition based on the legislation?

Generally, in the absence of an individual's consent, the legislation does explicitly allow for some exceptions, particularly in the interests of national security. However, there is a lack of clarity and consistency, specifically around processing for public health purposes. Article 35 of the *Charter of Fundamental Rights of the European Union* emphasises the right to health care, and states "A high level of human health protection shall be ensured in the definition and implementation of all Union policies and activities" [11]. In almost all cases, exceptions are also made for research, as long as the individuals whose data is processed are not identified in the results. Generally, the individual whose information has been disclosed should be informed; however, provisions are also made for cases where doing so is impossible or unreasonable.

The decision around whether or not the processing of the information is permitted under these exceptions is somewhat vague and inconsistent. In Canada, for example, the four provinces with health information legislation delegate the decision making authority to research ethics boards; otherwise, it is generally delegated to the head of the data-holding organisation. In the case of EU institutions, processing is only permissible after consultation with the European Data Protection Supervisor [16], whereas the UK *Data Protection Act 1998* exception for research (S. 4(33)) is unclear as to the decision-making authority. This leads to issues around governance.

### **Governance**

In Canada, the Office of the Privacy Commissioner (OPC) is responsible for protecting and promoting the privacy rights of Canadians by overseeing compliance with Canadian federal privacy legislation. Each province and territory also has its own privacy commissioners who oversee their respective jurisdictions. As previously noted, health information legislation in Alberta, Saskatchewan, Manitoba and Ontario also delegates decision-making authority on these matters to research ethics boards.



The EU, as previously mentioned, has established the office of the European Data Protection Supervisor [26] for oversight of EU institution activities. The UK's equivalent of Canada's Office of the Privacy Commissioner is the Information Commissioner's Office (ICO) [27]. The legislation does not specifically mention research ethics boards or committees, and is unclear as to decision-making authority – in most cases, it seems to lie with the data controllers.

### **Implications and final thoughts**

The privacy of personal information is a recognised and important human right, protected through multiple intertwined acts and regulations in Canada, the EU and the UK. In the absence of informed consent, the legislation generally allows for the processing of an individual's personal information – which is any information that can identify the individual, and therefore includes health and disaggregate location information – for research purposes, subject to approval by the appropriate authority. However, guidelines are lacking, and authorities tend to err on the conservative side, resulting in much expressed frustration by health researchers. In the absence of frameworks to inform the processing of personal information, the only other alternative (besides seeking informed consent from every individual) for health researchers is the use of de-identification techniques, such as might be applied through privacy-preserving solutions involving disaggregate geographic data.

It has been suggested that privacy in the United States, Canada and the European Union have their bases in slightly different philosophical constructs: in the United States, privacy is anchored in protection from the government; in Canada, in principles of autonomy and control; and in the European Union, the focus is more on dignity and public image [28]. The argument is made that the Canadian model offers the appropriate “middle-ground” – after all, if individuals truly do have control over their own personal information, then they can choose to protect it from the government and others, and their dignity as far as public image is concerned is in their own hands. If we accept this definition of privacy – that is, having control over one's own personal information – then one might ask whether de-identification really solves the issue. Perhaps what is really needed is public health specific clarification in the legislation, public and practitioner education, and clear and concise frameworks and guidelines.

Public health practitioners around the world are increasingly recognising the importance of having some understanding of the legal system, and a working relationship with the legal profession [29]. Unfortunately, the relationship typically tends to be unidirectional. Just as privacy is a multifaceted and complex concept, so too is the required collaboration resulting from the interdependency of public health and legislation. And yet, the legal profession has not fully recognised the interdependence of the two fields [29]. While the privacy debate in public health may be fuelled in part by misperceptions of public health practitioners, it is very much coupled with a lack of understanding of the requirements of public health by legal practitioners. “Privacy laws are most burdensome and least effective when they apply broadly, without proper concern for the settings in which they operate, the types of information that they cover, the obligations that they impose and the purposes they were designed to serve” [30]. The issue can only be truly addressed through interdisciplinary collaboration. Until that happens, and until we recognise the importance and value of public health research and its implications on the health of individuals, we will continue to grapple with alternate de-identification solutions and sub-optimal data.

**Table 1: Inclusion of health and location information in the definitions of "personal information" in Canadian legislation**

Jurisdiction	Act	Reference	In Definition	
			Health	Location
Canada	The Privacy Act [5]	R.S.C. 1985, c. P-21	✓	✓
Canada	Personal Information Protection and Electronic Documents Act [6]	S.C. 2000, c. 5 P-8.6	✓	
B.C.	Freedom of Information and Protection of Privacy Act [31]	R.S.B.C. 1996, c. 165		
B.C.	Personal Information Protection Act [32]	S.B.C. 2003, c. 63		
B.C.	Freedom of Information and Protection of Privacy Regulation [33]	B.C. Reg 323/93		
B.C.	Personal Information Protection Act Regulations [34]	B.C. Reg. 473/2003		✓
B.C.	British Columbia Cancer Agency Research Information Regulation [35]	B.C. Reg. 286/91	✓	✓
B.C.	Privacy Act [36]	R.S.B.C. 1996, c. 373		
AB	Health Information Act [37]	R.S.A. 2000, c. H-5	✓	✓
AB	Freedom of Information and Protection of Privacy Act [38]	R.S.A. 2000, c. F-25	✓	✓
AB	Personal Information Protection Act [23]	S.A. 2003 c. P-6.5		
AB	Personal Information Protection Act Regulation [39]	AR 366/2003		✓
SK	The Health Information Protection Act [40]	S.S. 1999, c. H-0.021	✓	
SK	The Freedom of Information and Protection of Privacy Act [41]	SS. 1990-91, c. F-22.01		✓
SK	The Local Authority Freedom of Information and Protection of Privacy Act [42]	SS. 1990-91, c. L-27.1	✓	✓
MB	The Personal Health Information Act [43]	C.C.S.M., c. P-33.5	✓	
MB	The Freedom of Information and Protection of Privacy Act [24]	C.C.S.M., c. F-175	✓	✓
ON	Personal Health Information Protection Act [44]	S.O. 2004, c. 3	✓	
ON	Freedom of Information and Protection of Privacy Act [45]	R.S.O. 1990, c. F-31	✓	✓
ON	Municipal Freedom of Information and Protection of Privacy Act [46]	R.S.O. 1990, c. M.56	✓	✓
QC	An Act respecting Access to documents held by public bodies and the protection of personal information [47]	R.S.Q., c. A-2.1		
QC	An Act respecting the Protection of personal information in the private sector [48]	R.S.Q., c. P-39.1		
N.B.	Protection of Personal Information Act [49]	S.N.B. 1998, c. P-19.1		
N.S.	Freedom of Information and Protection of Privacy Act [50]	S.N.S. 1993, c. 5, s. 1	✓	✓
N.S.	Health Protection Act [51]	S.N.S. 2004, c. 4, s. 1		
P.E.I.	Freedom of Information and Protection of Privacy Act [52]	R.S.P.E.I. 1988, c. F-15.01	✓	✓
NL	Access to Information and Protection of Privacy Act [53]	S.N.L. 2002, c. A-1.1	✓	✓
YK	Access to Information and Protection of Privacy Act [54]	R.S.Y. 2002, c. 1	✓	✓
N.T.	Access to Information and Protection of Privacy Act [55]	S.N.W.T. 1994, c. 20	✓	✓
NU	Access to Information and Protection of Privacy Act [56]	S.N.W.T. 1994, c.20	✓	✓

## References

1. AbdelMalik P, Kamel Boulos MN, Jones R: The perceived impact of location privacy: a web-based survey of public health perspectives and requirements in the UK and Canada. *BMC Public Health* (2008), 8:156. [<http://www.biomedcentral.com/1471-2458/8/156>]
2. Canadian Charter of Rights and Freedoms. (March 1982)  
[<http://laws.justice.gc.ca/en/charter/1.html>]
3. R. v. Morgentaler, [1988] 1 S.C.R. 30. Supreme Court of Canada: January 28, 1988  
[<http://scc.lexum.umontreal.ca/en/1988/1988rcs1-30/1988rcs1-30.html>]
4. Hunter et al. v. Southam Inc., [1984] 2 S.C.R. 145. Supreme Court of Canada: September 17, 1984  
[<http://csc.lexum.umontreal.ca/en/1984/1984rcs2-145/1984rcs2-145.html>]
5. The Privacy Act, R.S.C. 1985, c. P-21. (1985)  
[[http://laws.justice.gc.ca/en/P-21/section-\[section-no\].html](http://laws.justice.gc.ca/en/P-21/section-[section-no].html)]
6. Personal Information Protection and Electronic Documents Act, S.C. 2000, c. 5 P-8.6. (2000)  
[[http://laws.justice.gc.ca/en/P-8.6/section-\[section-no\].html](http://laws.justice.gc.ca/en/P-8.6/section-[section-no].html)]
7. European Communities Act 1972. (1972)
8. European Judicial Network in civil and commercial matters: Legal Order - England and Wales. Last Updated: 19-8-2004; Accessed: 5-7-2009  
[[http://ec.europa.eu/civiljustice/legal\\_order/legal\\_order\\_eng\\_en.htm](http://ec.europa.eu/civiljustice/legal_order/legal_order_eng_en.htm)]
9. Universal Declaration of Human Rights. (December 1948)  
[<http://www.un.org/Overview/rights.html>]
10. Convention for the Protection of Human Rights and Fundamental Freedoms as amended by Protocol No. 11. (September 2003)  
[<http://www.echr.coe.int/NR/rdonlyres/D5CC24A7-DC13-4318-B457-5C9014916D7A/0/EnglishAnglais.pdf>]
11. Charter of Fundamental Rights of the European Union (2000/C 364/01). (2000)  
[<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2000:364:0001:0022:EN:PDF>]
12. Directive 95/46/EC of the European Parliament and of the council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. (November 1995)  
[<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML>]
13. Directive 97/66/EC of the European Parliament and of the Council of 15 December 1997 concerning the processing of personal data and the protection of privacy in the telecommunications sector. (January 1998)  
[<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:1998:024:0001:0008:EN:PDF>]
14. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). (July 2002)  
[<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:NOT>]
15. Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC. (May 2006)  
[<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006L0024:EN:NOT>]
16. Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the

Community institutions and bodies and on the free movement of such data. (January 2001)  
<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:008:0001:0022:EN:PDF>

17. Naomi Campbell and MGN Limited, [2002] EWCA Civ 1373. Supreme Court of Judicature, Court of Appeal (Civil Division): October 14, 2002  
<http://www.bailii.org/ew/cases/EWCA/Civ/2002/1373.html>
18. Health and Social Care Act 2008 (c. 14). (2008)  
[http://www.opsi.gov.uk/acts/acts2008/ukpga\\_20080014\\_en\\_1](http://www.opsi.gov.uk/acts/acts2008/ukpga_20080014_en_1)
19. Office of Public Sector Information. Last Updated: 16-6-2009; Accessed: 2009  
<http://www.opsi.gov.uk/>
20. Public Health etc. (Scotland) Act 2008 (asp 5). (2008)  
[http://www.opsi.gov.uk/legislation/scotland/acts2008/asp\\_20080005\\_en\\_1](http://www.opsi.gov.uk/legislation/scotland/acts2008/asp_20080005_en_1)
21. Health and Social Care (Reform) Act (Northern Ireland) 2009 (c. 1). (2009)  
[http://www.opsi.gov.uk/legislation/northernireland/acts/acts2009/nia\\_20090001\\_en\\_1](http://www.opsi.gov.uk/legislation/northernireland/acts/acts2009/nia_20090001_en_1)
22. Campbell v. MGN Limited, [2004] UKHL 22. House of Lords Appellate Committee: May 6, 2004  
<http://www.bailii.org/uk/cases/UKHL/2004/22.html>
23. Personal Information Protection Act, S.A. 2003, c. P-6.5. (2003)  
<http://www.canlii.org/en/ab/laws/stat/sa-2003-c-p-6.5/latest>
24. Freedom of Information and Protection of Privacy Act, C.C.S.M. c. F-175. (1997)  
<http://www.canlii.org/mb/laws/sta/f-175/20090324/whole.html>
25. R. v. Patrick, [2009] S.C.C. 17. Supreme Court of Canada: April 9, 2009  
<http://scc.lexum.umontreal.ca/en/2009/2009scc17/2009scc17.html>
26. European Data Protection Supervisor. Last Updated: 2009; Accessed: 2009  
<http://www.edps.europa.eu/EDPSWEB/>
27. Information Commissioner's Office. Last Updated: 2009; Accessed: 2009  
<http://www.ico.gov.uk/>
28. Levin A, Nicholson MJ: Privacy law in the United States, the EU and Canada: the allure of the middle ground. *University of Ottawa Law and Technology Journal* (2005), 2(2):357-395.
29. Hoffman RE, Lopez W, Matthews GW, Rothstein MA, Foster KL: Law in Public Health Practice. New York: Oxford University Press, 2007
30. Cate FH: Privacy in perspective. Washington, D.C.: The AEI Press, 2001
31. Freedom of Information and Protection of Privacy Act, R.S.B.C. 1996, c. 165. (1996)  
[http://www.qp.gov.bc.ca/statreg/stat/F/96165\\_01.htm](http://www.qp.gov.bc.ca/statreg/stat/F/96165_01.htm)
32. Personal Information Protection Act, S.B.C. 2003, c. 63. (2003)  
[http://www.qp.gov.bc.ca/statreg/stat/P/03063\\_01.htm](http://www.qp.gov.bc.ca/statreg/stat/P/03063_01.htm)
33. Freedom of Information and Protection of Privacy Regulation, B.C. Reg 323/93. (1993)  
[http://www.qp.gov.bc.ca/statreg/reg/F/323\\_93.htm](http://www.qp.gov.bc.ca/statreg/reg/F/323_93.htm)
34. Personal Information Protection Act Regulations, B.C. Reg. 473/2003. (2003)  
[http://www.qp.gov.bc.ca/statreg/reg/P/473\\_2003.htm](http://www.qp.gov.bc.ca/statreg/reg/P/473_2003.htm)
35. British Columbia Cancer Agency Research Information Regulation, B.C. Reg. 286/91. (1991)  
[http://www.qp.gov.bc.ca/statreg/reg/h/health/286\\_91.htm](http://www.qp.gov.bc.ca/statreg/reg/h/health/286_91.htm)
36. Privacy Act, R.S.B.C. 1996, c. 373. (1996) [http://www.qp.gov.bc.ca/statreg/stat/P/96373\\_01.htm](http://www.qp.gov.bc.ca/statreg/stat/P/96373_01.htm)
37. Health Information Act, R.S.A. 2000, C. H-5. (2000)  
[http://www.assembly.ab.ca/HIARReview/Health\\_Information\\_Act.pdf](http://www.assembly.ab.ca/HIARReview/Health_Information_Act.pdf)

38. Freedom of Information and Protection of Privacy Act, R.S.A. 2000, c. F-25. (2000)  
[\[http://www.iijcan.org/ab/laws/sta/f-25/20080818/whole.html\]](http://www.iijcan.org/ab/laws/sta/f-25/20080818/whole.html)
39. Personal Information Protection Act Regulation, AR 366/2003. (2003)  
[\[http://www.qp.gov.ab.ca/documents/Regs/2003\\_366.cfm?frm\\_isbn=0779725050\]](http://www.qp.gov.ab.ca/documents/Regs/2003_366.cfm?frm_isbn=0779725050)
40. Health Information Protection Act, S.S. 1999, c. H-0.021. (1999)  
[\[http://www.qp.gov.sk.ca/documents/english/Statutes/Statutes/H0-021.pdf\]](http://www.qp.gov.sk.ca/documents/english/Statutes/Statutes/H0-021.pdf)
41. The Freedom of Information and Protection of Privacy Act, SS. 1990-91, c. F-22.01. (1990)  
[\[http://www.qp.gov.sk.ca/documents/English/Statutes/Statutes/F22-01.pdf#\]](http://www.qp.gov.sk.ca/documents/English/Statutes/Statutes/F22-01.pdf#)
42. The Local Authority Freedom of Information and Protection of Privacy Act, SS. 1990-91, c. L-27.1. (1990) [\[http://www.qp.gov.sk.ca/documents/English/Statutes/Statutes/L27-1.pdf\]](http://www.qp.gov.sk.ca/documents/English/Statutes/Statutes/L27-1.pdf)
43. The Personal Health Information Act, C.C.S.M. c. P-33.5. (1997)  
[\[http://web2.gov.mb.ca/laws/statutes/ccsm/p033-5e.php\]](http://web2.gov.mb.ca/laws/statutes/ccsm/p033-5e.php)
44. Personal Health Information Protection Act, S.O. 2004, c. 3. (2004)  
[\[http://www.e-laws.gov.on.ca/html/statutes/english/elaws\\_statutes\\_04p03\\_e.htm\]](http://www.e-laws.gov.on.ca/html/statutes/english/elaws_statutes_04p03_e.htm)
45. Freedom of Information and Protection of Privacy Act, R.S.O. 1990, c. F-31. (1990)  
[\[http://www.e-laws.gov.on.ca/html/statutes/english/elaws\\_statutes\\_90f31\\_e.htm\]](http://www.e-laws.gov.on.ca/html/statutes/english/elaws_statutes_90f31_e.htm)
46. Municipal Freedom of Information and Protection of Privacy Act, R.S.O. 1990, c. M.56. (1990)  
[\[http://www.e-laws.gov.on.ca/html/statutes/english/elaws\\_statutes\\_90m56\\_e.htm\]](http://www.e-laws.gov.on.ca/html/statutes/english/elaws_statutes_90m56_e.htm)
47. An Act respecting access to documents held by public bodies and the protection of personal information, R.S.Q., c. A-2.1. (1982)  
[\[http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/A\\_2\\_1/A2\\_1\\_A.html\]](http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/A_2_1/A2_1_A.html)
48. An Act respecting the protection of personal information in the private sector, R.S.Q., c. P-39.1. (1993)  
[\[http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/P\\_39\\_1/P39\\_1\\_A.html\]](http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/P_39_1/P39_1_A.html)
49. Protection of Personal Information Act, S.N.B. 1998, c. P-19.1. (1998)  
[\[http://www.gnb.ca/acts/acts/p-19-1.htm\]](http://www.gnb.ca/acts/acts/p-19-1.htm)
50. Freedom of Information and Protection of Privacy Act, S.N.S. 1993, c. 5, s. 1. (1993)  
[\[http://www.gov.ns.ca/legislature/legc/statutes/freedom.htm\]](http://www.gov.ns.ca/legislature/legc/statutes/freedom.htm)
51. Health Protection Act, S.N.S. 2004, c. 4, s. 1. (2004)  
[\[http://www.gov.ns.ca/legislature/legc/statutes/healthpr.htm\]](http://www.gov.ns.ca/legislature/legc/statutes/healthpr.htm)
52. Freedom of Information and Protection of Privacy Act, R.S.P.E.I. 1988, c. F-15.01. (1988)  
[\[http://www.gov.pe.ca/law/statutes/pdf/f-15\\_01.pdf\]](http://www.gov.pe.ca/law/statutes/pdf/f-15_01.pdf)
53. Access to Information and Protection of Privacy Act, S.N.L. 2002, c. A-1.1. (2002)  
[\[http://www.assembly.nl.ca/legislation/sr/statutes/a01-1.htm\]](http://www.assembly.nl.ca/legislation/sr/statutes/a01-1.htm)
54. Access to Information and Protection of Privacy Act, R.S.Y. 2002, c. 1. (2002)  
[\[http://www.gov.yk.ca/legislation/acts/atipp.pdf\]](http://www.gov.yk.ca/legislation/acts/atipp.pdf)
55. Access to Information and Protection of Privacy Act, S.N.W.T. 1994, c. 20. (1994)  
[\[http://www.justice.gov.nt.ca/pdf/ACTS/Access\\_to\\_Information.pdf\]](http://www.justice.gov.nt.ca/pdf/ACTS/Access_to_Information.pdf)
56. Access to Information and Protection of Privacy Act, S.N.W.T. 1994, c. 20. (1994)  
[\[http://action.attavik.ca/home/justice-gn/attach-en\\_conlaw\\_prediv/Type002.pdf\]](http://action.attavik.ca/home/justice-gn/attach-en_conlaw_prediv/Type002.pdf)

# A Method for Aggregating Small Geographic Areas to Protect Privacy

19<sup>th</sup> March 2009

**Khaled El Emam**  
*CHEO Research Institute*

**Fida Dankar**  
*CHEO Research Institute*

**Philip AbdelMalik**  
*Public Health Agency of Canada*



## **Document Information**

**Document Title:** A Method for Aggregating Small Geographic Areas to Protect Privacy

**Original Document Date:** 13<sup>th</sup> March 2009

**Document Version:** Version 3

**Copyright:** CHEO Research Institute, 401 Smyth Road, Ottawa, ON K1H 8L1, Canada

**Contact:** Khaled El Emam (kelemam@ehealthinformation.ca)

**More Information:** <http://www.ehealthinformation.ca/>

## **Other Relevant Publications and Reports**

- K. El Emam, A. Brown, and P. AbdelMalik: "Evaluating predictors of geographic area population size cutoffs to manage re-identification risk." In *Journal of the American Medical Informatics Association*, March/April, 2009.
- K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk: "Evaluating Patient Re-identification Risk from Hospital Prescription Records." In the *Canadian Journal of Hospital Pharmacy*, June 2009.
- K. El Emam: "Heuristics for de-identifying health data." In *IEEE Security and Privacy*, July/August, 6(4):58-61, 2008.
- K. El Emam, and F. Dankar: "Protecting privacy using k-anonymity." In the *Journal of the American Medical Informatics Association*, September/October, 15:627-637, 2008.
- K. El Emam, E. Neri, and E. Jonker: "An evaluation of personal health information remnants in second hand personal computer disk drives." In *Journal of Medical Internet Research*, 9(3):e24, 2007.
- K. El Emam, S. Jabbouri, S. Sams, Y. Drouet, M. Power: "Evaluating common de-identification heuristics for personal health information." In *Journal of Medical Internet Research*, 2006;8(4):e28, November 2006.
- K. El Emam: "Overview of Factors Affecting the Risk of Re-Identification in Canada", Access to Information and Privacy, Health Canada, May 2006.
- K. El Emam: "Data Anonymization Practices in Clinical Research: A Descriptive Study", Access to Information and Privacy, Health Canada, May 2006.

*More information is available from*  
<http://www.ehealthinformation.ca/>

## Table of Contents

<b>ABSTRACT .....</b>	<b>2</b>
<b>1 INTRODUCTION .....</b>	<b>3</b>
<b>2 METHODS.....</b>	<b>4</b>
2.1 DEFINITIONS .....	4
2.1.1 <i>Quasi-identifiers</i> .....	4
2.1.2 <i>Equivalence Classes</i> .....	4
2.1.3 <i>Definition of a Small Area</i> .....	4
2.2 THE GEOLEADER ALGORITHM .....	5
2.2.1 <i>The Homogeneity Metric</i> .....	7
2.2.2 <i>Information Loss</i> .....	7
2.2.3 <i>GeoLeader Algorithm Walk-Through</i> .....	8
2.3 EMPIRICAL EVALUATION .....	11
<b>3 RESULTS.....</b>	<b>13</b>
<b>4 DISCUSSION.....</b>	<b>16</b>
4.1 SUMMARY .....	16
4.2 LIMITATIONS .....	16
<b>5 ACKNOWLEDGEMENTS .....</b>	<b>17</b>
<b>6 REFERENCES .....</b>	<b>18</b>



## Abstract

---

**Background:** Many health data sets contain geographic information. However, geographic information makes it easier to re-identify the individuals in the data, especially if the geographic areas are small. A common way to manage re-identification risk is to aggregate small areas into larger ones. Aggregation does result in the loss of information and reduces the utility of the data. The most commonly used geocodes in health data sets have been postal/ZIP codes. Thus far, a simplistic removal of characters/numbers from the end of the postal/ZIP codes has been used as the recommended form of aggregation, and this may result in aggregated areas that are too large, with the commensurate loss of utility.

**Objective:** Develop and test a clustering algorithm to aggregate areas.

**Design:** We developed a clustering algorithm which searches for an optimal way to aggregate geographic areas such that they are not too small. The algorithm was compared to the common aggregation methods for all Forward Sortation Areas (the first three characters of the postal code) in the largest eight Canadian provinces. The aggregated areas were evaluated in terms of a penalty metric that measured deviation from the minimum area size.

**Results:** The clustering algorithm always had a lower penalty than the common aggregation method. This was the case across all provinces and two penalty metrics.

**Conclusion:** Given that geographic area aggregation is a common technique for protecting the privacy of data sets with geographic information, the clustering algorithm described in this paper will ensure that areas are not too small but at the same time will limit the amount of aggregation to maximize the utility of the data.

# 1 Introduction

---

Location information is critical for many health data sets [1-9]. For example, a common patient residence location indicator is the postal/ZIP code [10-15]. However, the inclusion of such location information makes it easier to determine the identity of the individuals in the data sets [16-18]. Specifically, patients living in small geographic areas (i.e., with small populations) tend to be more easily re-identifiable because they are more likely to be unique on their demographics [19-21].

A common way to address this privacy risk is to stipulate a minimum population size for geographic areas (or a population size cutoff) [22-27]. The larger the population in the area, the less likely that an individual living there would be unique. For example, the US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule defines a small geographic area as one having a population smaller than 20,000.

Such population size cutoff rules are implemented by either aggregating small geographic areas into larger ones or removing records about individuals in the small geographic areas. A commonly used method for aggregation consists of removing the final characters/digits from the postal/ZIP codes [28, 29]. Any records in areas that are still too small after aggregation are suppressed. This simplistic approach, however, does not consider the population size of the areas and can result in too much aggregation or suppression. Aggregation can reduce the ability to perform meaningful analysis and conceal variations that would otherwise be visible at smaller geographical scales [30-35]. Suppression results in the loss of data and hence reduces the statistical power of any analysis, and can also result in bias if the suppressed records are different in some important characteristics from the rest of the data.

Furthermore, the simple aggregation method described above does not account for other variables in a data set that can be used for re-identification. For example, a male living in an area with 20,000 people is less likely to be unique. Whereas if we knew his date of birth, ethnicity, and years of schooling as well, then that person is much more likely to be uniquely identifiable in that area. In the latter situation, a stronger case can be made for aggregation than in the former situation. The other variables included in the data set do matter.

In this paper we propose and empirically evaluate a clustering algorithm to find optimal aggregations of geographic areas. This method takes into account the population in areas being aggregated, and by using the models in [36] to compute the cutoff population size, also takes into account the other variables in the data set.

## 2 Methods

---

We first describe a clustering algorithm, GeoLeader, which aggregates geographic areas in a more optimal manner than commonly used methods of aggregation. After that we describe a study on Canadian Forward Sortation Areas (FSAs - the first three characters of the postal code) to demonstrate that the algorithm is superior to these traditional methods.

### 2.1 Definitions

#### 2.1.1 Quasi-identifiers

The variables that can potentially re-identify a patient in a data set are called the *quasi-identifiers* [37]. In the current paper we exclude geographic information from the definition of quasi-identifiers. Examples of common quasi-identifiers are [28, 38-41]: dates (such as, birth, death, admission, discharge, visit, and specimen collection), race, ethnicity, languages spoken, aboriginal status, and gender.

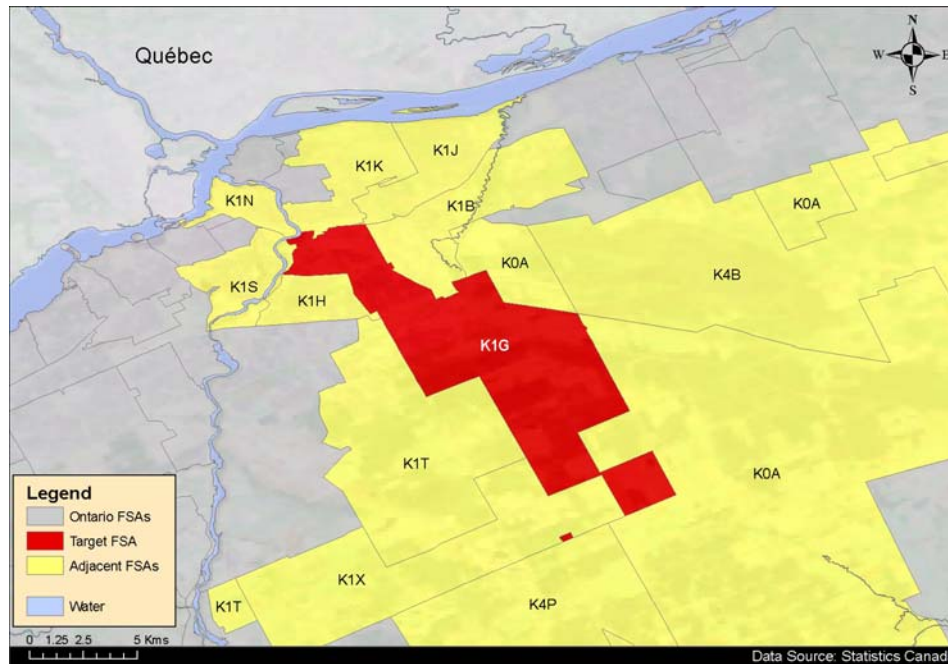
#### 2.1.2 Equivalence Classes

All the records that have the same values on the quasi-identifiers are called an *equivalence class*. For example, all records in data set for 17 year old males admitted on 1<sup>st</sup> January 2008 are an equivalence class. The number of equivalence classes in a data set is denoted by *MaxCombs* .

#### 2.1.3 Definition of a Small Area

In a recent study [36] we developed models (one for each region of Canada) using census data that determine when an area becomes sufficiently large that the risk of re-identification is negligible. The models take as input the *MaxCombs* value, and estimate the acceptable smallest population size for that area. Any geographic area that is smaller than this estimated cutoff would have to be aggregated or suppressed.

For example, if a data set contains age and gender, as well as the FSA, then the *MaxCombs* is 86 (the number of years up to the life expectancy) multiplied by 2 (gender values), to give 172. The 172 value would be input into the model and the model would provide an estimate of the minimum population size in a geographic area. For Ontario the minimal area size for a *MaxCombs* of 172 would be 13,135 people.



**Figure 1:** An example of an FSA “K1G” and the adjacent FSAs to it. Note that one FSA can be a collection of multiple nested, contiguous or non-contiguous polygons.

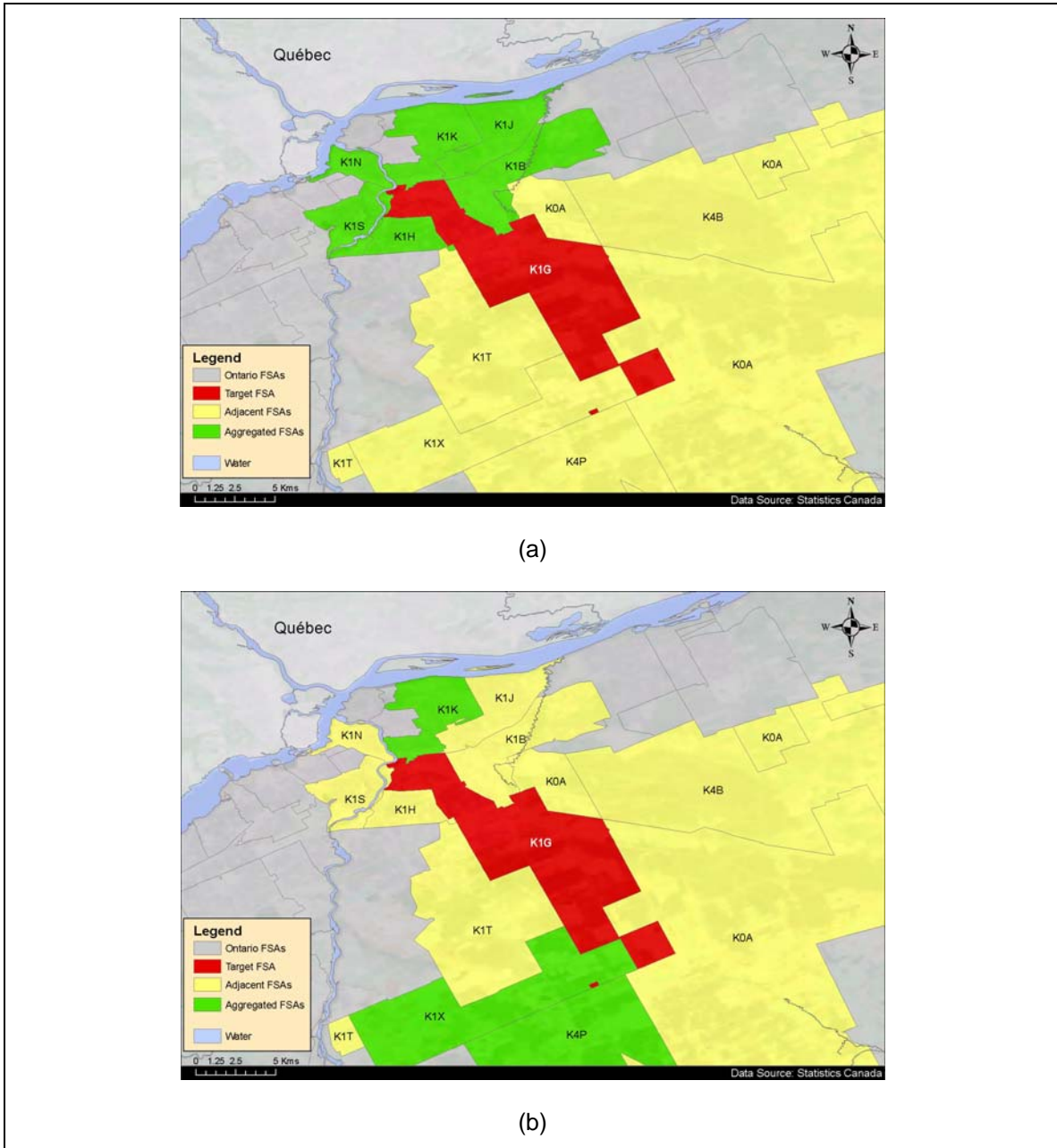
## 2.2 The GeoLeader Algorithm

The algorithm is based on the Leader clustering technique. Our adaptation of the basic Leader algorithm had to meet two criteria:

- Only adjacent areas should be aggregated. If non-adjacent areas are allowed to be aggregated then there would be no geographic relationship among the new larger areas. The example in Figure 1 shows an FSA and its adjacent FSAs.
- The aggregated areas should be as close as possible geographically or clustered (i.e., the aggregated areas should be *all* close to each other to the extent possible rather than, say, stretched out in a thin strip over a long distance). This is illustrated in Figure 2, where example (a) shows a localized clustering, and example (b) shows an aggregation that is stretched out in a longer strip that is less localized. The aggregation showed in (a) would be more desirable.

In our study the basic area was the Canadian FSA. We constructed an adjacency matrix which indicates, for each FSA, all other adjacent FSAs in the same province. This adjacency matrix is then used within the algorithm. It should be noted that the same algorithm will work with any

geographic unit for which an adjacency matrix can be created. We just happen to use FSAs in this paper.

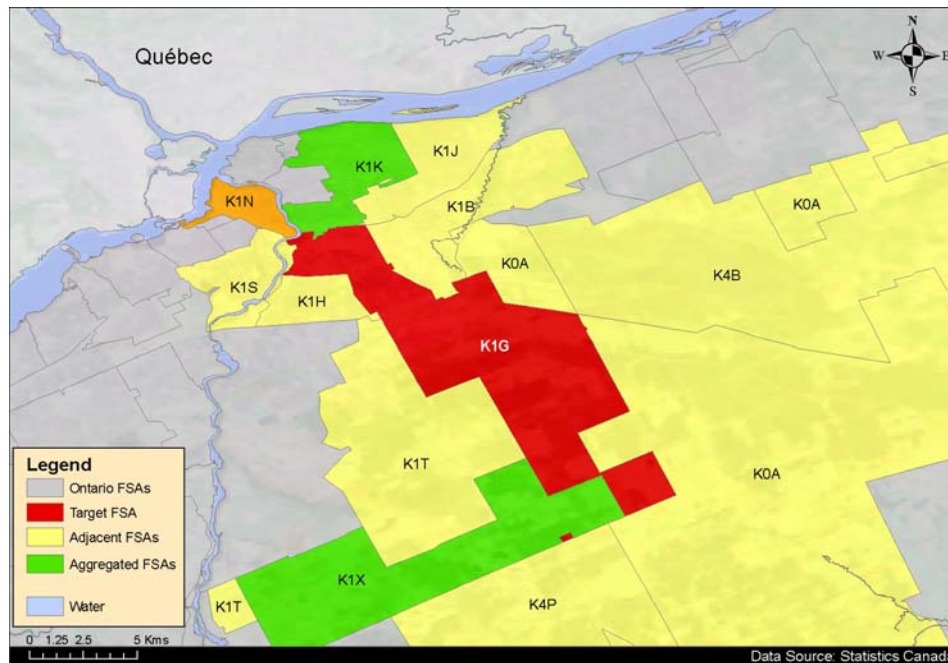


**Figure 2:** Aggregation options for FSA polygons adjacent to K1G: (a) “clustered”, minimizing distance between aggregated FSAs, and (b) “string”, where aggregation is stretched based on other parameters, irrespective of geography.

We use the term “cluster” to denote a group of FSAs that have been merged together.

### 2.2.1 The Homogeneity Metric

A homogeneity metric was defined to ensure that areas that were aggregated together are geographically close to each other. It measures the proportion of FSAs in a cluster that a separate FSA is adjacent to. For example (see Figure 3), consider two adjacent FSAs, K1N and K1G, that are merged together during aggregation into a single cluster. A third FSA, K1X, is only adjacent to K1G. Therefore, its homogeneity with the cluster is 0.5 since it is only adjacent to half the FSAs in the cluster. However, the FSA K1K is adjacent to both K1N and K1G, and therefore its homogeneity with the cluster is one.



**Figure 3:** Considering two adjacent FSAs, K1G and K1N, this figure shows how K1K is adjacent to both, and therefore has a homogeneity metric of 1. K1X, however, is only adjacent to K1G and therefore has a homogeneity metric of 0.5.

Our GeoLeader algorithm sets a minimum value of homogeneity that is required for a merge to occur. For example, if the minimum homogeneity is 0.75, then the K1X FSA would not be merged with the cluster in our example above, but the K1K FSA would be merged with the cluster.

### 2.2.2 Information Loss

FSAs are merged if the information loss from doing so is small. Information loss is defined as a weighted population size. Let  $H$  be the homogeneity metric value, then information loss is

defined as  $1 - H$  multiplied by the cluster population size. This metric penalizes merges that are less homogeneous, and penalizes merges with clusters that are larger.

For example, the 2001 census population of the cluster K1N and K1G is 59,209. The population of K1X is 843. Therefore the information loss from merging K4P to this cluster would be 0.5 multiplied by the total population of the merged areas,  $60,052 = 30,026$ .

This information loss metric is used within the algorithm to select between alternative merges.

### 2.2.3 GeoLeader Algorithm Walk-Through

We define the smallest homogeneity that is allowed for an FSA to be merged with another FSA/cluster. Let this minimum homogeneity be denoted by  $L$ . The search for good mergers is iterative and starts with a homogeneity of 1, and then decrements it to  $L$  in subsequent iterations. We therefore have the most homogeneous mergers of FSAs happening first.

We first consider the FSAs that are smaller than our area size cutoff, and we put them in a set  $A$ . The remaining FSAs, that are equal to or larger than the cutoff, are put in another set  $B$ .

One FSA from the set  $A$  is selected randomly to start off with and is removed and put in a new set  $Z$ . A second FSA is then selected randomly from  $A$ . If the homogeneity with the FSA in  $Z$  is one, then they are merged. If the homogeneity is zero, then that second FSA is put in  $Z$  and removed from the set  $A$ . A third FSA is selected from the set  $A$  and compared to the FSAs/clusters in  $Z$  with a homogeneity equal to one, and merged with the FSA/cluster where the merged cluster would have the smallest information loss. Then it is removed from the set  $A$ .

This process is repeated until all of the FSAs in the set  $A$  are exhausted, and then the acceptable homogeneity value is decremented from one to say 0.95. All areas in the set  $Z$  which are smaller than the cutoff are put back in the set  $A$ , and the remaining areas moved to the set  $B$ . Then the search for a good merger is started over again with the decremented minimum homogeneity.

While this algorithm does not find the globally optimal aggregation of FSAs, it only needs to perform better than the current simplistic method in use today to provide tangible value in terms of improved data quality (and still protect privacy).

A more precise definition of the GeoLeader algorithm is described below.

Function	Description
$a \parallel b$	Merge area $a$ and area $b$ together,
Adjacent( $a$ )	Returns a set of all areas that are geographically adjacent to area $a$ .
Population( $a$ )	Returns the population of area $a$ .
isBig( $a$ )	Determines whether an area $a$ is bigger than the cutoff population size or not. Returns True or False. It is computed as: Population( $a$ ) > Cutoff, where Cutoff is computed for a particular set of variables.
H( $a,b$ )	Computes the homogeneity of the two areas $a$ and $b$ .
InfoLoss( $a,b$ )	Computes the information loss if area $a$ is merged with area $b$ . This is computed as: $(1 - H(a,b)) \times \text{Population}(b)$ .

The full GeoLeader algorithm is then defined as follows:



**GeoLeader Algorithm**

```

// The set  $D$  consists of all of the areas
// The set  $A$  consists of all areas that are too small
 $A = \{a \mid a \in D \wedge \neg isBig(a)\}$ 

// The  $L_1$  value is decremented to  $L$ , which is the minimal allowable homogeneity
For  $L_1 = 1$  to  $L$  do
   $Z = \emptyset$ 
  For every  $a \in A$  do
    If  $|Z| = 0$  then
       $Z = \{a\}$ 
       $A = A - a$ 
    Else
      // find all areas adjacent to  $a$ 
       $Q = \{q \mid q \in Z \wedge q \in Adjacent(a)\}$ 
      //  $q'$  is the cluster that is being merged with  $a$ 
       $q' = (q' \parallel a) \mid q' \in Q \wedge H(a, q') \geq L_1 \wedge$ 
         $InfoLoss(a, q') = \min_{q \in Q} (InfoLoss(a, q))$ 
      If  $q' = \emptyset$  then
         $Z = Z + a$ 
      End If
       $A = A - a$ 
    End If
  End For
  // Recompute  $A$  as the set of all small areas
   $A = \{a \mid a \in Z \wedge \neg isBig(a)\}$ 
  // The set  $B$  consists of all areas that are larger than the cutoff
   $B = \{b \mid b \in Z \wedge isBig(b)\}$ 
End For

For every  $a \in A$  do
   $R = \{r \mid r \in B \wedge r \in Adjacent(a)\}$ 
   $r' = (r' \parallel a) \mid r' \in R \wedge H(a, r') \geq L \wedge InfoLoss(a, r') = \min_{r \in R} (InfoLoss(a, r))$ 
End For

```

## 2.3 Empirical Evaluation

We conducted an empirical evaluation to compare the GeoLeader algorithm to the commonly used method for aggregating geographic areas to protect privacy: the removal of characters/digits at the end of the postal code [28, 29]. We used the three character FSA as the area being aggregated. The three approaches that were compared were: (a) the GeoLeader aggregation of FSAs, (b) removing the last character of the FSA, and (c) removing the last two characters of the FSA.

FSA adjacency matrices were created for the eight largest provinces in Canada. The FSA adjacency matrices was created using the geographic information systems (GIS) software ArcMap 9.2. First order adjacency matrices were created, by province [42]. The adjacency matrices were used to compute aggregations using GeoLeader. Each province was analyzed separately to account for variations in the mix of population density across the country.

To evaluate the three approaches, we used two penalty metrics that:

- penalized areas that were smaller than the estimated cutoff computed from the models,
- and
- penalized areas that were too large.

Therefore, any aggregated areas that diverge too much from the cutoff value would be penalized. This balances the need to protect privacy with the negative consequences on the utility of the data if the geographic areas are excessively aggregated.

The penalties were computed relative to the three character FSA (the baseline penalty) and expressed as a percentage. For example, if the penalty is 50% then this means that it is 50% less than the penalty with the baseline three character FSA.

The first penalty metric weights the total population in a cluster. If a cluster is smaller than the cutoff value then its weight is one. For clusters that are larger than the cutoff, we use a weight based on the homogeneity metric. Each FSA in a cluster is taken out and its homogeneity with the remainder of its cluster is computed. The average across all FSAs in the cluster is computed. One minus the average homogeneity is used as a weight, and multiplied by the population size of the cluster. The weighted cluster population is summed across all clusters. For example, if all FSAs are smaller than the cutoff then their weights are all one, and the penalty would be equal to the population of the province. This penalty is referred to as the Homogeneity Information Loss (HIL).

The second penalty metric is derived from the commonly used discernability metric (DM) in the computational disclosure control literature [43-51]. This is defined as:

$\sum_{pop(a) \geq cutoff} pop(a)^2 + \sum_{pop(a) < cutoff} pop(D)^2$  where  $a$  is an FSA within the province,  $pop(a)$  is

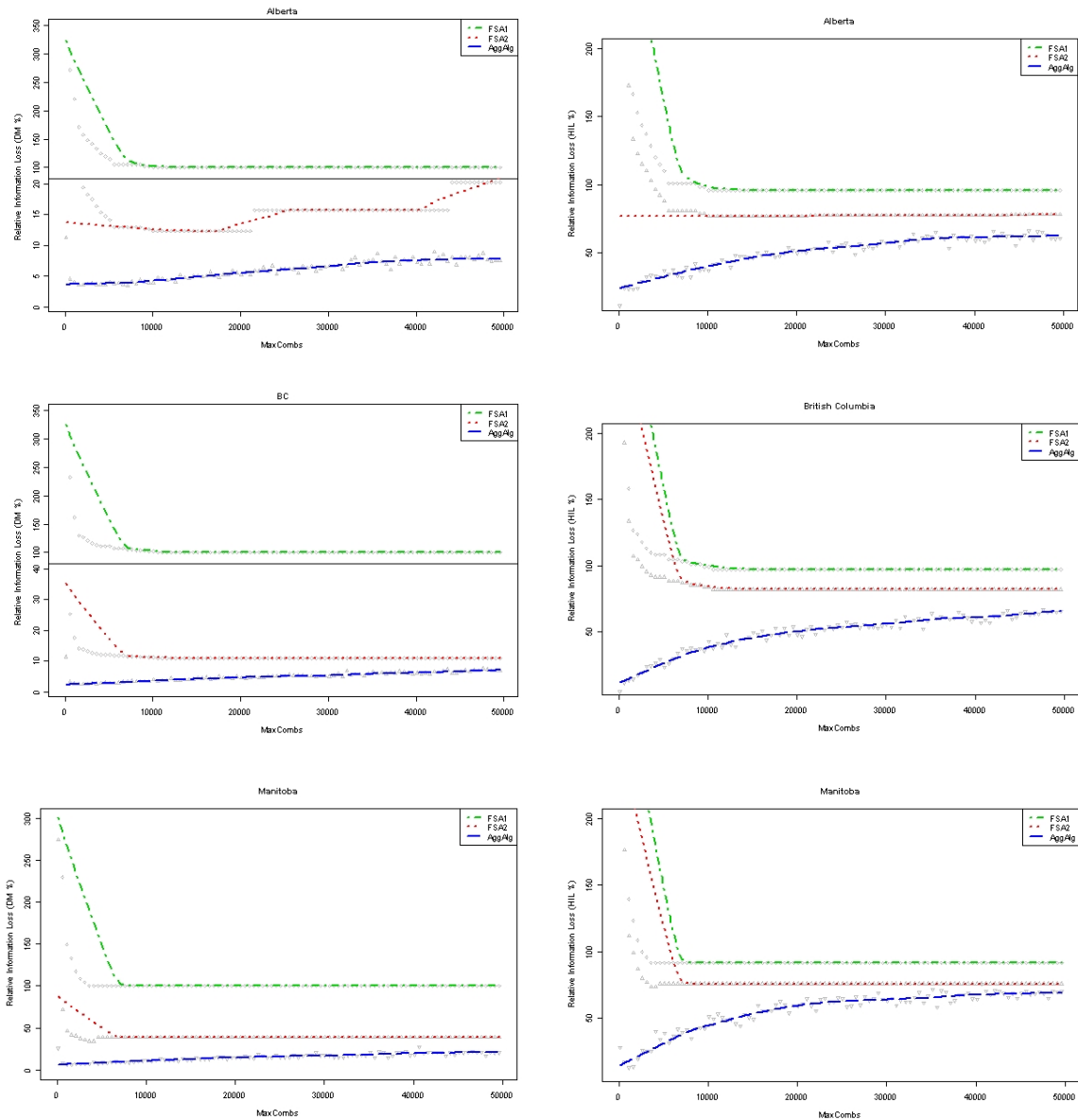
the population of an area, and  $D$  signifies the whole province (i.e.,  $pop(D) = \sum pop(a)$ ).

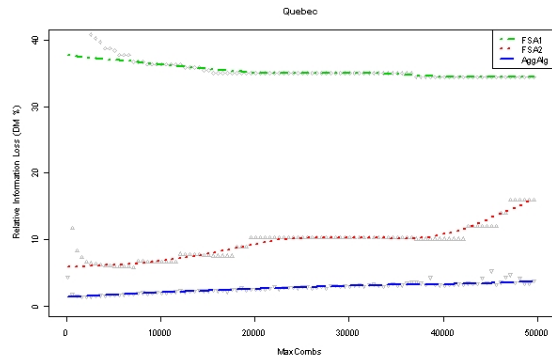
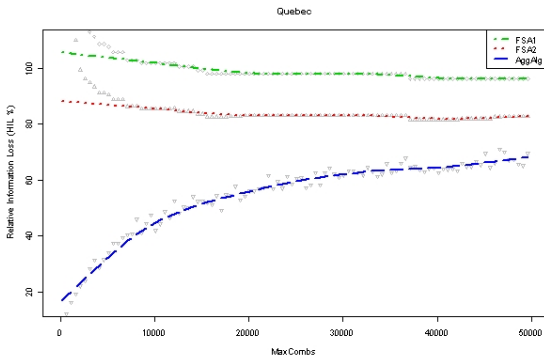
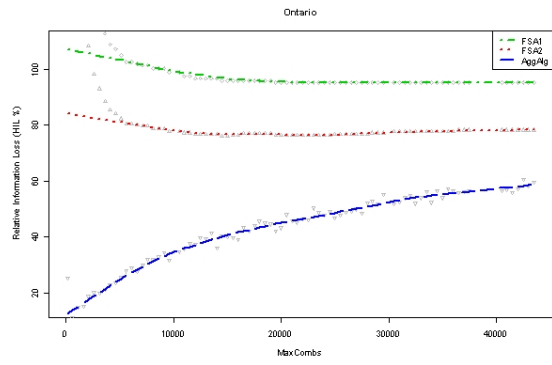
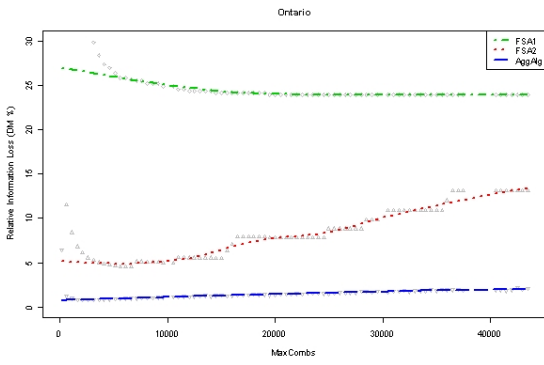
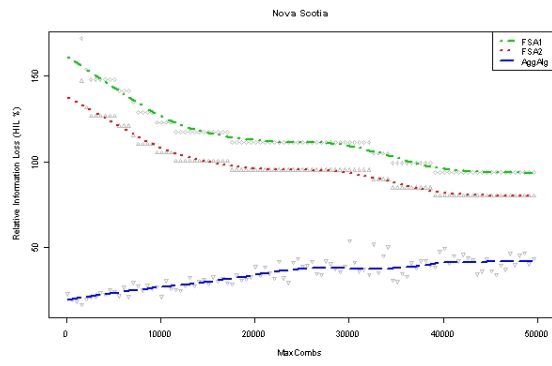
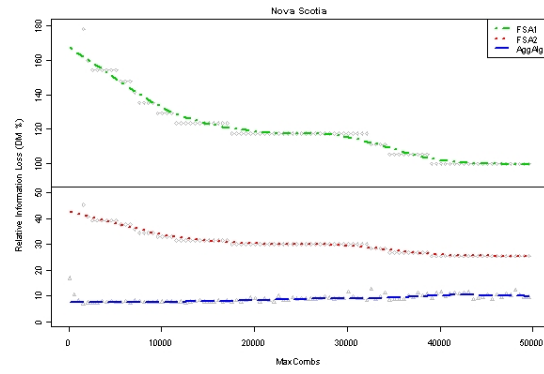
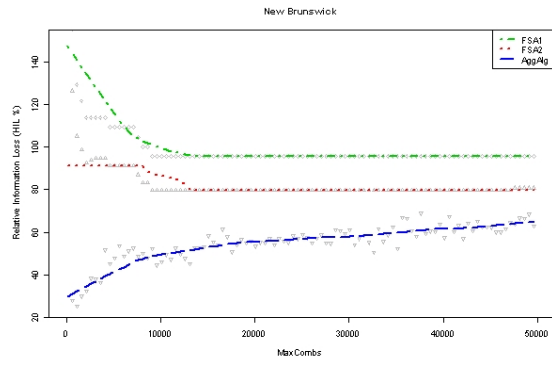
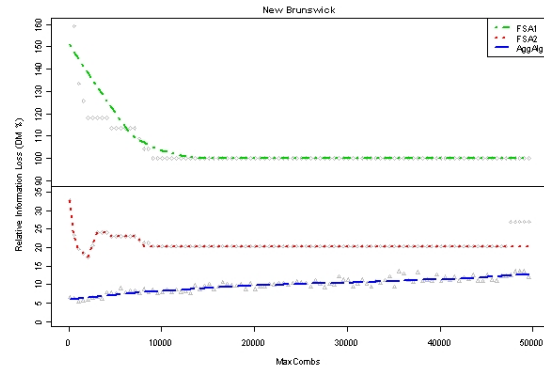
This metric severely penalizes areas that are too small and would therefore need to be suppressed.

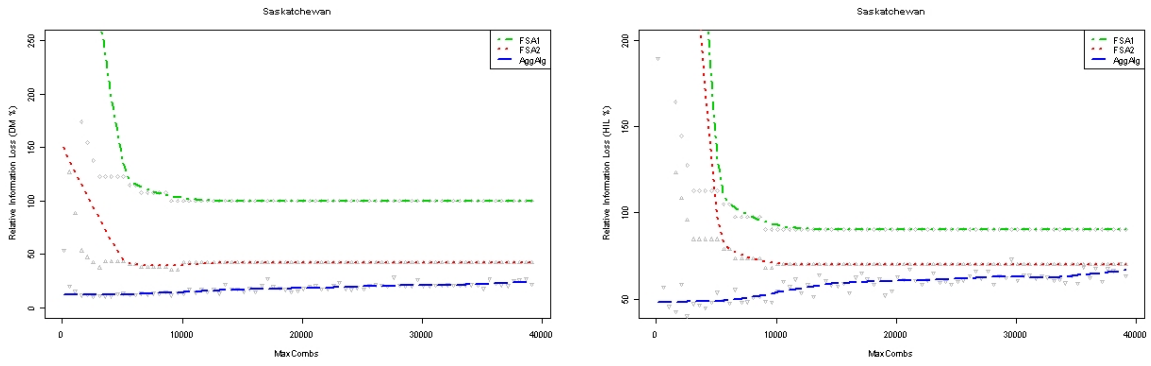
The *MaxCombs* value was varied from 500 to 50,000 to account for different possible variables that may be included in a data set.

### 3 Results

The results are shown in Figure 4. These plot the two penalty metrics for the algorithm, aggregation by removing the last character of the FSA, and aggregation by removing the last two characters of the FSA. As is clear, the GeoLeader algorithm always performs better than the other two commonly used approaches across all the provinces we evaluated.







**Figure 4:** Evaluation results of the GeoLeader algorithm on the two penalty metrics for eight Canadian provinces. The FSA1 plots show the penalty when only the first character of the FSA is used, and the FSA2 plots show the penalty when only the first two characters of the FSA are used.

## 4 Discussion

---

### 4.1 Summary

Because individuals living in small geographic areas tend to be easier to re-identify, the aggregation of smaller areas into larger ones is a common method for protecting privacy. In this paper we have presented a clustering algorithm for performing this aggregation. This algorithm builds on recent work to empirically define when an area becomes too small [36], and we have shown that the algorithm works better than currently used simplistic methods of aggregation. Our evaluation was done for all FSAs in Canada's eight largest provinces.

This clustering algorithm has two main advantages. First, it takes into account the other (non-geographic) variables in a data set that can be used for re-identification. It does so through the *MaxCombs* value that is used to estimate the area size cutoff. Secondly, it results in less aggregation and suppression. This means data sets will be of higher utility for subsequent analysis that requires geographic detail, for example, health services research and public health investigations.

While our analysis used FSAs as the geographic area, the algorithm itself is not limited to FSAs and can be used with any definition of an area, whether it is based on political boundaries, service provision boundaries, or some other criterion.

### 4.2 Limitations

To apply this clustering algorithm requires an adjacency matrix to be constructed. An adjacency matrix shows for each area what other areas are physically adjacent to it. We have found that such matrices are not readily available and we therefore had to construct them for FSAs in the eight provinces ourselves.

## **5 Acknowledgements**

---

This work was funded by the GeoConnections (Natural Resources Canada), the Public Health Agency of Canada, the Ontario Centers of Excellence, and the Natural Sciences and Engineering Research Council of Canada.



## 6 References

---

1. Boulos M. *Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom*. International Journal of Health Geographics 2004; 3(1).
2. O'Dwyer LA, Burton DL. *Potential meets reality: GIS and public health research in Australia*. Australian and New Zealand Journal of Public Health, 1998; 22(7):819-823.
3. Ricketts TC. *Geographic information systems and public health*. Annual Review of Public Health, 2003; 24:1-6.
4. Cromley EK. *GIS and Disease*. Annual Review of Public Health, 2003; 24:7-24.
5. Brindley P, Maheswaran R. *My favourite software: geographic information systems*. Journal of Public Health Medicine, 2002; 24(2):149.
6. Richards TB, Croner CM, Rushton G, Brown CK, Fowler L. *Geographic information systems and public health: mapping the future*. Public Health Reports, 1999; 114:359-373.
7. Ricketts T. *Geographic information systems and public health*. Annual Review of Public Health, 2003; 24:1-6.
8. McLafferty S. *GIS and health care*. Annual Review of Public Health, 2003; 24:25-42.
9. Cromley E. *GIS and disease*. Annual Review of Public Health, 2003; 24:7-24.
10. Bow C, Waters N, Faris P, Seidel J, Galbraith P, Knudtson M, Ghali W. *Accuracy of city postal code coordinates as a proxy for location of residence*. International Journal of Health Geographics, 2004; 3(5).
11. Ng E, Wilkins R, Perras A. *How far is it to the nearest hospital ? Calculating distances using the Statistics Canada Postal Code Conversion file*. Health Reports, 1993; 5:179-183.
12. Mackillop W, Zhang-Salomons J, Groome P, Pazat L, Holowaty E. *Socioeconomic status and cancer survival in Ontario*. Journal of Clinical Oncology, 1997; 15:1680-1689.
13. Spasoff A, Gilkes D. *Up-to-date denominators: Evaluation of taxation family for public health planning*. Canadian Journal of Public Health, 1994; 85:413-417.
14. Demissie K, Hanley J, Menzies D, Joseph L, Ernst P. *Agreement in measuring socio-economic status: Area-based versus individual measures*. Chronic Diseases in Canada, 2000; 21:1-7.
15. Guernsey J, Dewar R, Weerasinghe S, Kirkland S, Veugelers P. *Incidence of cancer in sydney and Cape breton County, Nova Scotia 1979-1997*. Canadian Journal of Public Health, 2000; 91:285-292.
16. Mugge R. *Issues in protecting confidentiality in national health statistics*. Proceedings of the Social Statistics Section, American Statistical Association. 1983.
17. Mackie C, Bradburn N. *Improving access to and confidentiality of research data: Report of a workshop*. 2000: The National Academies Press.
18. Croner C. *Public health, GIS, and the Internet*. Annual Review of Public Health, 2003; 24:57-82.

19. Hawala S. *Enhancing the "100,000" rule: On the variation of percent of uniques in a microdata sample and the geographic area size identified on the file*. Proceedings of the Annual Meeting of the American Statistical Association 2001.
20. Greenberg B, Voshell L. *Relating risk of disclosure for microdata and geographic area size*. Proceedings of the Section on Survey Research Methods, American Statistical Association. 1990.
21. Greenberg B, Voshell L. *The geographic component of disclosure risk for microdata*. 1990; Bureau of the Census.
22. Zayatz L, Massell P, Steel P. *Disclosure limitation practices and research at the US Census Bureau*. Netherlands Official Statistics, 1999; 14(Spring):26-29.
23. Zayatz L. *Disclosure avoidance practices and research at the US Census Bureau: An update*. 2005; US Census Bureau.
24. Hawala S. *Microdata disclosure protection research and experiences at the US census bureau*. Proceedings of Workshop on Microdata. 2003; Available from: [\[http://www.census.gov/srd/sdc/microdataprotection.pdf\]](http://www.census.gov/srd/sdc/microdataprotection.pdf). Archived at: [\[http://www.webcitation.org/5b7mPeVPi\]](http://www.webcitation.org/5b7mPeVPi).
25. Marsh C, Dale A, Skinner C. *Safe data versus safe settings: Access to microdata from the British census*. International Statistical Review, 1994; 62(1):35-53.
26. Statistics Canada. *Canadian Community Health Survey (CCHS) Cycle 3.1 (2005) Public Use Microdata File (PUMF) User Guide*. 2006.
27. Willenborg L, de Waal T. *Statistical Disclosure Control in Practice*. 1996: Springer-Verlag.
28. Canadian Institutes of Health Research. *CIHR best practices for protecting privacy in health research*. 2005; Canadian Institutes of Health Research.
29. Pabrai U. *Getting Started with HIPAA*. 2003: Premier Press.
30. Fefferman N, O'Neil E, Naumova E. *Confidentiality and confidence: Is data aggregation a means to achieve both ?* Journal of Public Health Policy, 2005; 26(4):430-449.
31. Willenborg L, Mokken R, Pannekoek J. *Microdata and disclosure risks*. Proceedings of the Annual Research Conference of US Bureau of the Census. 1990.
32. Olson K, Grannis S, Mandl K. *Privacy protection versus cluster detection in spatial epidemiology*. American Journal of Public Health, 2006; 96(11):2002-2008.
33. Marceau D. *The scale issue in social and natural sciences*. Canadian Journal of Remote Sensing, 1999; 25(4):347-356.
34. Bivand R. *A review of spatial statistical techniques for location studies*. 1998; Norwegian School of Economics and Business Administration.
35. Ratcliffe J. *The Modifiable Areal Unit Problem*. 2005; Available from: [\[http://www.jratcliffe.net/research/maup.htm\]](http://www.jratcliffe.net/research/maup.htm).
36. El Emam K, Brown A, Abdelmalik P. *Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk*. Journal of the American Medical Informatics Association, 2009; 16(2):256-266.
37. Dalenius T. *Finding a needle in a haystack or identifying anonymous census records*. Journal of Official Statistics, 1986; 2(3):329-336.
38. El Emam K, Brown A, Abdelmalik P. *Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk*. Journal of the American Medical Informatics Association, 2008; (accepted).

39. El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. *Evaluating common de-identification heuristics for personal health information*. Journal of Medical Internet Research, 2006; 8(4):e28.
40. El Emam K, Jonker E, Sams S, Neri E, Neisa A, Gao T, Chowdhury S. *Pan-Canadian De-Identification Guidelines for Personal Health Information (report prepared for the Office of the Privacy Commissioner of Canada)*. 2007; Available from: [\[http://www.ehealthinformation.ca/documents/OPCReportv11.pdf\]](http://www.ehealthinformation.ca/documents/OPCReportv11.pdf). Archived at: [\[http://www.webcitation.org/5Ow1Nko5C\]](http://www.webcitation.org/5Ow1Nko5C).
41. ISO/TS 25237. *Health Informatics: Pseudonymization*. 2008.
42. Hitchen M. *Nth order polygon neighbour analysis*. ESRI Support Center; Available from: [\[http://arcscripsts.esri.com/details.asp?dbid=15048\]](http://arcscripsts.esri.com/details.asp?dbid=15048).
43. Bayardo R, Agrawal R. *Data Privacy through Optimal k-Anonymization* Proceedings of the 21st International Conference on Data Engineering. 2005.
44. LeFevre K, DeWitt D, Ramakrishnan R. *Mondrian multidimensional k-anonymity*. Proceedings of the 22nd International Conference on Data Engineering. 2006.
45. Hore B, Jammalamadaka R, Mehrotra S. *Flexible anonymization for privacy preserving data publishing: A systematic search based approach*. Proceedings of SIAM International Conference on Data Mining. 2007.
46. Xu J, Wang W, Pei J, Wang X, Shi B, Fu A. *Utility-based anonymization for privacy preservation with less information loss*. ACM SIGKDD Explorations Newsletter 2006; 8(2):21 - 30.
47. Nergiz M, Clifton C. *Thoughts on k-anonymization*. Second International Workshop on Privacy Data Management. 2006.
48. Poletini S. *A note on the individual risk of disclosure*. 2003; Istituto nazionale di statistica (Italy).
49. Machanavajjhala A, Gehrke J, Kifer D, Venkatasubramanian M. *L-Diversity: Privacy Beyond k-Anonymity*. International Conference on Data Engineering. 2006.
50. Bayardo R, Agrawal R. *Data Privacy through Optimal k-Anonymization*. Proceedings of the 21st International Conference on Data Engineering. 2005.
51. El Emam K, Dankar F. *Protecting privacy using k-anonymity*. Journal of the American Medical Informatics Association, 2008; 15:627-637.

# Uniqueness in the Canadian Population

*29<sup>th</sup> January 2009*

**Khaled El Emam**

*CHEO Research Institute*

**Ann Brown**

*CHEO Research Institute*

**Philip AbdelMalik**

*Public Health Agency of Canada*

**Angelica Neisa**

*CHEO Research Institute*



## **Document Information**

**Document Title:** Uniqueness in the Canadian Population  
**Original Document Date:** 14<sup>th</sup> January 2009  
**Document Version:** Version 7  
**Copyright:** CHEO Research Institute, 401 Smyth Road, Ottawa, ON K1H 8L1, Canada  
**Contact:** Khaled El Emam (kelemam@ehealthinformation.ca)  
**More Information:** <http://www.ehealthinformation.ca/>

## **Other Relevant Publications and Reports**

- K. El Emam, A. Brown, and P. AbdelMalik: "Evaluating predictors of geographic area population size cutoffs to manage re-identification risk." In *Journal of the American Medical Informatics Association*, March/April, 2009.
- K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk: "Evaluating Patient Re-identification Risk from Hospital Prescription Records." In the *Canadian Journal of Hospital Pharmacy*, June 2009.
- K. El Emam: "Heuristics for de-identifying health data." In *IEEE Security and Privacy*, July/August, 6(4):58-61, 2008.
- K. El Emam, and F. Dankar: "Protecting privacy using k-anonymity." In the *Journal of the American Medical Informatics Association*, September/October, 15:627-637, 2008.
- K. El Emam, E. Neri, and E. Jonker: "An evaluation of personal health information remnants in second hand personal computer disk drives." In *Journal of Medical Internet Research*, 9(3):e24, 2007.
- K. El Emam, S. Jabbouri, S. Sams, Y. Drouet, M. Power: "Evaluating common de-identification heuristics for personal health information." In *Journal of Medical Internet Research*, 2006;8(4):e28, November 2006.
- K. El Emam: "Overview of Factors Affecting the Risk of Re-Identification in Canada", Access to Information and Privacy, Health Canada, May 2006.
- K. El Emam: "Data Anonymization Practices in Clinical Research: A Descriptive Study", Access to Information and Privacy, Health Canada, May 2006.

*More information is available from*  
<http://www.ehealthinformation.ca/>

## Table of Contents

<b>EXECUTIVE SUMMARY .....</b>	<b>2</b>
<b>1 INTRODUCTION .....</b>	<b>3</b>
<b>2 METHODS.....</b>	<b>5</b>
2.1 DEFINITIONS .....	5
2.1.1 <i>Quasi-identifiers</i> .....	5
2.1.2 <i>Equivalence Classes</i> .....	5
2.1.3 <i>Focus on Forward Sortation Area (FSA)</i> .....	5
2.2 DATA SET .....	5
2.3 QUASI-IDENTIFIER MODELS .....	6
2.4 ESTIMATING UNIQUENESS.....	8
2.5 PREDICTION MODELS .....	8
<b>3 RESULTS.....</b>	<b>10</b>
<b>4 DISCUSSION.....</b>	<b>11</b>
4.1 SUMMARY .....	11
4.2 USING THE MODELS .....	11
4.3 APPLICATION OF RESULTS.....	12
4.4 SELECTION OF THRESHOLD .....	13
4.5 LIMITATIONS .....	15
4.6 RELATIONSHIP TO PREVIOUS WORK.....	15
<b>5 ACKNOWLEDGEMENTS .....</b>	<b>17</b>
<b>6 APPENDIX A: MAPPING CENSUS GEOGRAPHY TO POSTAL GEOGRAPHY USING A GRIDDING METHODOLOGY .....</b>	<b>18</b>
6.1 BACKGROUND .....	18
6.2 METHODS .....	18
6.3 RESULTS .....	22
6.4 CONCLUSIONS .....	24
<b>7 REFERENCES .....</b>	<b>25</b>

## Executive Summary

---

**Objective:** A common disclosure control practice for health data sets is to identify small geographic areas and either suppress records from these areas or aggregate them into larger ones. A recent study provided a method for deciding when an area is too small based on the uniqueness criterion. This uniqueness criterion stipulates that an area is no longer too small when the proportion of unique individuals on the relevant variables (the quasi-identifiers) approaches zero. However, using a uniqueness value of zero is quite a stringent threshold, and is only suitable when the risks from disclosure are quite high. Other uniqueness thresholds that have been proposed for health data are 5% and 20%.

**Design:** We estimated uniqueness for urban Forward Sortation Areas (FSAs) by using the 2001 long form Canadian census data representing 20% of the population. We then constructed two logistic regression models to predict when the uniqueness is greater than the 5% and 20% thresholds, and validated their predictive accuracy using 10-fold cross-validation. Model parameters included the population size of the FSA and the maximum number of possible values on the quasi-identifiers.

**Results:** All model parameters were significant and the models had very high prediction accuracy, with sensitivity above 0.9, and specificity at 0.87 and 0.74 for the 5% and 20% threshold models respectively. The application of the models was illustrated with an analysis of the Ontario newborn registry. At the 5% and 20% thresholds less than 1% of the records would have to be suppressed, but the majority of the records would have to be suppressed at the 0% threshold. We have also included checklists to provide guidance for data custodians in deciding which one of the three uniqueness thresholds to use (0%, 5%, 20%), depending on the mitigating controls that the data recipients have in place, the potential invasion of privacy if the data is disclosed, and the motives and capacity of the data recipient to re-identify the data.

**Conclusion:** The models we developed can be used to manage the re-identification risk from small geographic areas. Being able to choose among three possible thresholds, a data custodian can adjust the definition of “small geographic area” to the nature of the specific data and recipient.

# 1 Introduction

---

The disclosure of health data for secondary purposes, such as research, public health, marketing, and quality improvement, are increasing [1-5]. In many instances it is impossible or impractical to obtain the consent of the patients ex post facto for such purposes. But if the data is de-identified then there is no legislative requirement to obtain consent.

The inclusion of geographic information in health data sets is critical for many analyses [6-14]. However, the inclusion of geographic details in a data set also makes it much easier to re-identify patients [15-17]. This is exemplified by a recent Canadian federal court ruling which implied that the inclusion of an individual's province of residence in a data set may re-identify individuals [18].

Records from individuals living in small geographic areas tend to have a higher probability of being re-identified [19-21]. Some general heuristics for deciding when a geographic area is too small have been applied by national statistical agencies [22-27]. For example, the US Health Information Portability and Accountability Act (HIPAA) Privacy Rules defines a small geographic area as one having a population smaller than 20,000.

Two common ways of managing the re-identification risks of small geographic areas are to: (a) remove records in the small geographic areas, or (b) aggregate the small geographic areas into larger ones. The former results in the loss of data and hence reduces the statistical power of any analysis, and can also result in bias if the suppressed records are different in some important characteristics from the rest of the data. The latter can reduce the ability to perform meaningful analysis and conceal variations that would otherwise be visible at smaller geographical scales [28-33].

The uniqueness of individuals is often used as a surrogate measure of re-identification risk [34]. An individual is unique if s/he is the only individual with a specific combination of values of their personal characteristics that are included in a data set. A recent study has provided a direct empirical link between the geographic area size heuristics used by the national statistical agencies to uniqueness [35]: if uniqueness within a geographic area is approximately zero then the geographic area is not too small.

However, using zero uniqueness as a threshold for disclosure control is quite stringent, and higher thresholds have been found acceptable and applied in practice. Specifically, previous reports have proposed thresholds of 5% and 20% population uniqueness as acceptable for public release and research use respectively [36-38].



In this paper we extend on this line of work by developing models to determine whether a Forward Sortation Area (the first three characters of the postal code) is too small based on the 5% and 20% uniqueness thresholds.

## 2 Methods

---

Our approach was to construct models to determine if the proportion of unique records in a particular FSA was above the 5% and the 20% thresholds. These models characterize the FSA in terms of its population size, and also take into account the characteristics of the variables in the data set that can be used for re-identification.

### 2.1 Definitions

#### 2.1.1 Quasi-identifiers

The variables in a data set that can be used to re-identify individuals are called the *quasi-identifiers* [39]. Examples of common quasi-identifiers are [35, 40, 41]: dates (such as, birth, death, admission, discharge, visit, and specimen collection), race, ethnicity, languages spoken, aboriginal status, and gender.

The uniqueness of records in the data set are based on the quasi-identifiers. For example, if our quasi-identifiers are age and gender, then say, the only 90 year old female in the FSA “N3E” would be a unique record on these quasi-identifiers within that geographic area.

#### 2.1.2 Equivalence Classes

An equivalence class is defined as the set of records having a given set of values on the quasi-identifiers. For example, “50 year old male” represents the equivalence class of records with the “50” value on the age quasi-identifier and “Male” on the gender quasi-identifier. The number of records that have these two values on the quasi-identifiers is the size of the equivalence class.

#### 2.1.3 Focus on Forward Sortation Area (FSA)

The postal code is the basic geographical unit that we will use in our analysis. While the postal code is often collected because it is readily available, and consequently, used as the geographical location of residence in health data sets [42-47], the full six character postal code is clearly too specific. Further, in combination with other variables the full postal code would make it easy to re-identify individuals, especially in residential urban areas [41].

While there are many potential ways of aggregating geographic regions to construct areas for analysis [33], the FSA, a higher level of the postal code geographic unit, is the unit that we considered.

### 2.2 Data Set

The data set we used is the 2001 long form census microdata file (collected from 20% of the households) made available by Statistics Canada through its Research Data Centers (RDCs).

The RDC data set only has geographic information at the level of the census tract. We developed a gridding methodology, described in Appendix A, to assign the FSAs from the census tracts.

Variable Name in the 2001 Census RDC File	Definition	# Response categories <sup>(*)</sup>
SEXP	Gender	2
BRTYR	Year of birth (from 1880 to 2001). Age: We defined age categories based on 5 year ranges.	24
HLNABDR	Language: Language spoken most often at home by the individual at the time of the census.	4
ETH1-6	Ethnic Origin: Refers to the six possible answers for the ethnic or cultural group(s) to which the respondent's ancestors belong.	26
ASRR	Aboriginal Identity: Persons identifying with at least one Aboriginal group.	8
RELIGWI	Religious denomination: Specific religious denominations, groups or bodies as well as sects, cults, or other religiously defined communities or systems of belief.	3
TOTYRSR	Total Years of Schooling: Total sum of the years (or grades) of schooling at the elementary, high school, university and college levels. Only available for individuals age 15+.	9
MARST	Marital Status (Legal)	5
TOTINC	Total income: Total money income received from all sources during the calendar year 2000 by persons 15 years of age and over. We defined categories in \$15K ranges.	22
DVISMIN	Visible minority status	4
DISABIL	Activity difficulties/reductions: Combinations of one or more activity difficulties/reduction.	4

<sup>(\*)</sup> The number of response categories excludes non-specific responses such as missing values, not available or "other".

**Table 1:** The list of quasi-identifiers that were analyzed from the census file.

### 2.3 Quasi-identifier Models

A quasi-identifier model consists of two or more quasi-identifiers (*qid*). The subset of quasi-identifiers from the census file that we analyzed is shown in Table 1. These were selected to be representative of commonly used quasi-identifiers.

To manage the scope of the analysis we consider only combinations of up to and including 5 *qids*. A total of 358 models were analyzed. This results from the following approach of combining the *qids*.

Initially, for the 11 *qids* listed in the above table, there are some similarities related to ethnicity and therefore they were treated as a group: HLNABDR, ETH1-6, RELIGWI, and DVISMIN. Whenever the ethnicity variable appears in a model it was replaced by one of its similar variables. Each substitution represented a different model.

Thus, this gives 8 distinct *qids*: gender, age, ethnicity, schooling, marital status, total income, aboriginal identity and activity difficulties.

Categorizing the 8 distinct *qids* by their sensitivity and availability for re-identification gives the following two types:

- Very sensitive and available: gender, and age
- Possibly used for re-identification/sensitive: ethnicity, schooling, marital status, total income, aboriginal identity and activity difficulties

The value for  $C_r^n$  gives the number of possible combinations of size  $r$  from a larger group of size  $n$ . The different models were defined by the number of *qids* in the model and by having at least one very sensitive *qid* included in each model.

For models including both age and gender, there are 42 models for the 8 distinct *qids* as follows:

- 5 *qids*: have age and gender and 20 combinations of 3 of the 6 sensitive *qids*.
- 4 *qids*: have age and gender and 15 combinations of 2 of the 6 sensitive *qids*.
- 3 *qids*: have age and gender and each of the 6 sensitive *qids*.
- 2 *qids*: have age and gender only – there is only one model.

Then substituting each of language, religion and visible minority for ethnicity gives an additional 48 models: 30 (3X10) models for 5 *qids* (ethnicity appears in 10 of the 20 models), 15 (3x5) models for 4 *qids* (ethnicity appears in 5 of the 15 models), and 3 (3x1) models for 3 *qids* (ethnicity appears in one of the 6 models).

The subtotal for this group of models containing both age and gender is 90 (42+48).

We repeated the above process for each *one* of age and gender in combination with the sensitive *qids*. That is there are 56 models containing:

- 5 *qids*: have age and 15 combinations of 4 of the 6 sensitive *qids*.
- 4 *qids*: have age and 20 combinations of 3 of the 6 sensitive *qids*.
- 3 *qids*: have age and 15 combinations of 2 of the 6 sensitive *qids*.
- 2 *qids*: have age and each of the 6 sensitive *qids* only.

Similarly to the previous group, by taking into account the ethnicity related variables, there are a sub-total of 134 models for this group.

Lastly, age is replaced with gender for an additional 134 models. Adding up the sub-totals gives a total number of 358 quasi-identifier models.

## 2.4 Estimating Uniqueness

Given that the data set is a sample from the census, we will use uniqueness estimators to determine the proportion of unique records for each FSA by quasi-identifier model combination.

The fraction of population uniques can be estimated by using the poisson–gamma model with the  $\alpha$  and  $\beta$  parameters estimated by the method of moments [34, 48]. However, this approach over-estimates with small sampling fractions and under-estimates as the sampling fraction increases [49]. We will therefore adopt a different estimation approach that is based on sub-sampling [21, 29, 50]. While this approach tends to over-estimate (positive bias) the number of unique population estimates for small sampling fractions, the 20% sampling fraction of households should alleviate concerns about bias.

## 2.5 Prediction Models

We developed one binary logistic regression model [51] to determine whether the estimated uniqueness for a particular FSA and quasi-identifier model was above 5%, and another to determine if the estimated uniqueness was above 20%. We denote the probability that the uniqueness on a particular FSA and quasi-identifier model is above 5% as  $\pi_{05}$ , and the probability that the uniqueness on a particular FSA and quasi-identifier model is above 20% as  $\pi_{20}$ . Whether the uniqueness is above the threshold or not depends on the population size of the FSA and on the characteristics of the quasi-identifiers. The population size of the FSA can be obtained from Statistics Canada from the 2001 census. We denote this variable as *POP*. In a previous study it was shown that the maximum number of combinations of values on the quasi-identifiers was a good predictor of uniqueness [35]. We denote this variable *MaxCombs*. For example, if we have two quasi-identifiers, age and gender, and age has 86 possible values (age range 0-85) and gender has two, then the value of *MaxCombs* is  $86 \times 2 = 172$ , which represents the maximum possible values for these two quasi-identifiers.

The 5% model was therefore:  $\text{logit}(\pi_{05}) \sim POP + MaxCombs + (POP \times MaxCombs)$ . To avoid collinearity with the interaction term in the model, both independent variables were centered [52]. The 20% model was similarly constructed. Because both independent variables in this model have large values, the interaction term can create overflow problems during computation. We therefore scaled the independent variables by 10,000.

An observation for these logistic regression models was an FSA by quasi-identifier model combination. For example, there is one observation for the “K1J” FSA for the quasi-identifier model “age x gender”. In total there were 957 FSAs and 358 quasi-identifier models, giving 342,606 observations.

This data set was unbalanced. This means that the proportion of observations with uniqueness less than 20% was quite small, and similarly for the proportion of observations with uniqueness less than 5%. Constructing regression models with an unbalanced data set can result in poor model fit, inaccuracy in predicting the less prevalent class, and may even impede the convergence of the numeric maximum likelihood estimation algorithms. We re-balanced the data set using down-sampling, and adjusted the parameter estimates accordingly [53-55].

To validate that the models can correctly predict which FSAs are above the 5% and 20% threshold respectively, we used 10-fold cross-validation [56, 57]. That is, we divided the data sets into deciles and used one decile in turn as the test data set, and the remaining nine deciles to build the model. The down-sampling was performed separately on the nine deciles each time a model was estimated.

If the predicted probability,  $\hat{\pi}_{05}$ , was greater than 0.5 then the FSA was deemed to have a uniqueness greater than 5%. A similar predicted probability cut-off was used for  $\hat{\pi}_{20}$ . The overall prediction accuracy was evaluated in terms of average sensitivity and specificity across the 10-folds.

### 3 Results

In this section we present the results of the regression models and the 10-fold cross-validation. Both models had a Hosmer & Lemshow goodness of fit  $p < 0.001$  [51]. The model parameters are shown in Table 2. All model parameters are significant, including the interaction term. The sensitivity and specificity values averaged across the 10-fold cross validation are shown in Table 3, demonstrating that both of the models have good predictive power.

Logistic Regression Model for 5% Threshold				
	Intercept	<i>POP</i>	<i>MaxCombs</i>	<i>POP</i> × <i>MaxCombs</i>
<b>Coefficient</b>	775.7	-37.35	137.8	-6.5
<b>p-value</b>	<0.0001	<0.0017	<0.001	0.0019
Logistic Regression Model for 20% Threshold				
	Intercept	<i>POP</i>	<i>MaxCombs</i>	<i>POP</i> × <i>MaxCombs</i>
<b>Coefficient</b>	63.3	-6	11.8	-1
<b>p-value</b>	<0.0001	<0.0001	<0.0001	<0.0001

**Table 2:** Logistic regression model results for the 5% and 20% thresholds.

	Sensitivity	Specificity
<b>5% Model</b>	0.996	0.87
<b>20% Model</b>	0.98	0.74

**Table 3:** Sensitivity and specificity results for the 10-fold cross-validation on the 5% and 20% models.

## 4 Discussion

---

### 4.1 Summary

A common disclosure control practice for health data sets is to identify small geographic areas and either suppress records from these areas or aggregate them into larger ones. A recent study provided a method for deciding when an area is too small based on the uniqueness criterion [35]. That is, the uniqueness criterion in that study stipulated that an area is no longer too small when the proportion of unique individuals on the quasi-identifiers approaches zero.

However, using a uniqueness value of zero is quite a stringent threshold. Thresholds of 5% and 20% uniqueness have been proposed for the disclosure of sensitive health data. Such higher thresholds would be preferred if the overall risk of disclosing the data can be managed.

In this paper we developed models to predict whether the population in a geographic area has a uniqueness above the 5% and 20% thresholds using data from the Canadian census. We also demonstrated that the prediction models are quite accurate with high sensitivity and specificity. The areal unit that we studied was the urban FSA.

### 4.2 Using the Models

The logistic regression models can be used to determine whether or not the FSAs in actual data sets are too small. The *MaxCombs* value is computed based on the quasi-identifiers in the data set. For each FSA, its population value can be determined from the Statistics Canada population tables. With these two values we can estimate the probability that the proportion of uniques is above 5% or 20%. If the estimated probability is above 0.5, then that FSA must be suppressed or combined with another FSA in the data set.

Because the independent variables in the models were centred and scaled, this also has to be done when using the models for actual prediction. Let the *MaxCombs* value for a particular data set be denoted by  $M$ . We index the FSAs in a data set by  $j$ . Let the population size for a particular FSA in the data set be denoted by  $S_j$ .

We have the centered and scaled *MaxCombs* value:

$$M' = (M - 59861) / 10000 \dots\dots\dots (1)$$

and the centered and scaled population size value:



$$S'_j = \frac{(S_j - 21120)}{10000} \dots\dots\dots (2)$$

Then an FSA is considered to be high risk under the 5% threshold if the following condition is true:

$$\frac{1}{1 + e^{-(779.1 + 137.8M' - 37.3S'_j - 6.5MS'_j)}} > 0.5 \dots\dots\dots (3)$$

and an FSA is considered to be high risk under the 20% threshold if the following condition is true:

$$\frac{1}{1 + e^{-(63.3 + 11.8M' - 6S'_j - MS'_j)}} > 0.5 \dots\dots\dots (4)$$

For the FSAs that are flagged through equations (3) or (4) the options are to then either aggregate the FSAs or to suppress the records in those FSAs.

### 4.3 Application of Results

We applied the models to evaluate whether the FSA sizes in the newborn registry of Ontario (Niday) were appropriate. This registry captures information about all births. We used a data extract for 2006-2007. There were 124,933 births in the registry during that period. The quasi-identifiers that were considered were: baby’s date of birth, mother’s date of birth, baby’s gender, and the primary language spoken at home.

The proportion of records in the Niday registry that would have to be suppressed under each of the three thresholds was computed. We computed this proportion for every combination of 1, 2, 3, and 4 quasi-identifiers. For the 2 and 3 quasi-identifiers we averaged that proportion across the quasi-identifier combinations.

	0% Threshold	5% Threshold	20% Threshold
1 quasi-identifier	0.69	0.0078	0.0039
2 quasi-identifiers	0.85	0.0125	0.00634
3 quasi-identifiers	0.86	0.0126	0.0064
4 quasi-identifiers	0.86	0.0126	0.0064

**Table 4:** The proportion of Niday records that would have to be suppressed for each of the uniqueness thresholds.

The results of this analysis are shown in Table 4. As can be seen there is a pronounced difference between using the 0% threshold and the others, with far less data having to be suppressed for the 5% and 20% thresholds. These results demonstrate that, where the risk profile is acceptably low, using a higher threshold can result in significantly more data being made available.

#### 4.4 Selection of Threshold

An important decision when using the above models is selecting which of the three uniqueness threshold to use: 0%, 5%, or 20%. The most stringent uniqueness threshold of zero would be appropriate for data sets that are released to the public. This threshold would result in the most suppression and aggregation. The most permissive 20% threshold can be used when disclosing data to trusted recipients where the overall risks are quite low. This larger threshold would result in the least suppression and aggregation.

To assist with deciding which of the thresholds is most appropriate under a broad set of conditions, three general criteria have been proposed in the context of secondary use [58]:

- Mitigating controls that are in place at the data recipient's organization.  
Mitigating controls evaluates the extent to which the data recipient has good security and privacy practices in place. A recent checklist can be used for evaluating the extent to which mitigating controls have been implemented [59]. The fewer security and privacy practices that the data recipient has in place, the lower the threshold that should be used.
- The extent to which a disclosure (inadvertent or otherwise) constitutes an invasion of privacy for the patients.  
Figure 1 contains a checklist has been developed based on the literature [60-63]. The greater the risk of an invasion of privacy, the lower the threshold that should be used.
- The extent to which the data recipient is motivated and capable of re-identifying the data.  
Figure 2 contains a checklist has been developed based on the literature [64, 65]. The greater the risk that the data recipient is motivated and has the capacity to re-identify the database, the lower the threshold that should be used.

Admittedly, the use of these checklists remains qualitative, but they do provide a starting point for deciding what an appropriate threshold should be.

<b>Criteria for Evaluating the Risk of Invasion of Privacy</b>
<b>Sensitivity of the data</b>
<ul style="list-style-type: none"> <li>• The personal information in the database is highly detailed</li> <li>• The information in the database is of a highly sensitive personal nature (e.g., sexual attitudes, practices, and orientation; use of alcohol, drugs, or other addictive substances; illegal activities; suicide; sexual abuse; sexual harassment; mental health; certain types of genetic information; HIV status)</li> <li>• The information in the database comes from a sensitive context (e.g., data about individuals participating in a youth employment program are less sensitive than a similar list containing names and addresses of Hepatitis C and HIV compensation victims)</li> </ul>
<b>Appropriateness of Consent</b>
<ul style="list-style-type: none"> <li>• The conditions that were established at the time the information was first collected from the patients are consistent with the intended purpose of the recipient</li> <li>• Consent for secondary uses was obtained at the time the data was originally collected</li> <li>• The information was unsolicited or given freely or voluntarily by the patients with little expectation of it being maintained in total confidence</li> <li>• The custodian has sought consultation from well-defined groups or communities (e.g., minority groups, family groups, band leaders, Aboriginal people, people with disabilities, consumer associations, community representatives, patient advisory councils) regarding the disclosure</li> <li>• A strategy for informing/notifying the public about the secondary uses is in place (e.g., posters)</li> <li>• There was a commitment or promise made to the patients not to disclose the database to any third party or institution</li> <li>• The patients from whom the information was originally collected have previously objected to having their data used for this purpose</li> <li>• Obtaining consent from the individuals at this point is inappropriate or impractical (e.g., making contact to obtain consent may reveal the individual's condition to others against their wishes, the size of the population is too large, many patients have relocated or died, there is a lack of existing or continuing relationship with the patients, the consent procedure itself may introduce bias, there is a risk of inflicting psychological, social or other harm by contacting individuals and/or their families in delicate circumstances. It would be difficult to contact individuals through advertisements and other public notices, and undue hardship that would be caused by the additional financial, material, human, organizational or other resources required to obtain consent)</li> </ul>
<b>Potential Injury to Patients</b>
<ul style="list-style-type: none"> <li>• The database is large / many people would be affected if there was an inappropriate disclosure/breach</li> <li>• Inappropriate disclosure of the information carries a probability of causing measurable injury (e.g., identity theft, fraud, etc)</li> <li>• There is a risk in terms of the possible application of foreign laws</li> <li>• Inappropriate disclosure of the data may cause harm to individuals, a defined community (e.g., neighbourhood, minority groups) or a family (for example, physical injury, emotional or psychological harm, social harm such as stigmatization, financial harm such as employment or insurability)</li> </ul>

**Figure 1:** A checklist that can be used to evaluate the invasion of privacy risk.

<b>Criteria for Evaluating the Recipients Motives and Capacity to Re-identify the Data</b>
<b>Motives to Re-identify the Database</b>
<ul style="list-style-type: none"> <li>• The recipient has directly or indirectly worked/collaborated with the data custodian in the past</li> <li>• The database has potential commercial value (e.g., the recipient or his/her family may receive financial benefits from using the data; a pharmaceutical company may want to contact the patients directly for marketing purposes or to recruitment them in a study)</li> <li>• The disclosed database has potential criminal value (e.g., the database has dates of birth and mother's maiden name and can potentially be useful for financial crimes)</li> <li>• There is a likely non-commercial motive for the recipient to try to re-identify the disclosed database (e.g., a reporter or researcher making the point that the data is not safe, or to reveal health information about a famous person)</li> <li>• The recipient may want to harm or embarrass the data custodian</li> <li>• If the recipient does have a possible motive to attempt re-identification, they can achieve their objectives through other means apart from re-identification</li> </ul>
<b>Capacity to Re-identify the Database</b>
<ul style="list-style-type: none"> <li>• The recipient has had a data breach in the last two years</li> <li>• The recipient has the technical expertise to attempt to re-identify the disclosed database</li> <li>• The recipient has the financial resources to attempt to re-identify the disclosed database</li> </ul>

**Figure 2:** A checklist that can be used to evaluate motives and capacity of the data recipient to re-identify data.

## 4.5 Limitations

The FSAs that were included in our analysis were from urban areas in Canada. As described in the appendix, the reason is that the census tract information from the census file that we used is only defined for urban areas. Therefore, FSAs from rural areas were not covered.

The prediction models that we constructed were based on the uniqueness of the quasi-identifiers in FSAs. There is wide variation in FSA population size and we take that into account in our models. It is an empirical question whether our models are suitable for other areal units within the same population size range as urban FSAs.

Although we contend that the ten quasi-identifiers we considered represent basic demographics that are quite common in health research, they will not cover all possible quasi-identifiers that may be used in practice. Thus, our results are limited to the specific variables that we have considered in our analysis.

## 4.6 Relationship to Previous Work

The first study to examine uniqueness in the general population was conducted in the US by Sweeney [66]. In that study Sweeney only considered date of birth and gender as the

demographic variables. Since she did not have access to census microdata, she used publicly released tabulations that indicate the population per zip code. Relying on the generalized Dirichlet drawer problem she was able to make inferences about uniqueness in the population. For example, if we consider date of birth and gender, there are 2 (gender) x 365 (year) x 76 (life expectancy) = 55,480 possible values on these two variables. If a zip code has less than 55,480 individuals living in it then all individuals in that zip code were considered potentially at risk. Based on such calculations. She concluded that 87% of the US population are uniquely identifiable by their date of birth, gender, and zip code. However, because of the approach used in the analysis, this number would be expected to be an overestimate.

A subsequent analysis by Golle [67], again using publicly available census tabulations, assumed that births are uniformly distributed throughout the year. His analysis concluded that only 63% of the US population is uniquely identifiable with such simple demographics.

Neither of the above studies examined more than two quasi-identifiers, and did not specifically address the problem of determining the appropriate population size for the geographic area.

The earlier study which predicted when a geographic area is too small, was based on the zero uniqueness threshold, utilized a public use census file, and made a number of assumptions about the relationship between uniqueness and area size [35]. As opposed to this current study, it was constrained by the small sampling fraction of the public use file which made it difficult to directly estimate the Canadian population uniqueness if it was above zero.

## **5 Acknowledgements**

---

This work was funded by the GeoConnections (Natural Resources Canada), Public Health Agency of Canada, the Ontario Centers of Excellence, and the Natural Sciences and Engineering Research Council of Canada.

We would like to thank Dr. Mark Walker (OHRI) and Dr. Jim Bottomley (CHEO) for making the Niday data available for this analysis.

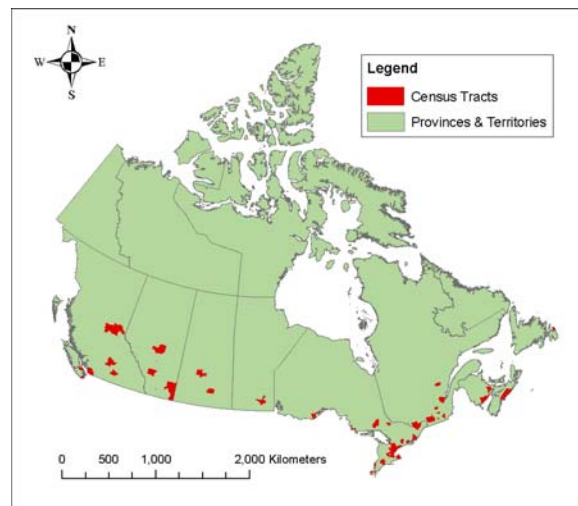
This work was approved by the research ethics board of the Children's Hospital of Eastern Ontario Research Institute.

## 6 Appendix A: Mapping Census Geography to Postal Geography Using a Gridding Methodology

---

### 6.1 Background

The smallest geographic unit provided in the census microdata file available through Statistics Canada's Research Data Centre (RDC) is the census tract (CT). CTs are only defined for census metropolitan areas and census agglomerations with urban core populations of at least 50,000 individuals. They are defined by Statistics Canada as "...small, relatively stable geographic areas that usually have a population of 2,500 to 8,000." [68]. The 2001 census contained a total of 4,798 CTs distributed over 9 provinces (no CTs are defined for the Territories or the province of PEI; see Figure 3).



**Figure 3:** Distribution of 2001 census tracts across Canada

In order to compute re-identification risk by Forward Sortation Area (FSA) in our current study, we needed to devise a method to estimate conversion between census and postal geography. A gridding methodology similar in nature to the Gridded Population of the World Project (GPW) [69] at the Center for International Earth Science Information Network at Columbia University [70] was utilized, allowing assignment of geography based on areal weighting using a population grid for Canada.

### 6.2 Methods

Population-based weights were assigned to CT-FSA unions based on a created population grid for all of Canada. The grid cell size was one kilometre by one kilometre, and assigned populations were based on the 2001 census profile at the dissemination area level (DA). This is

the smallest geography at which census profile information is released by Statistics Canada [71]. Similar to the PCCF+, these population weights were then used to randomly assign census tracts to their associated FSAs. Details of the steps taken to create the population grid are described below.

Twenty six (26) complete grids of dimensions 1554 by 546 Kilometres were created using a script in ESRI's ArcMap 9.2 [72], as specified in Table 5. This created 848,484 one kilometre square cells per grid, for a total of 22,909,068 cells covering the Canadian landmass.

Once the grids were created, the next task was to assign an estimated population to each cell. This was done using the Statistics Canada DA file [73]. First, all DA polygons identified as water were removed. A new DA shape file containing only land DAs was created. DA boundaries were then dissolved so that DAs with disparate polygons were captured within one record. Areas and perimeters were summed for each polygon to give the total DA area and perimeter. This reduced the number of records from 62,015 to 52,924, which matches the number of DAs as reported by Statistics Canada. Total population, as well as sex and age-stratified populations were extracted for all DAs across Canada, using four separate profile files (Western Canada and the Territories, Ontario, Quebec, and Atlantic Canada). Next, the 2001 DA population file was joined with the 2001 DA boundary file, to create a 2001 Canada DA boundary file containing total and sex and age stratified populations.

A "Select by attributes" function where population was not zero (0) was completed on the above file to create a new boundary file containing only DA polygons with reported populations. This further reduced the number of records to 49,153, creating a boundary file for non-water, populated DAs only. A "Select by location" function was completed on all 26 grids, for any cells that intersected the boundary file from the previous function. The resultant grids had a combined total cell count of 2,367,457.

A model was created using the ArcGIS model builder, and run for each of the 26 grids, to create grid section intersects with the 2001 DAs, FSAs and CTs. The model also calculated proportional grid sub-section areas and the corresponding population, based on underlying DA population and an assumption of uniform population distribution within each of the geographic areas.

A summary was done by each CT-FSA combination, to create unique CT-FSA records with the corresponding sum of the calculated grid-section populations. These summed populations were then divided by the total sum of the gridded-CT population to give the proportion of the population in each CT that lay within the corresponding FSA. In essence, this creates a population-based weight for each CT-FSA combination, allowing us to randomly assign any given record within a CT to its most likely (population-weighted) FSA.



A simplified hypothetical example of the end result is given in Table 6 and Figure 4. In this example, 64.07% of the population in CT16003 is found in FSA *K2S*, and 35.93% in FSA *K2T*. For CT 16004, 49.35% of its population is in *K2R*, 19.48% in *K2S* and 31.17% in *K2T*. This reduces the table to five rows, with a population-based weight for each unique CT-FSA combination. If, for example, there were then 28 records from the microdata file falling in CT 16003, 18 (~65.86%) would be allocated to *K2S*, and 10 (~34.14%) to *K2T*.

Grid Section	x	y	rows	columns	# Cells	# Cells (DA-clipped)	# Cells (populated DA-clipped)
00	-2341699	310266	1554	546	848,484	147,282	95,225
01	-1795699	310266	1554	546	848,484	323,759	292,052
02	-1249699	310266	1554	546	848,484	400,335	352,048
03	-703699	310266	1554	546	848,484	421,104	252,417
04	-157699	310266	1554	546	848,484	442,583	112,863
05	388301	310266	1554	546	848,484	444,187	47,006
06	934301	310266	1554	546	848,484	588,000	220,587
07	1480301	310266	1554	546	848,484	514,762	202,006
08	2026301	310266	1554	546	848,484	222,848	139,035
09	2572301	310266	1554	546	848,484	79,825	30,635
10	-2341699	1864266	1554	546	848,484	490,304	181,644
11	-1795699	1864266	1554	546	848,484	843,129	253,796
12	-1249699	1864266	1554	546	848,484	753,391	84,386
13	-703699	1864266	1554	546	848,484	749,156	802
14	-157699	1864266	1554	546	848,484	563,822	1,239
15	388301	1864266	1554	546	848,484	192,569	1,005
16	934301	1864266	1554	546	848,484	587,718	1,420
17	1480301	1864266	1554	546	848,484	342,289	683
18	2026301	1864266	1554	546	848,484	220,305	48,694
19	2572301	1864266	1554	546	848,484	55,829	25,720
20	-2341699	3418266	1554	546	848,484	21,506	0
21	-1795699	3418266	1554	546	848,484	168,942	531
22	-1249699	3418266	1554	546	848,484	135,498	686
23	-703699	3418266	1554	546	848,484	229,560	0
24	-157699	3418266	1554	546	848,484	424,214	1,101
25	388301	3418266	1554	546	848,484	258,726	210
26	934301	3418266	1554	546	848,484	26,188	160
<b>TOTAL</b>					<b>22,909,068</b>	<b>9,647,831</b>	<b>2,345,951</b>

Table 5: Canadian grid development table.

CT	FSAsa	FSAsa Pop Density (per Sq. Km.)	CT Area in FSA (Sq. Km.)	Pop	CT Pop	Weight
16003	K2S-1	50	0.95	48	128	0.3750
16003	K2S-2	25	0.56	14	128	0.1094
16003	K2S-3	42	0.48	20	128	0.1563
16003	K2T-1	20	1.23	25	128	0.1953
16003	K2T-2	56	0.37	21	128	0.1641
16004	K2R-1	37	1.03	38	77	0.4935
16004	K2S-1	42	0.36	15	77	0.1948
16004	K2T-2	56	0.42	24	77	0.3117

FSAsa = FSA sub-area

Pop = Population

Table 6: Simplified hypothetical example of the weighted association between CTs and FSAs.

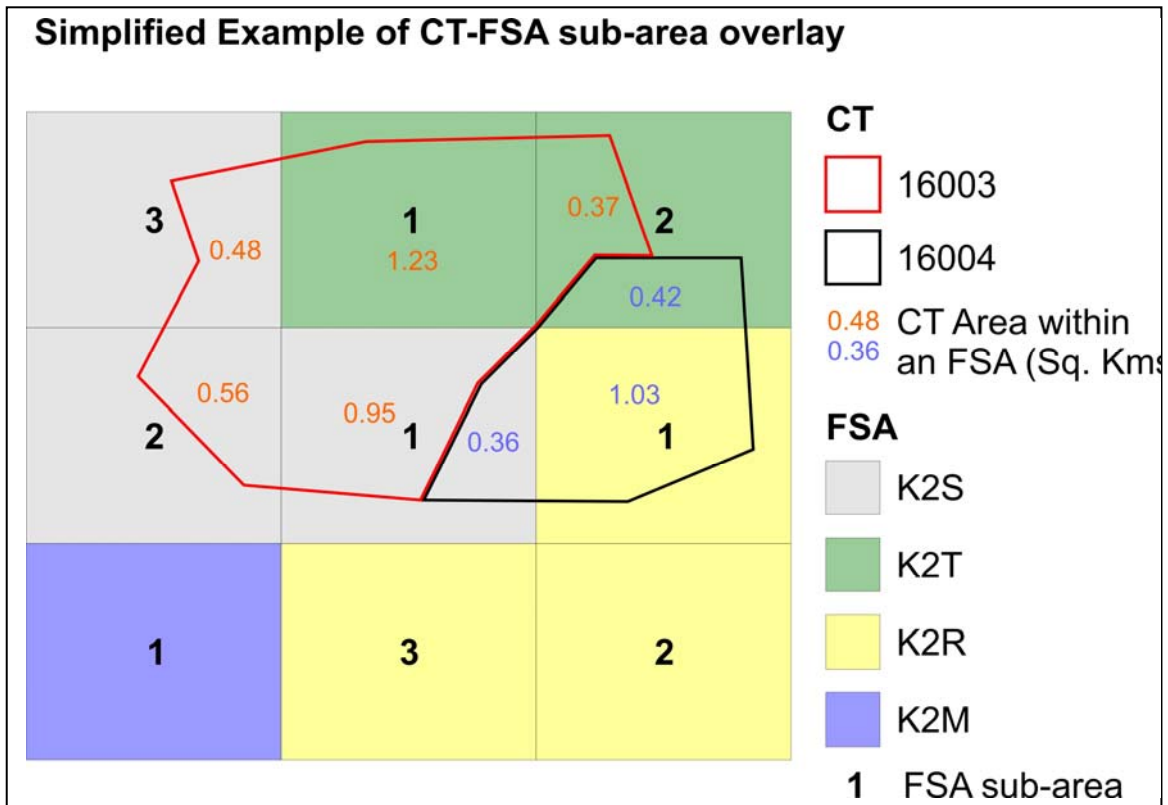


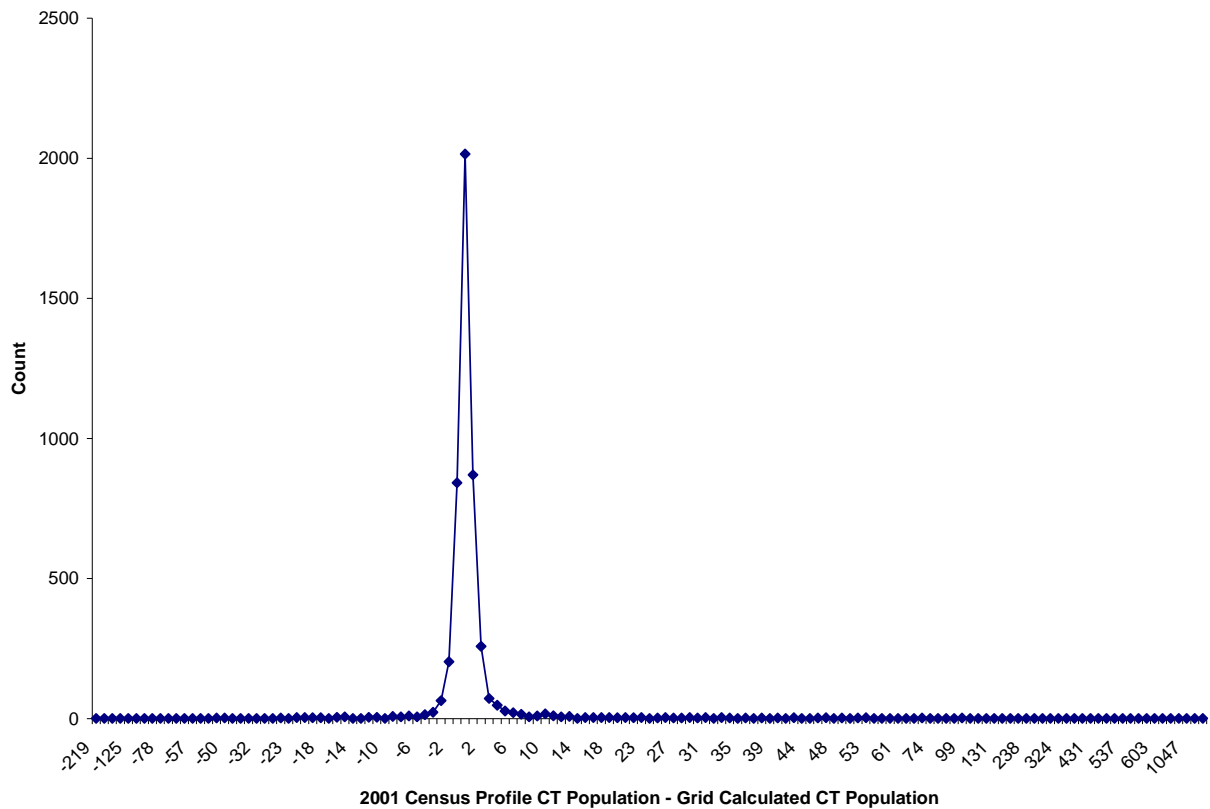
Figure 4: Example CT-FSA sub-area overlay to illustrate the hypothetical example.

### 6.3 Results

The CT population assignments based on the gridding methodology proved to be very similar to the 2001 Statistics Canada Census Tract population profile (Table 7). The mean difference between the populations was 3.45 individuals, with a standard deviation of 48.96 individuals (median was 0). A graphical representation of the distribution of the population differences, by census tracts, is given in Figure 5.

	2001 Statistics Canada Population Profile Census Tract	Canada Population Grid Project Census Tract
<b>Total n</b>	4757	4757
<b>Mean population</b>	4413.99	4410.54
<b>Standard Deviation</b>	1911.77	1911.33
<b>Minimum population</b>	40	0
<b>Median population</b>	4290	4287
<b>Maximum population</b>	20635	20636

**Table 7:** Census tract population comparison between created population grid and 2001 census profile.



**Figure 5:** Distribution of Census Tract Population Difference between Grid-Calculated Population and 2001 Census Profile.

Provincial analyses also showed a high concordance between the CT populations using the gridding methodology as compared to the 2001 Statistics Canada Census Tract population profile (Table 8). The greatest differences were in New Brunswick (mean difference = 6.97 individuals, standard deviation = 75.26 individuals) and Alberta (mean difference = 6.75 individuals, standard deviation = 81.67 individuals).

	NL	NS	NB	QC	ON	MB	SK	AB	BC
<b>N</b>	45	85	70	1246	2001	164	101	449	596
<b>Mean</b>	3.71	2.6	6.97	1.55	3.68	2.93	-1.18	6.75	4.79
<b>Std Dev</b>	12.01	19.37	75.26	26.19	51.38	27.14	37.86	81.67	51.38
<b>Median</b>	0	0	0	0	0	0	0	0	0

**Table 8:** Provincial differences between Profile and grid CT populations.

## 6.4 Conclusions

The population grid created in this study provides a means for linking census geography to postal geography in Canada. While creating population grids in and of itself is not a novel idea, the created grid in this project allows the mapping of census geography to postal geography, based on population weights. The procedure assumes a uniform population distribution within the geography being used. However, since CTs only occur in highly populated urban areas, this was felt to be an appropriate assumption. A similar assumption would not hold in rural or less densely populated areas, and this technique would therefore not be appropriate. However, it could be utilized, and further refined, by incorporating additional information, such as ecumene areas, satellite imagery for residential and inhabited areas, address data, etc.

## 7 References

---

1. Safran C, Bloomrosen M, Hammond E, Labkoff S, S K-F, Tang P, Detmer D. *Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper*. Journal of the American Medical Informatics Association, 2007; 14:1-9.
2. Roy D, Fournier F. *Secondary use of personal information held on national electronic health record systems*. 2007; Centre for Bioethics, Clinical Research Institute of Montreal.
3. Kosseim P, Brady M. *Policy by procrastination: Secondary use of electronic health records for health research purposes*. McGill Journal of Law and Health, 2008; 2:5-45.
4. Black C, McGrail K, Fooks C, Baranek P, Maslove L. *Data, Data, Everywhere -- Improving access to population health and health services research data in Canada*. 2005; Centre for Health Services and Policy Research and Canadian Policy Research Networks.
5. Willison D, Gibson E, McGrail K. *A roadmap to research uses of electronic health information*. CIHR Health Information Summit. 2008.
6. Boulos M. *Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom*. International Journal of Health Geographics 2004; 3(1).
7. O'Dwyer LA, Burton DL. *Potential meets reality: GIS and public health research in Australia*. Australian and New Zealand Journal of Public Health, 1998; 22(7):819-823.
8. Ricketts TC. *Geographic information systems and public health*. Annual Review of Public Health, 2003; 24:1-6.
9. Cromley EK. *GIS and Disease*. Annual Review of Public Health, 2003; 24:7-24.
10. Brindley P, Maheswaran R. *My favourite software: geographic information systems*. Journal of Public Health Medicine, 2002; 24(2):149.
11. Richards TB, Croner CM, Rushton G, Brown CK, Fowler L. *Geographic information systems and public health: mapping the future*. Public Health Reports, 1999; 114:359-373.
12. Ricketts T. *Geographic information systems and public health*. Annual Review of Public Health, 2003; 24:1-6.
13. McLafferty S. *GIS and health care*. Annual Review of Public Health, 2003; 24:25-42.
14. Cromley E. *GIS and disease*. Annual Review of Public Health, 2003; 24:7-24.
15. Mugge R. *Issues in protecting confidentiality in national health statistics*. Proceedings of the Social Statistics Section, American Statistical Association. 1983.
16. Mackie C, Bradburn N. *Improving access to and confidentiality of research data: Report of a workshop*. 2000: The National Academies Press.
17. Croner C. *Public health, GIS, and the Internet*. Annual Review of Public Health, 2003; 24:57-82.
18. Mr. Justice Gibson. *Mike Gordon and The Minister of Health and Privacy Commissioner of Canada*. February 27, 2008; Federal Court of Canada.

19. Hawala S. *Enhancing the "100,000" rule: On the variation of percent of uniques in a microdata sample and the geographic area size identified on the file*. Proceedings of the Annual Meeting of the American Statistical Association 2001.
20. Greenberg B, Voshell L. *Relating risk of disclosure for microdata and geographic area size*. Proceedings of the Section on Survey Research Methods, American Statistical Association. 1990.
21. Greenberg B, Voshell L. *The geographic component of disclosure risk for microdata*. 1990; Bureau of the Census.
22. Zayatz L, Massell P, Steel P. *Disclosure limitation practices and research at the US Census Bureau*. Netherlands Official Statistics, 1999; 14(Spring):26-29.
23. Zayatz L. *Disclosure avoidance practices and research at the US Census Bureau: An update*. 2005; US Census Bureau.
24. Hawala S. *Microdata disclosure protection research and experiences at the US census bureau*. Proceedings of Workshop on Microdata. 2003; Available from: [\[http://www.census.gov/srd/sdc/microdataprotection.pdf\]](http://www.census.gov/srd/sdc/microdataprotection.pdf). Archived at: [\[http://www.webcitation.org/5b7mPeVPi\]](http://www.webcitation.org/5b7mPeVPi).
25. Marsh C, Dale A, Skinner C. *Safe data versus safe settings: Access to microdata from the British census*. International Statistical Review, 1994; 62(1):35-53.
26. Statistics Canada. *Canadian Community Health Survey (CCHS) Cycle 3.1 (2005) Public Use Microdata File (PUMF) User Guide*. 2006.
27. Willenborg L, de Waal T. *Statistical Disclosure Control in Practice*. 1996; Springer-Verlag.
28. Fefferman N, O'Neil E, Naumova E. *Confidentiality and confidence: Is data aggregation a means to achieve both ?* Journal of Public Health Policy, 2005; 16:430-449.
29. Willenborg L, Mokken R, Pannekoek J. *Microdata and disclosure risks*. Proceedings of the Annual Research Conference of US Bureau of the Census. 1990.
30. Olson K, Grannis S, Mandl K. *Privacy protection versus cluster detection in spatial epidemiology*. American Journal of Public Health, 2006; 96(11):2002-2008.
31. Marceau D. *The scale issue in social and natural sciences*. Canadian Journal of Remote Sensing, 1999; 25(4):347-356.
32. Bivand R. *A review of spatial statistical techniques for location studies*. 1998; Norwegian School of Economics and Business Administration.
33. Ratcliffe J. *The Modifiable Areal Unit Problem*. 2005; Available from: [\[http://www.jratcliffe.net/research/maup.htm\]](http://www.jratcliffe.net/research/maup.htm).
34. Bethlehem J, Keller W, Pannekoek J. *Disclosure control of microdata*. Journal of the American Statistical Association, 1990; 85(409):38-45.
35. El Emam K, Brown A, Abdelmalik P. *Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk*. Journal of the American Medical Informatics Association, 2009; (to appear).
36. Howe H, Lake A, Shen T. *Method to assess identifiability in electronic data files*. American Journal of Epidemiology, 2007; 165(5):597-601.
37. Howe H, Lake A, Lehnerr M, Roney D. *Unique record identification on public use files as tested on the 1994-1998 CINA analytic file*. 2002; North American Association of Central Cancer Registries.
38. El Emam K. *Heuristics for de-identifying health data*. IEEE Security and Privacy, 2008:72-75.

39. Dalenius T. *Finding a needle in a haystack or identifying anonymous census records*. Journal of Official Statistics, 1986; 2(3):329-336.
40. El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. *Evaluating common de-identification heuristics for personal health information*. Journal of Medical Internet Research, 2006; 8(4):e28.
41. El Emam K, Jonker E, Sams S, Neri E, Neisa A, Gao T, Chowdhury S. *Pan-Canadian De-Identification Guidelines for Personal Health Information (report prepared for the Office of the Privacy Commissioner of Canada)*. 2007; Available from: [<http://www.ehealthinformation.ca/documents/OPCReportv11.pdf>]. Archived at: [<http://www.webcitation.org/5Ow1Nko5C>].
42. Bow C, Waters N, Faris P, Seidel J, Galbraith P, Knudtson M, Ghali W. *Accuracy of city postal code coordinates as a proxy for location of residence*. International Journal of Health Geographics, 2004; 3(5).
43. Ng E, Wilkins R, Perras A. *How far is it to the nearest hospital ? Calculating distances using the Statistics Canada Postal Code Conversion file*. Health Reports, 1993; 5:179-183.
44. Mackillop W, Zhang-Salomons J, Groome P, Pazat L, Holowaty E. *Socioeconomic status and cancer survival in Ontario*. Journal of Clinical Oncology, 1997; 15:1680-1689.
45. Spasoff A, Gilkes D. *Up-to-date denominators: Evaluation of taxation family for public health planning*. Canadian Journal of Public Health, 1994; 85:413-417.
46. Demissie K, Hanley J, Menzies D, Joseph L, Ernst P. *Agreement in measuring socio-economic status: Area-based versus individual measures*. Chronic Diseases in Canada, 2000; 21:1-7.
47. Guernsey J, Dewar R, Weerasinghe S, Kirkland S, Veugelers P. *Incidence of cancer in sydney and Cape breton County, Nova Scotia 1979-1997*. Canadian Journal of Public Health, 2000; 91:285-292.
48. Skinner C, Holmes D. *Estimating the re-identification risk per record in microdata*. Journal of Official Statistics, 1998; 14(4):361-372.
49. Chen G, Keller-McNulty S. *Estimation of identification disclosure risk in microdata*. Journal of Official Statistics, 1998; 14(1):79-95.
50. Zayatz L. *Estimation of the percent of unique population elements on a microdata file using the sample*. 1991; US Bureau of the Census.
51. Hosmer D, Lemeshow S. *Applied Logistic Regression*. 1989: John Wiley & Sons.
52. Jaccard J. *Interaction Effects in Logistic Regression*. 2001: Sage Publications.
53. King G, Zeng L. *Logistic regression in rare events data*. Political Analysis, 2001; 9(2):137-163.
54. Lowe W. *Rare events research*, in *Encyclopedia of Social Measurement*, K. Kempf-Leonard, Editor. 2004; Academic Press.
55. Ruiz-Gazen A, Villa N. *Storms prediction: Logistic regression vs. random forests for unbalanced data*. Case Studies in Business, Industry and Government Statistics, 2007; 1(2):91-101.
56. Cherkassky V, Muller F. *Learning from data*. 1998: Wiley.
57. Alpaydin E. *Introduction to machine learning*. 2004: MIT Press.
58. El Emam K. *De-identifying health data for secondary use: A framework*. 2008; Available from: [<http://www.ehealthinformation.ca/documents/SecondaryUseFW.pdf>].



59. El Emam K, Dankar F, Vaillancourt R, Roffey T, Lysyk M. *Evaluating patient re-identification risk from hospital prescription records*. Canadian Journal of Hospital Pharmacy (to appear), 2009.
60. Treasury Board of Canada Secretariat. *Privacy impact assessment guidelines: A framework to manage privacy risks*. 2002; Government of Canada.
61. Treasury Board of Canada Secretariat. *Guidance document: Taking privacy into account before making contracting decisions*. 2006; Government of Canada.
62. Canadian Institutes of Health Research. *CIHR best practices for protecting privacy in health research*. 2005; Canadian Institutes of Health Research.
63. Canadian Institutes of Health Research. *Secondary use of personal information in health research: Case studies*. 2002.
64. Elliot M, Dale A. *Scenarios of attack: the data intruders perspective on statistical disclosure risk*. Netherlands Official Statistics, 1999; 14(Spring):6-10.
65. Sweeney L. *Guaranteeing anonymity when sharing medical data: The Datafly system*. Proceedings of the American Medical Informatics Association Symposium. 1997.
66. Sweeney L. *Uniqueness of Simple Demographics in the US Population*. 2000; Carnegie Mellon University, Laboratory for International Data Privacy.
67. Golle P. *Revisiting the uniqueness of simple demographics in the US population*. Workshop on Privacy in the Electronic Society. 2006.
68. Statistics Canada. *Cartographic boundary files: 2001 census*. 2002.
69. Yetman G, Deichmann U, Balk D. *Creating a global grid of human population*. [cited; Available from: <http://gis.esri.com/library/userconf/proc00/professional/papers/PAP552/p552.htm>. Archived at: [Web [69]].
70. *Center for International Earth Science Information Network (CIESIN)*. [cited; Available from: <http://beta.sedac.ciesin.columbia.edu>. Archived at: [Web [70]].
71. Statistics Canada. *Profile of all levels of geography in Canada, 2001 census*. 2003.
72. Nicholas R. *ESRI Support Center: Create a grid polygon shapefile (FISHNET)*. 2003; Available from: [<http://arcscripts.esri.com/details.asp?dbid=12807>].
73. Statistics Canada. *Dissemination Areas Cartographic Boundary Files (Geography Products: Spatial Information Products, 2001 Census)*. 2002.

*Model Formulation* ■

# Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk

KHALED EL EMAM, ANN BROWN, PHILIP ABDELMALIK

**Abstract** **Objective:** In public health and health services research, the inclusion of geographic information in data sets is critical. Because of concerns over the re-identification of patients, data from small geographic areas are either suppressed or the geographic areas are aggregated into larger ones. Our objective is to estimate the population size cut-off at which a geographic area is sufficiently large so that no data suppression or further aggregation is necessary.

**Design:** The 2001 Canadian census data were used to conduct a simulation to model the relationship between geographic area population size and uniqueness for some common demographic variables. Cut-offs were computed for geographic area population size, and prediction models were developed to estimate the appropriate cut-offs.

**Measurements:** Re-identification risk was measured using uniqueness. Geographic area population size cut-offs were estimated using the maximum number of possible values in the data set and a traditional entropy measure.

**Results:** The model that predicted population cut-offs using the maximum number of possible values in the data set had  $R^2$  values around 0.9, and relative error of prediction less than 0.02 across all regions of Canada. The models were then applied to assess the appropriate geographic area size for the prescription records provided by retail and hospital pharmacies to commercial research and analysis firms.

**Conclusions:** To manage re-identification risk, the prediction models can be used by public health professionals, health researchers, and research ethics boards to decide when the geographic area population size is sufficiently large.

■ *J Am Med Inform Assoc.* 2009;16:256–266. DOI 10.1197/jamia.M2902.

## Introduction

Privacy legislation in Canada applies to identifiable information. This means that if health information is deemed sufficiently de-identified, then there is no legislative requirement to obtain consent from patients to collect it and use it.<sup>1</sup> In addition, Research Ethics Boards (REBs) are more likely to waive the consent requirement if the information collected

for research is deemed de-identified.<sup>2</sup> The option to waive consent is important as there is evidence that currently used methods for obtaining opt-in consent can result in low recruitment and selection bias in health research.<sup>3–10</sup> The ability to make precise claims about identifiability therefore is needed to inform this consent waiver decision.

It is obvious that variables such as name and address would have to be removed, or not collected to start off with, to de-identify a data set. However, beyond the elimination of such variables, the definition of identifiability is often vague and remains an active area of research.<sup>11</sup>

The inclusion of geographic information (geocoding) in health data sets is critical for public health investigations and health services research.<sup>12–17</sup> However, the inclusion of geographic details in a data set also makes it much easier to re-identify patients.<sup>18,19</sup> The more specific the geographic detail included, the easier it is to use the other variables/information in the data to uniquely identify an individual. In fact, recently the federal court accepted evidence that the inclusion of the “Province” field in Health Canada’s adverse drug events database can potentially re-identify individuals.<sup>20</sup> Therefore, the province where the adverse event occurred cannot be disclosed by Health Canada in response to an access to information request. It has also been shown that patient addresses can be re-identified from published maps.<sup>21–23</sup> Consequently, there is a risk that geographic detail in health data sets makes Canadians identifiable.

To protect privacy one can mask geocodes,<sup>24,25</sup> or control geographic area population size (GAPS) to minimize the risk of re-identification. Due to its relative simplicity, controlling

---

Affiliations of the authors: Children’s Hospital of Eastern Ontario Research Institute (KEE, AB), Ottawa, ON, Canada; Pediatrics, Faculty of Medicine, University of Ottawa (KEE), Ottawa, ON, Canada; GIS Infrastructure, Office of Public Health Practice, Public Health Agency of Canada (PA), Ottawa, ON, Canada

This work was funded by the Public Health Agency of Canada, the Ontario Centers of Excellence, GeoConnections (Natural Resources Canada), and the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank Anita Fineberg from IMS Health Canada, Inc for providing us with information about the record layout for the prescription data. The authors also would like to thank David Paton (Canadian Institute for Health Information), Bradley Malin (Vanderbilt University), Jean-Louis Tambay (Statistics Canada), and Don Willison (McMaster University) for their detailed feedback on an earlier version of this paper. Comments from the anonymous review were also of considerable help in improving and clarifying the paper.

This work was approved by the research ethics board of The Children’s Hospital of Eastern Ontario Research Institute.

Correspondence: Khaled El Emam, CHEO Research Institute, 401 Smyth Road, Ottawa, ON K1H 8L1, Canada; e-mail: <kelemam@uottawa.ca>.

Received for review: 06/18/2008; accepted for publication: 11/30/2008.

GAPS has been adopted widely in practice. Controlling GAPS means either that data about individuals living in areas with small populations are suppressed, or that areas with small populations are aggregated into larger ones. Suppression results in the direct loss of data, and aggregation reduces the utility of a data set.<sup>26–28</sup> This is justified because of the demonstrated empirical relationship between GAPS and re-identification risk<sup>29–31</sup>: re-identification risk tends to be higher in areas with smaller populations.

Examples of GAPS cut-off use include the United States Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. The HIPAA Privacy Rule defines 18 variables in the Safe Harbor List that need to be removed or generalized to ensure that a data set is de-identified. One of these 18 items stipulates that the first three numbers of the ZIP code can be collected/disclosed if the population living within that geographic area is greater than 20,000 people. The US Bureau of the Census has a 100,000 GAPS cut-off for releasing public use microdata files.<sup>32–34</sup> That same cut-off is used for making disclosure control decisions with public health data sets.<sup>35,36</sup> Only data from areas with a population of 120,000 or more are released as microdata from the British census.<sup>37</sup> Similarly, Statistics Canada uses a 70,000 population size cut-off for health regions to control the risk of disclosure when releasing data from the Canadian Community Health Survey (CCHS).<sup>38</sup> It has been suggested that different GAPS cut-offs should be applied depending on the user, with a 25,000 cut-off for data disclosed to researchers, and a 100,000 cut-off for data disclosed to the public.<sup>39</sup>

The dearth of evidence supporting the specific cut-offs that are used in practice, and the “real research need to develop empirical evidence to justify recommendations regarding geographic specificity”<sup>19</sup> make the continued search for GAPS cut-offs important. Furthermore, existing GAPS cut-offs do not account for the fact that a cut-off is inherently dependent on the number and nature of the variables under consideration.<sup>31,40</sup> For example, the cut-off to apply when one has two variables will be smaller than a cut-off to apply when there are 15 variables. When the variables have few response categories, the cut-off will be smaller than when they have many response categories. Therefore, many GAPS cut-offs in current use (summarized above), may be over-protecting data sets or not protecting them enough depending on the specific variables in question.

The purpose of our study is to provide an empirically grounded basis for using GAPS cut-offs. The primary contributions of this work are to (a) provide models for predicting the GAPS cut-offs that explicitly account for re-identification risk and the variable characteristics based on two simple metrics: the number of possible combinations of data fields and entropy, (b) validating these models using Canadian census data, and (c) demonstrating their applicability with two examples of pharmacy prescription data.

## Methods

### Definitions and Preliminaries

#### *Quasi-identifiers*

When considering re-identification risk, we are only interested in a subset of variables in a data set.<sup>41</sup> These are called the quasi-identifiers.<sup>42</sup> They are variables that make individ-

uals unique in the population and are possibly publicly known. Therefore, they do not directly identify an individual, but can be used for indirect re-identification. While there is no universal definition of what constitutes a quasi-identifier, there are some quasi-identifiers that have been studied more extensively than others such as gender, date of birth, ethnicity, income, years of education, and geocodes. In addition, quasi-identifiers may differ across data sets. For example, gender will not be a meaningful quasi-identifier if all of the individuals in a data set are female. Lastly, in this study, the quasi-identifiers that are assessed have a finite set of possible discrete values.

#### *Uniqueness as a Measure of Re-identification Risk*

We define a unique individual as the one individual with specific values on the quasi-identifiers in a particular geographic area. For example, if there is only one 95-year-old male in a postal code, then that individual is unique within that postal code. The uniqueness of individuals is often used as a surrogate measure for re-identification risk: unique records in a data set are more likely to be re-identified by an intruder than non-unique records.<sup>43</sup> We therefore use uniqueness as our measure of re-identification risk.

#### *Nested Geographic Areas*

Geographic area aggregation implies a nesting relationship among those areas. For example, if we decide that re-identification risk is too high when we geocode using full postal codes, then we can aggregate the geographic area to Forward Sortation Areas (FSA), which are the first three characters of the postal code. Postal codes are nested within FSAs.

### Determining the GAPS Cut-offs

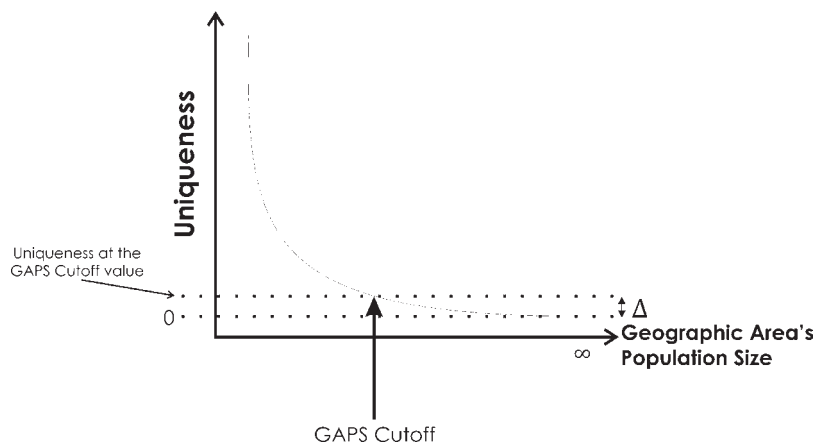
Geographic areas can be measured in terms of the physical area or population size. In this paper we refer only to the population size of the geographic area.

Previous research has identified two characteristics of the relationship between uniqueness and GAPS:<sup>29–31</sup>

- Uniqueness in a data set is inversely proportional to the population size of the geographic area. This means that the proportion of unique individuals in a large area will be smaller than in a nested smaller area. As smaller areas are aggregated into larger areas, the proportion of uniques goes down (see Fig 1).
- Once GAPS reaches a certain point, uniqueness tends to plateau. This trend applies irrespective of the quasi-identifiers in question.

A case has been made that the 100,000 GAPS cut-off used by the Census Bureau is justified by computing the uniqueness plateau noted above (i.e., the point at which uniqueness no longer changes).<sup>29</sup> The rationale is that increasing the size of the geographic area any further has little impact on uniqueness, and hence little impact on re-identification risk.<sup>29–31</sup> For example, if the uniqueness plateau is reached at 100,000 then this means the re-identification risk changes insignificantly between 100,000 and 110,000. Therefore, there is no disclosure control benefit in increasing the size of the geographic region or of aggregation beyond 100,000, and a reasonable cut-off would be 100,000.

In our analysis we build on a methodology used in a previous study at the Census Bureau<sup>29,31</sup> and proceed as follows:



**Figure 1.** Illustration of how the GAPS cutoff is calculated. Uniqueness is computed as the proportion of individuals who are unique on the values of the quasi-identifiers. For example, a uniqueness of 0.02 for a geographic area of 10,000 individuals on age, ethnicity, and gender means that 200 individuals have unique values on the combination of these three variables. At the limit, with an infinitely sized area, the uniqueness approaches zero. The delta value is the uniqueness at the GAPS cutoff value.

- Define a quasi-identifier model as a specific quasi-identifier or combination of quasi-identifiers and evaluate its uniqueness.
- Plot uniqueness against GAPS and compute the cut-off point as the point where the derivative approaches zero (illustrated in Fig 1).

Let the geographic areas under consideration be indexed by  $1..K$ , and their population size denoted by  $S_i$  where  $i:1..K$ . The area indexed by  $i$  is nested within the area indexed by  $i+1$ . Consequently, we also have  $S_i < S_{i+1}$  for all  $i$ . We denote the percentage of individuals on a particular quasi-identifier model that are unique in an area  $i$  by  $U(S_i)$ . Because of the monotonically decreasing relationship between GAPS and uniqueness, we expect the following relationship to hold:  $U(S_i) > U(S_{i+1})$ . The GAPS cut-off was then defined as the value of  $S_i$  where the approximate derivative, the change in the percentage of uniques, is close to zero<sup>31</sup>:

$$GAPS\_CUTOFF = S_i \left| \left\{ \frac{(U(S_i) - U(S_{i+1}))}{(S_{i+1} - S_i)} \cong 0 \right\} \right. \quad (1)$$

This approach, however, may identify local plateaus where the uniqueness remains temporarily steady, followed by a more substantial decrease to reach the asymptotic value. To address this we adopted a model building approach where the uniqueness function is defined as  $U(S_i) = \beta_0 \times S_i^{\beta_1}$ , where the  $\beta_0$  and  $\beta_1$  are estimated using ordinary least squares regression. We then take the derivative of this function and compute the cut-off as the size value where the derivative approaches zero:

$$GAPS\_CUTOFF = S_i \left| \left\{ \beta_0 \times \beta_1 \times S_i^{(\beta_1 - 1)} \cong 0 \right\} \right. \quad (2)$$

The cut-off values were computed separately for central Canada (which includes Ontario and Quebec), western Canada (which includes all territories and provinces west of Ontario), and eastern Canada (which includes all provinces east of Quebec).

#### Data Source

The data set used for our study is the 2001 Canadian census Public Use Microdata File (PUMF) made available by Statis-

tics Canada.<sup>44</sup> The PUMF represents approximately 2.7% of the Canadian population. The variable subset that is analyzed is shown in Table 1. These are common demographics that are often available in health data sets. There are 10 quasi-identifiers. These variables were selected because they can be used to link with other databases, because they describe attributes which are visible on individuals, or because they describe attributes which would make individuals easily identifiable.<sup>41</sup>

Disclosure control was already applied to the PUMF by Statistics Canada. The specifics that are relevant to this study consist of: (a) suppression for some variables for the Eastern region of Canada, and (b) the age variable was top coded at 85 years. As a result, there were three variables in the Eastern region, as seen in Table 1, which corresponded to variables in the West and Central regions but with a smaller number of response categories, where these response categories were coarsened.

#### Quasi-identifier Models

A quasi-identifier model consists of one or more quasi-identifiers (qids). To manage the scope, we only consider combinations of up to five quasi-identifiers.

There are some similarities among the ethnicity related variables, and therefore they were treated as a group: variables ETHNICRA, HLNPA, RELIGRPA, VISMNP. Whenever the ethnicity variable appears in a model it was replaced by one of the above individual variables. Each substitution represented a different model. This gives 7 distinct qids: sex, age, ethnicity, schooling, marital status, total income, and aboriginal identity.

Categorizing the 7 distinct qids by their sensitivity and availability to an intruder gives the following two types:

- Easily used and available for re-identification: sex and age
- Possibly usable for re-identification/sensitive: ethnicity, schooling, marital status, total income, and aboriginal identity

Table 1 ■ Quasi-identifiers to be Included in the Models for the Three Regions of Canada

Variable Name in the Census File	Definition	Number Response Categories*	
		Western and Central Canada	Eastern Canada
SEXP	sex	2	2
AGEP	single years of age from 0 to 84, 85+	86	86
HLNPA	language: the language spoken most often at home by the individual	14	4
ETHNICRA	ethnic or cultural group to which respondent's ancestors belong	41	26
ABSRP	aboriginal identity	4	4
TOTSCHP	total years of schooling	9	9
MARST	marital status (legal)	5	5
RELIGRPA	religious denomination	11	3
TOTINCP	total income: we defined categories of total income in \$ 15-K intervals	11	11
VISMIMP	visible minority	4	4

\*The Number of response categories excludes nonspecific responses such as missing value, not available, or "other".

The value for  $C_r^n$  gives the number of possible combinations of size  $r$  from a larger group of size  $n$ . The different models will be defined by the number of qids in the model with both age and gender being included in each model. That is, models containing:

- 5 qids: have age and gender and 10 combinations of 3 of the 5 sensitive qids.
- 4 qids: have age and gender and 10 combinations of 2 of the 5 sensitive qids.
- 3 qids: have age and gender and each of the 5 sensitive qids.
- 2 qids: have age and gender only—there is only one model.

This gives 26 models for the 7 distinct qids. Substituting each of home language, religion and visible minority for ethnicity then gives us 18 ( $3 \times 6$ ) models for 5 qids (ethnicity appears in 6 of the 10 models), 12 ( $3 \times 4$ ) models for 4 qids (ethnicity appears in 4 of the 10 models), and 3 ( $1 \times 3$ ) models for 3 qids. The subtotal for this group is 59 models.

We repeated the above process by using each one of age or gender in combination with the sensitive qids. That is, models containing:

- 5 qids: have age and 5 combinations of 4 of the 5 sensitive qids.
- 4 qids: have age and 10 combinations of 3 of the 5 sensitive qids.
- 3 qids: have age and 10 combinations of 2 of the 5 sensitive qids.
- 2 qids: have age and each of the 5 sensitive qids only.

This gives 30 models. Similarly to the previous group, by taking into account the ethnicity related variables gives a subtotal for this group of 75 models. For the last group, age is replaced with gender for an additional 75 models.

Therefore, in total we tested 209 different quasi-identifier models.

### Varying Region Size

We performed a simulation following the nested sampling method described by Greenberg and Voshell.<sup>30,31</sup> We took a simple random sample of 200,000 individuals from western Canada, 200,000 from central Canada, and 60,000 from

eastern Canada. For each of these three regions of Canada, we varied the size of the region by randomly removing individuals in 5,000 decrements. For example, for central Canada, we started with a random sample of 200,000 individuals, then a subsample of 195,000 was randomly selected, and then another subsample with 190,000 individuals, and so on. For each subsample we computed the proportion of unique records on each of the 209 quasi-identifier models described above. The cut-off was selected when the derivative was less than 0.001 using Eq (2).

This simulation approach has been shown to produce results that are quite similar to using actual contiguous areas (e.g., Census Tracts).<sup>30,31</sup> Furthermore, it has been argued that this simulation approach ensures that the results are controlled, replicable, and generalizable.<sup>31</sup>

When computing the cut-off using the derivative (Eq 2), the potential cut-offs were evaluated only within the GAPS range in our data set (i.e., 5–200 k for western and central Canada, and 5–60 k for eastern Canada) to ensure that we did not extrapolate beyond the original data used to build the models.

### Predicting the GAPS Cut-off

We developed a prediction model to have the results of the simulation be more practical for an end-user, such as a privacy analyst or epidemiologist, to calculate the GAPS cut-off for their particular study or data set. As noted earlier, we expected that a cut-off is related to the quasi-identifiers that are being considered. The following are two traditional ways used to characterize the quasi-identifiers:

**Entropy.** A previous study formulated an entropy measure that captures the dispersion in the quasi-identifiers.<sup>31</sup> This was found to be strongly related to uniqueness within a region. We computed the standard information theoretical entropy measure from the full samples using  $-\sum_{k=1}^L t_k \times (k/N) \times \log(k/N)$  where  $t_k$  is the number of equivalence classes of size  $k$ ,  $L$  is the size of the largest equivalence class, and  $N$  the total number of records in the sample. An equivalence class is defined as a possible value on the quasi-identifiers, for example, "50 year old male" is an equivalence class. We found that entropy computed from sub-samples were very strongly correlated, therefore, they produce similar results as full sample entropy.

**MaxCombs.** The maximum number of possible different values for the quasi-identifiers. For example, if we have two quasi-identifiers, say, age and gender, and assume that age has 86 possible values and gender has 2 values,  $86 \times 2 = 172$  is the maximum number of different possible combinations of values for these two quasi-identifiers. It is expected that the greater the maximum number of combinations the more uniques will be in a data set.<sup>31</sup>

We constructed two prediction models, each with a single independent variable: Entropy, or MaxCombs. An examination of the data indicated an obvious logarithmic relationship between each of these variables and the GAPS cut-off, giving us the following two linear models:  $\log(\text{GAPS\_CUTOFF}) \sim \beta_0 + \beta_1 \log(\text{Entropy})$  and  $\log(\text{GAPS\_CUTOFF}) \sim \beta_0 + \beta_1 \log(\text{MaxCombs})$ . For each of the two prediction models we had 209 observations representing the quasi-identifier models.

The GAPS cut-off value is truncated from below at 5,000 because that is the smallest subsample that was selected. It is also truncated at the top at 200,000 for central and western Canada, and 60,000 for eastern Canada because that was the size of the total sample that we used. Neither Entropy nor MaxCombs is truncated. A suitable modeling technique for such a censored data set is Tobit regression.<sup>45-47</sup>

Let  $y$  denote the actual value of the GAPS cut-off, the point at which the approximate derivative is close to zero, produced during our simulations. We have  $y \geq c_1$  and  $y \leq c_2$ , where  $c_1$  and  $c_2$  are the bottom and top truncation threshold values respectively. Also, let there be an underlying latent variable  $y^*$  of which  $y$  is the realized observation, such that  $y_i^* = x_i \beta + \varepsilon_i$ , where  $x_i$  is a matrix with the first column equal to 1 and the second value is the independent variable we are using to predict the GAPS cut-off,  $\beta$  is a vector of parameters, and  $\varepsilon_i$  are independent and normally distributed errors with zero mean and constant variance. The latent variable is the value that we would expect to observe if there was no censoring.

The Tobit model takes the form:

$$y_i = y_i^* \text{ if } c_1 \leq y_i^* \leq c_2$$

$$y_i = c_1 \text{ if } c_1 > y_i^*$$

$$y_i = c_2 \text{ if } c_2 < y_i^*$$

Maximum likelihood estimators were computed using SAS version 9.1 (proc LIFEREG).

To determine the goodness of fit of the models, we used the pseudo- $R^2$  of McKelvey and Zavoina,<sup>48</sup> which was shown to be valid for the Tobit model.<sup>49</sup> A Monte Carlo simulation compared different pseudo- $R^2$  measures for the Tobit model and found this one to be the best,<sup>50</sup> with the main criterion being equivalence to the  $R^2$  measure that would be obtained using ordinary least squares regression if there was no censoring in the data.

**Validation of GAPS Cut-off Predictions Models**

To validate the GAPS cut-off values that we used, the delta score was computed for each of the three regions of Canada. This score indicates how far the uniqueness at the GAPS cut-off was from the asymptotic value. Small values of the delta score indicate that uniqueness is close to zero, and that

any additional geographic area aggregation would have an insignificant impact on uniqueness.

An end-user can enter either the Entropy or MaxCombs values in the Tobit models to predict the GAPS cut-off value for their study. To validate the accuracy of the prediction models, we used the Tobit models to predict the GAPS cut-off using 10-fold cross-validation.<sup>51,52</sup> That is, we divided the data sets into deciles and used one decile in turn for validation, and the remaining nine deciles to build the model.

The predicted cut-off used for validation was the unconditional value of the realized variable  $\hat{y}$ —the full equation for this estimate is provided in the literature.<sup>45-47</sup> Using  $\hat{y}$  in the validation ensured that the predicted value was also censored. The quality of the prediction was evaluated by considering the median and trimmed mean of the error  $(y - \hat{y})$  and the relative error, defined as  $(y - \hat{y})/y$ .

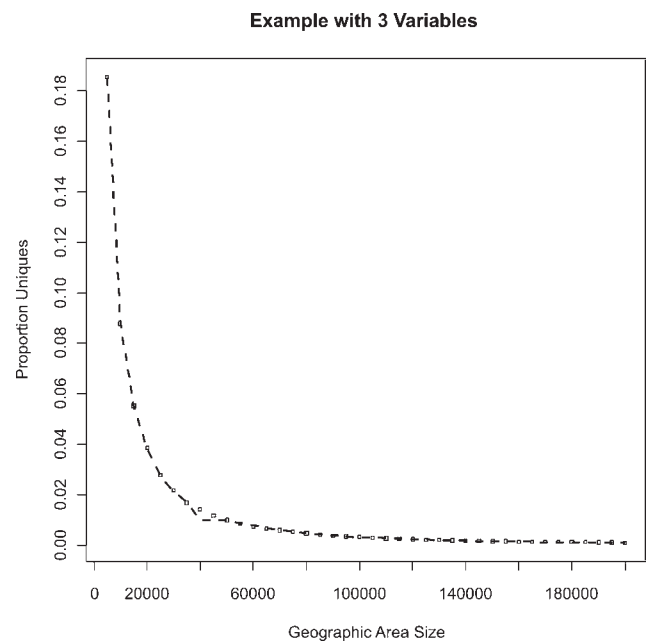
**Applying the Prediction Models**

Since an end-user does not need to worry about censoring (which is an artifact of our simulation), the predicted value of the latent variable would be used instead,  $\hat{y}^*$ . This is given by  $\hat{y}^* = e^{\beta_0} \text{Entropy}^{\beta_1}$  or  $\hat{y}^* = e^{\beta_0} \text{MaxCombs}^{\beta_1}$  where  $\beta_0$  and  $\beta_1$  are the model parameter estimates.

After presenting the results in the next section, the application of the prediction models in several real examples pertaining to the disclosure of retail and hospital pharmacy data to commercial data aggregators is illustrated in the discussion.

**Results**

An example of the relationship between GAPS and proportion uniqueness is shown in Fig 2. A similar pattern was observed for all regions and variable combinations. As illustrated in Fig 1, the cut-off was calculated from such a



**Figure 2.** Example showing the actual relationship between geographic area size and proportion uniques in the central region for the three variables: age, gender, and ethnicity.

**Table 2** ■ Table Showing the Delta Scores for the Three Regions. The Delta Score Represents the Proportion of Uniques at the Computed Geographical Area Population Size (GAPS) Cutoff Value. For Example, 0.0036 of the Individuals in Western Canada Were Unique at the GAPS Cutoff (median value)

	West	Central	East
Trimmed mean	0.007	0.0068	0.0061
Median	0.0036	0.0033	0.0037

graph by fitting a model and taking its derivative. The cut-off values were then used to develop the prediction models, as described in the previous section.

Table 2 shows the delta scores, which indicate how far uniqueness was from the asymptotic value at the various GAPS cut-offs that were calculated. As can be seen, there is very little difference in uniqueness across the regions, suggesting that there is little disclosure control benefit in increasing area sizes beyond the cut-offs that were calculated.

In Tables 3 and 4 we show the model parameters and diagnostics to predict the GAPS cut-off as a function of Entropy and MaxCombs, respectively. As is clear, all of the parameters are statistically significant, and the goodness of fit is high.

For both the Entropy and MaxCombs prediction models, the prediction errors are quite small. While the MaxCombs models have a slightly higher goodness-of-fit than the entropy models, the accuracy of the prediction for both are very similar.

## Discussion

The results suggest that the three regional models we have constructed for predicting the GAPS cut-off from both the Entropy and MaxCombs values can be quite accurate. They also make clear that having a single GAPS cut-off would be a serious oversimplification and that the appropriate cut-off

is a function of the quasi-identifiers that will be collected and the region of Canada.

Geographic areas that are larger than the GAPS cut-off represent low re-identification risk since they are close to the asymptotic risk value of zero, and there is also no disclosure control benefit in aggregating areas beyond the cut-off.

The prediction accuracy results were similar for MaxCombs and Entropy. One would expect Entropy to perform better given that it represents more information about the data distribution. However, there may be a ceiling effect in that the accuracy for either variable is sufficiently high that it is difficult for Entropy to outperform MaxCombs.

In practice, the MaxCombs value is easier to compute than the Entropy value. It is also possible to compute MaxCombs at the outset of a study during the design phase before any data are collected. We therefore recommend using the MaxCombs results in practice since in terms of accuracy they are very comparable to the Entropy results.

To apply these results an analyst first needs to compute the maximum number of combinations for the quasi-identifiers in the data set. Once this MaxCombs value is determined, the prediction models in Table 5 can be used to compute the GAPS cut-off. If the cut-off is deemed too large then the analyst can look at ways to reduce the value of MaxCombs by collapsing or coarsening the response categories. This process can be repeated until the cut-off is sensible for the particular study.

## Applying the Results

The following disclosure control example is about the re-identification of patients from their prescription records—it illustrates the application of our results. Many retail and hospital pharmacies across Canada provide prescription data to commercial data aggregators (we will refer to these data as “prescription records”). Prescription records are used to produce reports on physician prescription patterns

**Table 3** ■ Tobit Model Results for the Three Canadian Regional Models Using Entropy and Validation Accuracy Expressed in Terms of the Prediction Error and Relative Prediction Error

Entropy Prediction Model (Western)			
Pseudo- $R^2$		0.89	
Intercept		6.3; $p < 0.0001$	
Log (entropy) parameter est.		2.8; $p < 0.0001$	
Prediction error (10-fold)		Relative prediction error (10-fold)	
Trimmed mean	-4,433	Trimmed mean	0.012
Median	-1,500	Median	-0.02
Entropy Prediction Model (Central)			
Pseudo- $R^2$		0.8	
Intercept		6.5; $p < 0.0001$	
Log (entropy) parameter est.		2.6; $p < 0.0001$	
Prediction error (10-fold)		Relative prediction error (10-fold)	
Trimmed mean	-1,218	Trimmed mean	-0.015
Median	-7,405	Median	0.019
Entropy Prediction Model (Eastern)			
Pseudo- $R^2$		0.9	
Intercept		7.0; $p < 0.0001$	
Log (entropy) parameter est.		1.8; $p < 0.0001$	
Prediction error (10-fold)		Relative prediction error (10-fold)	
Trimmed mean	-1,284	Trimmed mean	0.0024
Median	-524	Median	-0.019

**Table 4** ■ Tobit Model Results Using MaxCombs for the Three Canadian Regions and Validation Accuracy Expressed in Terms of the Prediction Error and Relative Prediction Error

MaxCombs Prediction Model (Western)			
Pseudo- $R^2$		0.9	
Intercept		7.4; $p < 0.0001$	
Log (MaxCombs) parameter est.		0.4; $p < 0.0001$	
Prediction error (10-fold)		Relative prediction error (10-fold)	
Trimmed mean	-2,175	Trimmed mean	-0.012
Median	-1,325	Median	-0.016
MaxCombs Prediction Model (Central)			
Pseudo- $R^2$		0.9	
Intercept		7.3; $p < 0.0001$	
Log (MaxCombs) parameter est.		0.4; $p < 0.0001$	
Prediction error (10-fold)		Relative prediction error (10-fold)	
Trimmed mean	-2,472	Trimmed mean	-0.0002
Median	-1,156	Median	-0.013
MaxCombs Prediction Model (Eastern)			
Pseudo- $R^2$		0.9	
Intercept		7.6; $p < 0.0001$	
Log (MaxCombs) parameter est.		0.3; $p < 0.0001$	
Prediction error (10-fold)		Relative prediction error (10-fold)	
Trimmed mean	-920	Trimmed mean	-0.007
Median	-445	Median	-0.015

and drug use<sup>53</sup> These reports are then sold primarily to the pharmaceutical industry and government agencies.

In practice, the prescription records provided to commercial data aggregators do not contain directly identifying information about the patients (e.g., patient names and telephone numbers). However, it has been argued that the patient information that is disclosed in such records can still re-identify patients<sup>54,55</sup> and that this possible re-identification jeopardizes the confidentiality of Canadians' health information.<sup>54</sup>

The relevant quasi-identifiers in the prescription record are summarized in Table 6. We relied on five sources to construct this table: (1) the Canadian Pharmacists Association (CPhA) Pharmacy Claim Standard which defines all fields in the pharmacy electronic record used for claims adjudication,<sup>56</sup> (2) a report provided to us on the variables collected by the data management group at IMS Health Canada Inc, one of the largest commercial data aggregators in Canada,<sup>57</sup> (3) the investigation report by the Office of the Information and Privacy Commissioner of Alberta (OIPCA) which listed the 37 fields that are collected by commercial data aggregators,<sup>58</sup> (4) the results of a survey of provincial pharmacy regulatory authorities,<sup>54</sup> and (5) a specification of the data collected by Brogan Inc from Canadian hospital pharmacies (Brogan is another large commercial data aggregator in Canada).<sup>59</sup>

Key variables that are disclosed pertaining directly to patients are gender and year of birth.

**Table 5** ■ Prediction Models to Use for Determining the Smallest Region Size Using MaxCombs

Region of Canada	GAPS Cut-off
Western	$1588 \times \text{MaxCombs}^{0.42}$
Central	$1436 \times \text{MaxCombs}^{0.43}$
Eastern	$1978 \times \text{MaxCombs}^{0.304}$

GAPS = geographical area population size.

Brogan also collects the patient FSA, but IMS Health does not do so directly. However, it is often possible to infer new information about individuals from variables that already exist in a record:<sup>11</sup> it may be possible to infer the patient (residence) postal code from the postal code of their pharmacy or the prescriber if one assumes that there is some regularity in the distances that patients travel to see their general practitioner, specialist, or pharmacist. A simulation concluded that a patient would have to live at most within a 100-m radius from the pharmacy or prescriber to be able to accurately predict the full postal code in urban areas.<sup>11</sup> For rural areas, the distance varies from 1 km in Nova Scotia, 5 km in Ontario, to 10 km in Alberta.<sup>11</sup> We conducted a similar simulation to determine the accuracy of inferring the FSA and concluded that this can be accurately predicted if the patient lives within 10 km of the pharmacist/prescriber for rural areas, and within 1 km for urban areas in Nova Scotia and Alberta, and 0.5 km in Ontario.

In our analysis, we therefore made the assumption that the FSA was being collected or that it was reasonable to accurately infer the FSA for some of the patients if it is not collected.

#### Example 1

In this example, the prediction models were applied to assess patient re-identification risk for pharmacy prescription records in ten Canadian provinces, for the two quasi-identifiers of age and gender. The MaxCombs value is 172; the number of all possible values of age (86)  $\times$  gender categories (2). For each of the three regions of Canada the GAPS cut-off was computed using the values in Table 5. The percentage of FSAs whose population size is above the predicted cut-off for each province along with the percentage of the population that resides in these FSAs was then calculated.

The results are summarized in Table 7, and compared to the three other cut-offs that were being used: the 20,000 cut-off used in HIPAA (in practice the HIPAA Privacy Rule is sometimes used in Canada<sup>60</sup>), the Statistics Canada 70,000



**Table 6 ■ Fields That can be Used to Re-identify Patients in the Prescription Record According to Our Five Sources. For Hospital Pharmacies Other Data, Such as Dates for Admission and Discharge, are Collected. However, Here We Focus on the Variables That are Common Between Retail and Hospital Pharmacies**

Variable	CPhA Standard		IMS <sup>57</sup>	Field in OIPCA Report? <sup>58</sup>	Disclosed According to Survey? <sup>54</sup>	Brogan <sup>59</sup>	Additional Explanations
	Defined in CPhA Std? <sup>56</sup>	CPhA Mandatory? <sup>56</sup>					
Patient gender	Y	O	R	Y	Y**	Y	All sources indicate that patient gender is collected.
Patient year of birth	Y	O	R	Y	Y**	Y	The survey suggests that some provinces collected the full date of birth. <sup>54</sup> But both the OIPCA report <sup>58</sup> as well as the IMS Health Reports <sup>57</sup> indicate that only the year of birth is collected.
Patient postal code	Y	O	—	—	n***	Y†	The survey indicated that only PEI allowed the collection of postal codes. <sup>54</sup> When we contacted the pharmacy registrar in PEI it was made clear that if geographic information was disclosed by pharmacies, only the FSA was being disclosed rather than the full postal code. The IMS health report indicated that neither the full postal code nor FSA are collected from any province. <sup>57</sup> The Brogan document indicated that the FSA was being collected. <sup>59</sup>
Pharmacy postal code	Y	M	Y	Y	—	Y	Brogan's data are from hospital pharmacies, therefore the pharmacist is known.
Prescriber postal code	Y	O	Y*	Y	Y¶	Y	Prescriber group is in the record layout for the Brogan data.

M = Mandatory field in the CPhA claims standard; O = optional field. These fields will not necessarily be available for every pharmacy submitting data; CPhA = Canadian Pharmacists Association; SD = standard deviation; OIPCA = Office of the Information and Privacy Commissioner of Alberta; R = The field is required by IMS health Canada from all pharmacies submitting data, but if it is missing that would not invalidate the record. The field is not defined or collected at all.

\*whether this field is collected depends on the arrangement with a particular pharmacy and on the province (not collected in BC, MN, QC).

\*\*except MN, QC, NS.

\*\*\*except PEI.

¶except BC, SK, MN, Nfld.

†Brogan collects the patient FSA as part of its record layout.

cut-off for the CCHS, and the Census Bureau 100,000 cut-off. These data show that, except for New Brunswick, the vast majority of the provincial populations live in FSAs that are larger than the GAPS cut-off and therefore there is no disclosure control benefit in aggregating the geography any further.

For commercial data aggregators, there are three possible options:

1. Suppress the FSAs that are smaller than the cut-off. For example, in Ontario data from 31% (100–69%) of FSAs would need to be suppressed. These 31% of FSAs represent 9% of the Ontario population.

**Table 7 ■ The Percentage of FSAs and the Provincial Populations That Would be Above the GAPS Cut-off for an Age × Gender Quasi-identifier Combination for All Ten Canadian Provinces. These Counts are Based on the 2001 Census FSA Population Numbers Provided by Statistics Canada**

Province	Our GAPS Models		20,000 Cut-off		70,000 Cut-off		100,000 Cut-off	
	FSA	Pop	FSA	Pop	FSA	Pop	FSA	Pop
Alberta	55%	84%	38%	71%	1.4%	5%	0.00	0
British Columbia	68%	87%	46%	70%	1.1%	4%	0.00	0
Manitoba	59%	88%	39%	68%	0	0	0.00	0
New Brunswick	20%	51%	4.5%	19%	0	0	0.00	0
New found land	55%	83%	30%	62%	0	0	0.00	0
Nova Scotia	47%	82%	16%	43%	0	0	0.00	0
Ontario	69%	91%	49%	76%	1.4%	5%	0.20%	1%
PEI	57%	90%	43%	79%	0	0	0.00	0
Quebec	59%	84%	36%	63%	1%	5%	0.25%	0
Saskatchewan	60%	93%	49%	84%	2%	7%	0.00	2%

FSA = forward station area; GAPS = geographical area population size; PEI = Prince Edward Island.

2. Given that sex and gender are collected, determine what level of geographic aggregation would be suitable to avoid suppression of any data.
3. The analyst coarsens or collapses the response categories of the quasi-identifiers given that the level of geography is fixed at the FSA.

Suppression of data from small FSAs means that pharmacists would not be permitted to provide that data to the commercial data aggregators. Nevertheless, there would be far less FSA suppression using our models compared to the other cut-offs in current use: our models take into account the characteristics of the variables and calibrate the cut-off. For some provinces, no data would be released at all if some of the other GAPS cut-offs are applied.

For the second option described above, one can aggregate FSAs to the postal region, the first character of the postal code. We found that all postal regions in the ten provinces are above the GAPS cut-off. Therefore, inclusion of the sex and gender variables in the prescription record is possible as long as the geographic detail is at the postal region level, since this level of geography is always higher than the cut-off. The advantage of this option is that no data needs to be suppressed at all; however the disadvantage is that the aggregated geographic unit is quite large.

For the third option described above, it is assumed that the FSA geographic detail needs to be retained—the question then is which one of sex and gender is to be coarsened and the interval for grouping the coarsened age categories. For example, instead of disclosing the age in years, age can be disclosed as part of a 2-year interval, a 5-year interval, or a 10-year interval. The results for such coarsened categories are shown in Table 8. As expected the percentage of FSAs that can be disclosed increases as the amount of coarsening increases. However, for smaller provinces, such as New Brunswick, the proportion of the population in large FSAs remains low even with 10-years age intervals. Table 8 also shows the effect of coarsening the categories for age in terms of the percentage of the population. With 5-years age intervals, 98% of the Ontario population would be living in regions that are larger than the cut-off.

### Example 2

In this example we consider a specific data set from a hospital pharmacy. Records for all prescriptions dispensed from the Children's Hospital of Eastern Ontario pharmacy during the period beginning January 2007 to the end of June 2008 were obtained following institutional ethics approval. In total there were 94,100 records. These represent 10,259 patient visits and 6,902 unique patients.

The MaxCombs value for these data are 54 since the patient ages in years range from 0 to 26. Also, almost all of the patients of the hospital come from Ontario and Quebec. Therefore, we used the Central Canada model from Table 5.

The results were that 95% of the patients in the prescription record database reside in FSAs that are larger than the cut-off. However, for pediatric hospital patients it is important to know the age in weeks for patients younger than 1 year. Here, the MaxCombs value is 156, and the result is that 89% of the patients live in FSAs that are larger than the Central Canada cut-off.

### Summary

These examples show that using the MaxCombs prediction models given in Table 5 provide a straightforward technique to determine the GAPS cut-offs for datasets so the re-identification risk is managed while allowing for an increased amount of data to be available to the health researcher.

### Relationship to Other Work

There have been previous descriptive studies of uniqueness in the United States population on basic demographic variables, such as age and gender.<sup>61,62</sup> However, these studies did not explicitly consider the impact of nested geographic areas and their population size on uniqueness.

We used uniqueness as the measure for re-identification risk. Another common criterion for evaluating re-identification risk is k-anonymity.<sup>63,64</sup> This criterion considers that non-unique records are also risky. However, even under k-anonymity, unique records are still those with the highest probability of re-identification. Therefore, managing the risk of re-identification from uniques remains an important objective in disclosure control.

**Table 8** ■ The Percentage of FSAs and the Provincial Populations That Would be Above the GAPS Cut-off for an Age × Gender Quasi-identifier Combination for All Ten Canadian Provinces When the Age Variable is Coarsened to Different Sized Intervals

Province	Original Variables		2-yrs Age Intervals		5-yrs Age Intervals		10-yrs Age Intervals	
	FSA	Pop	FSA	Pop	FSA	Pop	FSA	Pop
Alberta	55%	84%	68%	92%	79%	96%	84%	98%
British Columbia	68%	87%	78%	93%	90%	99%	93%	99%
Manitoba	59%	88%	66%	92%	72%	95%	78%	98%
New Brunswick	20%	51%	26%	59%	37%	70%	45%	75%
Newfoundland	55%	83%	70%	91%	79%	95%	88%	98%
Nova Scotia	47%	82%	54%	86%	66%	93%	72%	95%
Ontario	69%	91%	78%	96%	84%	98%	87%	99%
PEI	57%	90%	71%	97%	71%	97%	71%	97%
Quebec	59%	84%	70%	91%	82%	96%	88%	99%
Saskatchewan	60%	93%	69%	97%	69%	97%	71%	98%

FSA = forward station area; GAPS = geographical area population size; PEI = Prince Edward Island.

Earlier work at the United States Census Bureau evaluated nested geographic areas, and provided the basic methodology for our study.<sup>29,31</sup> This work did not document a general model that can be applied by individuals outside the bureau and that takes into account the characteristics of the quasi-identifiers, which is what we did in this study.

### Limitations

The prediction models we present here should be considered as one element in a toolbox of heuristics that can be used for disclosure control. Some other heuristics have been described in previous work.<sup>65,66</sup>

Although we contend that the ten quasi-identifiers we considered represent basic demographics that are quite common in health research, they will not cover all possible quasi-identifiers that may be used in practice. Thus, our results are limited to the specific variables that we have considered in our analysis.

### Conclusions

Data custodians often use general population size cut-offs to determine the level of geographic information to disclose in a data set. For example, the HIPAA Privacy Rule's Safe Harbor list allows the release of the first three digits of the ZIP code only if that area has 20,000 or more individuals living in it. National statistical agencies in the United States, UK, and Canada also use such cut-offs as part of their disclosure control practices. The primary rationale for such cut-offs is that there is no disclosure control benefit for aggregating geographic areas beyond that size.

In this paper we performed an empirical evaluation of such cut-offs using Canadian census data. Our results indicate that the appropriate cut-off depends on characteristics of the variables included in the data set; therefore there is not a single cut-off. We developed and validated models to predict such population size cut-offs for Canada. The model which predicted population cut-offs using the maximum number of possible values in the data set had  $R^2$  values approaching 0.9, and relative error of prediction less than 0.02 across all regions of Canada. Our prediction models were then applied in a risk assessment of the prescription records that are provided by Canadian pharmacies to commercial data aggregators. This assessment indicated that for most of the Canadian population, that there is no disclosure control benefit to aggregating geography beyond the FSA when releasing patient age and gender.

### References ■

- Platt P, Hendlisz L, Intrator D. Privacy Law in the Private Sector: An Annotation of the Legislation in Canada, Canada Law Book, 2004.
- Willison D, Emerson C, Szala-Meneok K, et al. Access to medical records for research purposes: Varying perceptions across Research Ethics Boards. *J Med Ethics* 2008;34:308–14.
- Woolf S, Rothemich S, JR, Marsland D. Selection bias from requiring patients to give consent to examine data for health services research. *Arch Fam Med* 2000;9:1111–8.
- Junghans C, Feder G, Hemingway H, Timmis A, Jones M. Recruiting patients to medical research: Double blind randomised trial of 'opt-in' versus 'opt-out' strategies. *Br Med J* 2005 Oct 22;331(7522):940. Epub 2005 Sep 12.
- Jacobsen S, Xia Z, Campion M, et al. Potential effect of authorization bias on medical records research. *Mayo Clin Proc* 1999;74(4):330–8.
- Nelson K, Rosa E, Brown J, et al. Do patient consent procedures affect participation rates in health services research? *Med Care* 2002;40(4):283–8.
- McKinney P, Jones S, Parslow R, et al. A feasibility study of signed consent for the collection of patient identifiable information for a national paediatric clinical audit database. *Br Med J* 2005 Apr 16;330(7496):877–9. Epub 2005 Mar 18.
- Tu J, Willison D, Silver F, et al. Impracticability of informed consent in the Registry of the Canadian Stroke Network. *N Engl J Med* 2004;350(14):1414–21.
- Armstrong D, Kline-Rogers E, Jani S, et al. Eagle. Potential impact of the HIPAA privacy rule on data collection in a registry of patients with acute coronary syndrome. *Arch Intern Med* 2005;165:1125–9.
- Al-Shahi R, Vousden C, Warlow C. Bias from requiring explicit consent from all participants in observational research: Prospective, population based study. *Br Med J* 2005;331:942.
- El Emam K, Jonker E, Sams S, et al. Pan-Canadian de-identification guidelines for Personal health information Available at: <http://www.ehealthinformation.ca/documents/OPCReportv11.pdf>. Archived at: <http://www.webcitation.org/5Ow1Nko5C>. Accessed: May 18, 2007.
- Boulos M. Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. *Int J Health Geographics* 2004;3(1).
- O'Dwyer LA, Burton DL. Potential meets reality: GIS and Public health research in Australia. *Aust N Z J Pub Health* 1998;22(7):819–23.
- Ricketts TC. Geographic information systems and public health. *Annu Review Pub Health* 2003;24:1–6.
- Cromley EK. GIS and disease. *Annu Review Pub Health* 2003;24:7–24.
- Brindley P, Maheswaran R. My favourite software: Geographic information systems. *J Pub Health Med* 2002;24(2):149.
- Richards TB, Croner CM, Rushton G, Brown CK, Fowler L. Geographic information systems and public health: Mapping the future. *Pub Health Rep* 1999;114:359–73.
- Mugge R. Issues in protecting confidentiality in national health statistics. In: Proceedings of the Social Statistics Section, American, Statistical Association, 1983.
- Mackie C, Bradburn N. Improving Access to and Confidentiality of Research Data: Report of a Workshop, National Academies, 2000.
- Justice Gibson. Mr. Mike Gordon and the Minister of Health and Privacy Commissioner of Canada, Federal Court of Canada, February 27 2008.
- Brownstein J, Cassa C, Mandl K. No place to hide—Reverse identification of patients from published maps. *N Engl J Med* 2006;355(16):1741–2.
- Brownstein J, Cassa C, Kohane I, Mandl K. An unsupervised classification method for inferring original case locations from low-resolution disease maps. *Int J Health Geographics* 2006;5(56).
- Curtis A, Mills J, Leitner M. Spatial confidentiality and GIS: Re-Engineering mortality locations from published maps about Hurricane Katrina. *Int J Health Geographics* 2006;5(44).
- Armstrong M, Rushton G, Zimmerman D. Geographically masking health data to preserve confidentiality. *Stat Med* 1999;18:497–525.
- Zimmerman D, Pavlik C. Quantifying the effects of masking metadata disclosure and multiple releases on the confidentiality of geographically masked health data. *Geogr Anal* 2008;40:52–76.

26. Fefferman N, O'Neil E, Naumova E. Confidentiality and confidence: Is data aggregation a means to achieve Both? *J Pub Health Pol* 2005;16:430–49.
27. Willenborg L, Mokken R, Pannekoek J. Microdata and disclosure risks. In: . Proceedings of the Annual Research Conference of United States Bureau of the Census, 1990.
28. Olson K, Grannis S, Mandl K. Privacy protection *versus* cluster detection in spatial epidemiology. *Am J Pub Health* 2006;96(11): 2002–8.
29. Hawala S. Enhancing the “100,000” rule: on the variation of percent of uniques in a microdata sample and the geographic area size identified on the file. In: Proceedings of the Annual Meeting of the American, Statistical Association, 2001.
30. Greenberg B, Voshell L. Relating risk of disclosure for microdata and geographic area size. In: Proceedings of the Section on Survey Research Methods, American Statistical Association, 1990.
31. Greenberg B, Voshell L. The geographic component of disclosure risk for microdata. Bureau of The Census. 1990.
32. Zayatz L, Massell P, Steel P. Disclosure limitation practices and research at the US Census Bureau. *Neth Off Statistics* 1999; 14(Spring):26–9.
33. Zayatz L. Disclosure Avoidance Practices and Research at the US Census Bureau: An Update, United States Census Bureau, 2005.
34. Hawala S. Microdata disclosure protection research and experiences at the US Census Bureau. Proceedings of the Workshop on Microdata, 2003 Available at: <http://www.census.gov/srd/sdc/microdataprotection.pdf>. Archived at: <http://www.webcitation.org/5b7mPeVPi>. Accessed: September 26, 2008.
35. Rudolph B, Shah G, Love D. Small numbers, disclosure risk, security, and reliability issues in web-based data query systems. *J Pub Health Manag Practice* 2006;12(2):176–83.
36. Stoto M. Statistical issues in interactive web-based public health data dissemination systems, 2003, RAND Health.
37. Marsh C, Dale A, Skinner C. Safe data *versus* safe settings: Access to microdata from the British census. *Int Stat Review* 1994;62(1):35–53.
38. Statistics Canada, Canadian Community. Health survey (CCHS). Cycles 2005;3:1 Public Use Microdata File (PUMF) User Guide. 2006.
39. Willenborg L, de Waal T. Statistical Disclosure Control in Practice, Springer-Verlag, 1996.
40. Standards for Privacy of Individually Identifiable Health Information, in Federal Register, Dec 28, 2000 (Volume 65, Number 250). 2000. p. 82,511–82,560.
41. El Emam K. Overview of factors affecting the risk of Re-identification in Canada, 2006, Access to Information and Privacy Division, Health Canada.
42. Dalenius T. Finding a needle in a haystack or identifying anonymous census records. *J Off Stat* 1986;2(3):329–36.
43. Bethlehem J, Keller W, Pannekoek J. Disclosure control of microdata. *J Am Stat Assoc* 1990;85(409):38–45.
44. Statistics Canada. Census public use Microdata file: Individuals file user Documentation. 2001.
45. Long S. Regression Models for Categorical and Limited Dependent Variables, Sage Publications, 1997.
46. Breen R. Regression Models for Censored, Sample-Selected, and Truncated Data, Sage Publications, 1996.
47. Maddala G. Limited-Dependent and Qualitative Variables in Econometrics, Cambridge University Press, 1983.
48. McKelvey R, Zavoina W. A statistical model for the analysis of ordinal level dependent variables. *J Math Sociol* 1975;4:103–20.
49. Laitila T. A pseudo-R<sup>2</sup> measure for limited and qualitative dependent variable models. *J Econ* 1993;56:341–56.
50. Veall M, Zimmermann K. Goodness of fit measures in the Tobit model. *Oxf Bull Econ Stat* 1994;56(4):485–99.
51. Cherkassky V, Muller F. Learning from Data, Wiley, 1998.
52. Alpaydin E. Introduction to Machine Learning, MIT Press, 2004.
53. Kallukaran P, Kagan J. Data mining at IMS health: How we turned a mountain of data into a few information-rich molehills. In: . Proceedings of the 24<sup>th</sup> Annual SAS Users Group International Conference, 1999.
54. Zoutman D, Ford B, Bassili A. The confidentiality of patient and physician information on pharmacy prescription records. *CMAJ* 2004;170(5):815–6.
55. Zoutman D, Ford B, Bassili A. Privacy of pharmacy prescription records (author response). *CMAJ* 2004;171(7):712.
56. Canadian Pharmaceutical Association. Pharmacy claim standard (version 03), 2006.
57. Fineberg A. Information requested for “Re-identification Study”, 2006, IMS Health Canada.
58. Office of the Information and Privacy Commissioner Order of Alberta. Order H2002-003: Alberta Pharmacies and Pharmacists. 2003.
59. Brogan, Inc. MedMap Drug Utilization Program, Program Overview, 2008.
60. El Emam K. Data Anonymization practices in clinical research: A descriptive Study. Health Canada, Access to Inf and Privacy Division. 2006.
61. Sweeney L. Uniqueness of simple demographics in the US population. Carnegie Mellon University, Lab for Int Data Privacy. 2000.
62. Golle P. Revisiting the uniqueness of simple demographics in the US population. In: Workshop on Privacy in the Electronic Society, 2006.
63. Samarati P. Protecting respondents' identities in microdata release. *IEEE Trans Knowl Data Eng* 2001;13(6):1010–27.
64. Sweeney L. k-anonymity: A model for protecting privacy. *Int J Uncertainty, Fuzziness and Knowl-Based Syst* 2002;10(5): 557–70.
65. El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. Evaluating common de-identification heuristics for personal health information. *J Med Internet Res* 2006;8(4):e28.
66. El Emam K. Heuristics for de-identifying health data. *IEEE Sec Privacy* 2008:72–5.

Research article

Open Access

## The perceived impact of location privacy: A web-based survey of public health perspectives and requirements in the UK and Canada

Philip AbdelMalik\*<sup>1,2</sup>, Maged N Kamel Boulos<sup>1</sup> and Ray Jones<sup>1</sup>

Address: <sup>1</sup>Faculty of Health and Social Work, University of Plymouth, Centre Court, 73 Exeter Street, Drake Circus, Plymouth, Devon PL4 8AA, UK and <sup>2</sup>Office of Public Health Practice, Public Health Agency of Canada, 120 Colonnade Road, AL6702A, Ottawa, Ontario, K1A 0K9, Canada

Email: Philip AbdelMalik\* - philip\_abdelmalik@phac-aspc.gc.ca; Maged N Kamel Boulos - maged.kamelboulos@plymouth.ac.uk; Ray Jones - ray.jones@plymouth.ac.uk

\* Corresponding author

Published: 9 May 2008

Received: 18 December 2007

BMC Public Health 2008, 8:156 doi:10.1186/1471-2458-8-156

Accepted: 9 May 2008

This article is available from: <http://www.biomedcentral.com/1471-2458/8/156>

© 2008 AbdelMalik et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** The "place-consciousness" of public health professionals is on the rise as spatial analyses and Geographic Information Systems (GIS) are rapidly becoming key components of their toolbox. However, "place" is most useful at its most precise, granular scale – which increases identification risks, thereby clashing with privacy issues. This paper describes the views and requirements of public health professionals in Canada and the UK on privacy issues and spatial data, as collected through a web-based survey.

**Methods:** Perceptions on the impact of privacy were collected through a web-based survey administered between November 2006 and January 2007. The survey targeted government, non-government and academic GIS labs and research groups involved in public health, as well as public health units (Canada), ministries, and observatories (UK). Potential participants were invited to participate through personally addressed, standardised emails.

**Results:** Of 112 invitees in Canada and 75 in the UK, 66 and 28 participated in the survey, respectively. The completion proportion for Canada was 91%, and 86% for the UK. No response differences were observed between the two countries. Ninety three percent of participants indicated a requirement for personally identifiable data (PID) in their public health activities, including geographic information. Privacy was identified as an obstacle to public health practice by 71% of respondents. The overall self-rated median score for knowledge of privacy legislation and policies was 7 out of 10. Those who rated their knowledge of privacy as high (at the median or above) also rated it significantly more severe as an obstacle to research ( $P < 0.001$ ). The most critical cause cited by participants in both countries was bureaucracy.

**Conclusion:** The clash between PID requirements – including granular geography – and limitations imposed by privacy and its associated bureaucracy require immediate attention and solutions, particularly given the increasing utilisation of GIS in public health. Solutions include harmonization of privacy legislation with public health requirements, bureaucratic simplification, increased multidisciplinary discourse, education, and development of toolsets, algorithms and guidelines for using and reporting on disaggregate data.

## Background

Although "place" has been coined one of the three pillars of epidemiological data, only relatively recently has it garnered significant attention in the public health field, as Geographic Information Systems (GIS) have increasingly become more affordable, accessible, and intuitive. Indeed, the public health community's "place-consciousness" is on the rise as spatial analyses and GIS, now defined as part of the medical and health literature [1-3], are rapidly becoming key components of the public health professional's toolbox [4].

Privacy, an evolving "principle as old as the common law" [5], has been cited as an issue in a variety of public health events, reports, and media releases [6-11]. So much so, in fact, that one sometimes cannot help but wonder if privacy is, indeed, the enemy of public health [12], and whether they could ever peacefully co-exist [13]. A distinction should here be made between the related concepts of *privacy*, *confidentiality*, and *security* within the context of the current discussion. *Privacy* is attributable to the individual about whom identifiable information pertains, and refers to that individual's right to control such information, thereby freeing the individual from un-invited intrusion and identification. *Confidentiality* obligates others who have been entrusted with such information to respect the individual's privacy, and is therefore attributable to third parties; a breach of confidentiality violates the privacy of the individual because the individual has had no control over the release of the data. Finally, *security* refers to tools and methods used to safeguard confidentiality and privacy [14,15]. This research deals specifically with privacy issues as regulated and defined by legislation and ethical guidelines surrounding consent. From within this context, an individual's privacy is not deemed to have been violated if data shared in the absence of consent cannot be used to identify the individual. Exception clauses generally exist in legislation, allowing authorities to release personally identifiable data under various circumstances – such as where it is deemed to be in the best interest of society or where it is impractical to obtain consent. Examples include Section 60 of the UK's *Health and Social Care Act 2001* [16], and Sections 8 and 7 of Canada's *Privacy Act* [17] and *Personal Information Protection and Electronic Documents Act* [18], respectively. While an analysis of privacy legislation as it pertains to health data and the concept of "place" is beyond the scope of this paper, suffice it to say that such clauses are often ambiguous and subjective, particularly when combined with vague definitions of "sensitive personal information" and the scale at which geographic data becomes "identifiable". The concept of *place*, for example, is not explicitly specified as "sensitive personal data" in the UK's *Data Protection Act 1988* [19], nor in the generic *EU Data Protection Directive* of 1995 [20] (though it is explicitly mentioned in var-

ious telecommunications directives), but postcodes are specifically mentioned in a 2005 NHS data protection and medical research POSTnote [21]. In Canada's *Privacy Act* [17], "address" is specifically listed as "personal information", while in the *Personal Information Protection and Electronic Documents Act* [18], it is not (though implied). Such ambiguities deter the sharing of data, causing organisations and authorities to err on the side of caution and not release identifying information [22], including spatial data.

It is no surprise, therefore, that the increasing popularity of "place" in public health has further exacerbated the public health research-privacy debate. Traditional health-data anonymisation techniques, such as pseudonymisation and aggregation, cannot be applied to spatial data without significantly altering or destroying the spatial relationships under investigation [23-26], and hence the very reason for which they are to be used in the first place. The problem with "place" is that it is most useful at its most precise, granular scale [15,23]. Yet with increasing spatial precision and accuracy comes a corresponding increase in the risk of identification, and therefore a breach of privacy [15]. This becomes particularly troublesome when the spatial data is linked to health, social or demographic data. The development of methods by which to mitigate these risks continues to be an active area of research, but thus far, proposed solutions have limitations, risks and tradeoffs, and lack guidelines on their appropriate use. Consequently, the acquisition of geographic data tends to be either limited, or at a sub-optimal or unusable scale. Not only do privacy issues impact data acquisition and use for analysis, but also visualisation and dissemination of the results. Researchers have been able to "reverse engineer" maps, for example, to successfully re-identify individuals [27-29].

While the debate between the fields of privacy and public health has raged on for decades [5] despite their interdependence on one another [14], tension continues to rise in concert with the rampant growth of information technology and e-Health. From a health research perspective, both Canada and the UK place strong emphasis on evidence-based public health policies and services [6], yet in both countries, this seems to be hampered by privacy issues. While some argue that this debate is the product of a lack of understanding of the legislation and regulations by the public health community [14,30,31], there is little in the way of formal collection and synthesis of the corresponding views and perspectives of those directly involved in public health activities. This paper describes the views and requirements of public health professionals in Canada and the UK on privacy issues and spatial data, as collected through a web-based survey. Given that Canada's health care and public health systems were both

largely modeled after those of the UK [6,32,33], that each continues to be studied by the other for improvements and lessons learned [6,34], and that privacy issues for public health have been cited in both, it is expected that survey responses in the two countries will also be similar.

**Methods**

**Development & Content**

The survey was first developed on paper in the summer of 2006, and piloted with select public health individuals in Canada and the UK. It was then submitted for privacy assessment by the Access to Information and Privacy Branch of Health Canada, and for ethics review and approval from the Health Canada Research Ethics Board and the Southwest Multicentre Research Ethics Committee in the UK. Throughout the process it was clear that the survey would be developed as a closed web-based survey, running between November 2006 and January 2007. The final paper versions of the survey are provided (see Additional files 1, 2, 3) and can also be found on the research website [35].

The paper survey was then converted to a web-version by the ALPHA Project [36] team at the Public Health Agency of Canada (PHAC), and piloted by the author and several colleagues within the PHAC. The survey launch was delayed by two weeks, with only some of the concerns identified during the pilot being implemented due to limitations of the ALPHA architecture. Issues and limitations with the design of the web-based survey are addressed in a later section.

Three versions of the survey were developed and launched: Canada-English, Canada-French and UK-Eng-

lish. A summary of the survey's structure and contents is given in Table 1.

**Target**

The survey targeted government, non-government and academic GIS labs and research groups involved in public health, as well as public health units (Canada), ministries, and observatories (UK). Potential participants were identified through web searches of public health sites, mailing databases, personal contact, referrals/word of mouth, and postings on the research website [35], a PHAC Public Health Portal website [37], and the NHS Public Health Informatics Community website [38].

**Participation**

Potential participants were invited to participate through a standardised but personally addressed email outlining the reason for the invite, the mechanisms by which their contact information was retrieved, a brief summary of the research and survey, a description of the data handling methods, an estimate of the time it would take to complete the survey (approximately 20 minutes), a unique user ID and password, the URL to the survey site, the URL to the research website, and the principle investigator's contact information.

The survey website had no other content. In order to participate, invitees were required to (1) successfully log in, and (2) consent to participation. Only the most recent responses for any given user ID were collected, ensuring only one survey was completed per participant. The consent screen outlined the voluntary and anonymous nature of the survey, indicated the approximate time it would take to complete the survey, the risks and benefits to the

**Table 1: Sections of the survey**

Section	Title	Description
I	A little about you...	Participant scope, roles, use of GIS, etc
II	Current access to data	Asks participants with current access to PID to score 15 kinds of PID* on various dimensions, such as ease and frequency of access, usefulness and importance, etc.
III	No current access to data	Asks participants without current access to PID to score same as above
IV	Privacy issues	Collects participant opinions on the overall impact of restricted access to PID on public health practice (research, surveillance, health service delivery, etc)
V	Current data holdings and provision to others...	Collects information on the sharing of PID within and between participant organisations
VI	Solutions and research	Presents two distinct solutions to overcome barriers posed by privacy to public health research, and gather participant views on usefulness, usability and preference for each
VII	Qualitative component	Allows participants to provide views and opinions on knowledge of privacy and confidentiality issues/legislation, impact of privacy, proposed research and solutions, and additional thoughts or comments
VIII	Further participation and contact	Allows participants to provide contact information if they choose, for follow-up, updates, or piloting of potential solution(s)

\* For all participants: first name; last name; initials; sex; date of birth; date of death; registered GP or family physician; street address; postal code; community name; city/town/village; region/geographic area; latitude/longitude.  
 For Canadian participants: provincial health insurance plan number; hospital ID.  
 For UK participants: old NHS number; new NHS number

participants, the intellectual property and ownership of all data collected, and the protection of any personal data provided under Canadian and UK law. Failure to successfully complete either of these two requirements resulted in termination of the survey. After consenting, participants were given the option to select their country and language of choice, and the relevant survey then commenced.

All questions included a "Skip" option. Progress through the survey required the selection of a response for each question, and participants could terminate the survey at any time or complete it over multiple sessions, at their convenience. Questions were not randomized or alternated, but adaptive questioning was utilized. Question types varied, and included single-choice, multiple-choice, scale, and free-form response questions, thereby collecting both quantitative and qualitative responses. There was typically only one question per screen with multiple potential responses, the maximum number of which was 17. Depending on the responses of the participants, the survey was distributed over approximately 40 screens.

Key questions addressed by the survey included the following:

- Is there a requirement for personally identifiable data, including spatial data?
- What spatial resolution is ideal for public health research?
- Is privacy perceived to be a significant obstacle to public health practice?
- How knowledgeable do public health professionals consider themselves on privacy?
- What is the most critical obstacle to the access and use of personally identifiable data?
- What are the views of the public health community on public awareness and perceptions?
- Which is preferred: raw, case-level data, or aggregated, anonymised data?

Collected responses were analysed using basic descriptive statistics and non-parametric methods in SAS 9.2. The Checklist for Reporting Results of Internet E-Surveys (CHERRIES) [39] was used as a guideline in the reporting of the web-based survey methodology.

## Results

Of 112 invitees in Canada and 75 in the UK, 66 (59%) and 28 (37%) participated in the survey, respectively. Of the Canadian participants, three responded to the French version. The completion proportion for Canada was 91%, and 86% for the UK.

There were no differences in the distribution of roles reported by participants in both countries, with most participants (49% in Canada; 64% in the UK) identifying their main role as falling within the research and analysis domain (Table 2). Participant expertise varied, and included aboriginal health (Canada only), chronic diseases, paediatric public health, infectious diseases, dental public health, emergency preparedness and response, environmental public health, ethics and public health law, food and nutrition, health services, injuries and disabilities, mental health and substance misuse, social determinants of health, surveillance, and education.

No response differences were observed between the two countries on each of the key questions, and the overall, combined results are therefore reported. A summary of the findings is given in Table 3.

### ***Is there a requirement for personally identifiable data, including spatial data?***

Almost all participants identified a need for personally identifiable data (PID) in their roles; only one Canadian participant indicated no need for PID. Five Canadian participants and one UK participant chose not to answer the question. In total 93% of participants indicated a requirement for PID in their public health activities.

### ***What spatial resolution is ideal for public health research?***

All participants identified geographic location of health data as a requirement for their roles or organisation. When asked "...what level of geography would you ideally like to visualise your data and/or conduct spatial analyses," 69% of respondents identified "latitude and longitude, exact street address, or exact household."

### ***Is privacy perceived to be a significant obstacle to public health practice? AND How knowledgeable do public health professionals consider themselves on privacy?***

When asked "Are you or have you been restricted in your use of GIS for any public health activity because of privacy concerns (i.e. map or data might identify an individual or community)?" 79% of respondents marked "YES".

Of 83 participants who responded to the question "In your opinion, do current restrictions to PID pose an obstacle to any aspects of public health practice?" 59 (71%) agreed, rating the obstacle severity at 6 or higher. Of these 59, 36 (61%) rated their knowledge of privacy



**Table 2: Number and percent of survey participants by main role and geographical scope**

Scope	Main Role					
	Strategic decision/ policy maker	Manager/ Coordinator	Consultant	Research & Analysis	Front-Line Responder/ Patient Care/Clinical	Other
<b>Canadian Participants</b>						
North American or National	3 (4.5%)	6 (9%)	-	9 (13.6%)	1 (1.5%)	2 (3.0%)
Provincial/Territorial	1 (1.5%)	3 (4.5%)	4 (6.1%)	6 (9.1%)	-	2 (3.0%)
Local/Regional	2 (3.0%)	7 (10.6%)	1 (1.5%)	17 (25.8%)	1 (1.5%)	1 (1.5%)
<b>Totals</b>	<b>6 (9.1%)</b>	<b>16 (24.2%)</b>	<b>5 (7.6%)</b>	<b>32 (48.5%)</b>	<b>2 (3.0%)</b>	<b>5 (7.6%)</b>
<b>UK Participants</b>						
European or National	1 (3.6%)	1 (3.6%)	-	1 (3.6%)	-	-
Regional	2 (7.1%)	1 (3.6%)	2 (7.1%)	12 (42.9%)	-	-
Local	2 (7.1%)	-	-	4 (14.3%)	-	1 (3.6%)
<b>Totals</b>	<b>5 (17.9%)</b>	<b>2 (7.1%)</b>	<b>2 (7.1%)</b>	<b>17 (60.7%)</b>	<b>0 (0.0%)</b>	<b>1 (3.6%)</b>

\*One UK participant who identified a main role in research and analysis declined a response to the question on scope.

and confidentiality issues/legislation at 6 out of 10 or higher, with a mean score of 7.5 (std = 1.0) and a median score of 7.

Using the median, respondents with a self-rated knowledge score lower than 7 were classified as "low" on knowledge (47%), while those at or above the median score were classified as "high" (53%). Those classified as high were more likely to rate privacy as an obstacle (one-sided Wilcoxon exact  $P < 0.001$ ). A trend was evident for the overall correlation between restriction score and self-rated privacy knowledge score (Spearman  $r = 0.22$ ,  $P = 0.057$ ).

**What is the most critical obstacle to the access and use of personally identifiable data?**

The most common obstacles were reported as bureaucracy and legislation by 33% and 25% of the participants, respectively. Other responses included public disapproval/paranoia (15%), practitioner paranoia (7%), lack of knowledge (6%), combination of these factors (4%), other (2%), and none (skipped question, 7%).

**What are the views of the public health community on public awareness and perceptions?**

Fifty seven percent of participants felt that under 10% of the public population is aware of the impact of restricted access to PID on public health practice; 74% felt it to be under 20%, and 84% felt the proportion to be less than 30% (cumulative frequencies). Most identified education

**Table 3: Summary of findings**

Question	Response Summary <sup>†</sup>
1. Is there a requirement for personally identifiable data?	Yes (93%)
2. What spatial resolution is ideal for public health research?	Lat/Long or address (69%)
3. Is privacy perceived to be a significant obstacle to public health practice?	Yes (71%)
4. How knowledgeable do public health professionals consider themselves on privacy?	High Knowledge* (53%)
5. What is the most critical obstacle to the access and use of personally identifiable data?	Bureaucracy (33%) Legislation (25%)
6. What are the views of the public health community on public awareness and perceptions?	Less than 30% of the public is aware (84%)
7. Which is preferred: raw, case level data, or aggregated, anonymised data?	Raw, case-level data (66%)

<sup>†</sup>Numbers in parentheses are the percent of participants who responded as described

\*Participants rating their knowledge as high were also more likely to rate privacy as a more severe obstacle ( $P < 0.001$ )

and awareness (through media, reports, case studies, scenarios, etc) as the best methods to increase this proportion. When then asked what proportion of the public they felt would allow the use of their PID if they were educated on the usefulness of such data to public health practice, 67% said 50% or higher.

**Which is preferred: raw, case-level data, or aggregated, anonymised data?**

More respondents identified a preference for having access to granular-level rather than aggregate data (53 vs. 27; 66% of those responding to this question).

**Discussion**

This survey and user-needs assessment on privacy and public health shows a definite requirement by public health professionals – in various fields and positions in both Canada and the UK – for personally identifiable data, including spatial data. The requirement for this spatial data is at its most granular level – latitude and longitude, or exact street address – which necessarily compromises patient privacy. It is not surprising, therefore, that public health professionals perceive privacy to be a significant obstacle to public health practice.

There are those who would argue that this perception is the product of a lack of understanding of the legislation and regulations by the public health community. The results of this research, however, indicate the contrary. Not only did public health professionals in both countries generally rate themselves high on knowledge of privacy legislation and related issues, but those with the highest self-rated scores also tended to rate privacy as more of an obstacle. That these self-ratings of knowledge are not representative of actual knowledge remains possible.

Participants perceived the most critical obstacles to sharing or acquisition of health data with PID to be bureaucracy, followed by legislation.

Bureaucracy surrounding health research in both Canada and the UK generally revolves around data ownership, academic competitiveness, ethics review boards or committees, and in particular, requirements for informed consent, even if they compromise public health, or are not in the best interests of the patients involved [40-42]. Since seeking subject consent with every new hypothesis to be tested or model to be developed is an impossible task, some have suggested that thought be given to "blanket" consent. At the Canadian Institutes for Health Research (CIHR) 2003 workshop on the legal and ethical issues facing the Canadian Lifelong Health Initiative [43], participants spent some time discussing such issues, only to emphasise the importance of the establishment of ethical governance and structure; essentially, more necessary

bureaucracy. Interestingly, while the debate continues, a relatively recent survey found that most of the British public did not consider the use of their National Cancer Registry PID for public health research and surveillance to be an invasion of their privacy [30]. While the ethics of blanket consent are not discussed in this study, it is nonetheless offered as a potential solution in light of the requirements of the public health community. This does not, however, address other issues of data ownership and control that contribute to the bureaucratic debate.

While many individuals recognised the importance of privacy legislation, participants generally indicated a concern and, in some cases, first-hand frustration that legislation unduly restricts public health activities, compromising surveillance and research. Many phrases were used by respondents to describe the implications of privacy legislation on public health, including, among others: "increasingly restrictive;" "serious;" "incomplete;" "fuzzy;" "does more harm than good;" "two-edged sword;" "causes challenges;" "delays and restricts access [to data];" "[is a] hindrance to the improvement and efficiency of public health;" "disappointing;" "frustrating;" "difficult to interpret;" "very worrisome;" "disadvantages the public interest;" "not properly understood;" "over-protective;" "limiting;" "hinders knowledge;" and "used as an excuse not to share data." A large proportion of the public health community represented in this sample clearly expressed major concerns with the impact of privacy legislation on their work – both in Canada, and in the UK – in spite of having a good understanding and acceptance of its purpose and necessity. It is also important for legislation to be written in an unambiguous manner that is clearly understood by both public health professionals and the general public [4].

Public health professionals are largely of the opinion that the general public's level of awareness of the impact of restricted access to PID on public health practice is extremely low. Surveys by the Office of the Privacy Commissioner in Canada [44] repeatedly show that the majority of Canadians surveyed (up to 80%) place an extremely high level of importance on strong laws to protect personal information, particularly health information, and that they feel that the level of protection of their personal information has declined over the past ten years. Yet interestingly, only 20% are clearly aware of existing laws, and even fewer (12%) are aware of their rights around the collection, use and disclosure of this information. The "need to raise Canadians' awareness about the current laws in place and what their rights are" [44] must therefore be coupled with the corresponding need to address this from within the context of public health requirements.

Educating the public, therefore, as well as practitioners, data users, policy makers and politicians, was not surprisingly identified by participants as a potential solution. Participants put emphasis on the utilisation of the media to educate and increase awareness, as well as demonstrating the impact of a lack of data, and the benefits of its use when available. Demonstration of the benefits to the individual (e.g. streamlining of the system, not being asked for personal information with every visit to a new clinician, improved dissemination of public health information and intelligence directly to the public) was also offered as a solution, and summed up by one participant in the phrase "seeing is believing". It is worth noting, however, that a number of participants displayed a certain level of pessimism that until a crisis or extreme event occurs, no amount of education or awareness-increasing activities would make a difference.

Public health professionals generally prefer disaggregate, case-level data, but access to this data is an issue. The limitations imposed by privacy on public health have resulted in the development of a variety of techniques for data anonymisation [15,23,45]. However, all unavoidably have their issues, risks and limitations, and there is currently no framework to guide public health professionals in their appropriate use and interpretation.

#### **Generalisability**

Although the findings of this paper may be generalisable to public health professionals in Canada and the UK, issues of privacy and public health are not unique to these countries. Privacy is defined as a fundamental human right in the legislation of many countries, and the concept is enshrined in Article 12 of the United Nations' *Universal Declaration of Human Rights* [46] and Article 8 of the *European Convention on Human Rights* [47]. Similarly, public health is an international discipline; both diseases and information are ubiquitous, and neither is constrained by political boundaries and oceans. The increasing requirement for spatial data and its inherent clash with privacy legislation therefore extend beyond the UK and Canadian contexts, and the results, requirements and conclusions drawn from this research can be generalised to wherever such a clash exists. The implementation of solutions by national governments may be further exacerbated by issues of social political trust. General public distrust in government initiatives and motives, such as in most countries of the European Union, Canada, and the United States [48,49], complicates changes that may be perceived by the public to be intrusions of privacy. Such issues may currently be less of a concern in countries such as Finland, Sweden, Denmark, and the Netherlands, where social political trust, though declining, has traditionally tended to be much higher [50-53]. However, even in such nations where privacy and health have traditionally not clashed,

increased international data sharing requirements and spatial data implications may pose unanticipated and challenging obstacles.

#### **Limitations**

No comprehensive lists of public health and health GIS professionals were found in either country, so it was not possible to invite a random sample. In addition, the response rate in the UK was relatively low, and it is therefore uncertain that the sample is representative of all public health professionals in the two countries. However, responses between the two countries were consistent, with no significant differences.

Since knowledge of privacy legislation and policies was based on self-rated scores, a thorough review and assessment of privacy legislation as it pertains to public health practice is required in both Canada and the UK to validate the findings of this survey.

A number of limitations and issues pertaining to the web-survey were identified. Most notable of these was the presence of a scroll bar in sections II and III which most participants missed, thereby eliminating the ability to capture items in reference to "place", such as usefulness. However, these items were also captured more broadly in other sections of the survey. Other issues involved the inability of the architecture to support various designs and types of questions that would have facilitated the completion of the survey, and shortened the length of time required. Participants also noted frustration with the navigation and structure of the survey pages. A document outlining these issues and others was submitted to the ALPHA team after the initial pilot for future enhancements to the architecture.

#### **Conclusion**

It is clear that privacy is perceived to be a major obstacle and issue for public health – the literature illustrates it, and the current study provides both quantitative and qualitative evidence. Together, these provide a more holistic portrayal of public health community viewpoints, and can be used to educate the public, and as evidence for decision makers to implement changes in policies and legislation. The clash between a requirement for personally identifiable data – including exact, individual location – by public health professionals, and the limitations imposed by privacy and its associated bureaucracy, must be addressed and appropriate solutions developed, particularly given the increasing utilisation of geographic information systems in public health and the imminent completion of comprehensive electronic health systems. Privacy legislation is critical for the protection of this fundamental human right, and to prevent the abuse of personal information, particularly in the health field.

However, the legislation must be harmonised with the requirements of public health practice if the health of societies and populations is to be maintained and improved. Since health is not limited by political boundaries, this must be pursued at an international level, and solutions must address these perceptions in the public health community, simplify the bureaucratic process, promote multidisciplinary discussions between legislators, bureaucrats and the public health community, educate communities, and develop and provide public health professionals with toolsets, algorithms and guidelines for using and reporting on disaggregate data. While the results of this study should inform and justify the development of techniques that better anonymise health data with minimal impact on its integrity and frameworks for implementing them, it seems fitting to echo the warning of Curtis et al: "...health and spatial scientists should be proactive and suggest a series of point level spatial confidentiality guidelines before governmental decisions are made which may be reactionary toward the threat of revealing confidential information, thereby imposing draconian limits on research using a GIS [27]."

### Competing interests

PA is an epidemiologist working directly with Geographic Information Systems, and pursuing PhD research on issues of privacy and spatial health information. There are no other competing interests.

### Authors' contributions

PA conceived, designed and carried out the survey (part of his PhD research at the Peninsula Postgraduate Health Institute, Plymouth, UK), and analysed the results. He also conducted the literature review and wrote the draft paper. MNKB and RJ reviewed and provided critical input to the study design and to the discussion and interpretation of its results and their implications. They also helped with the ethical approval process in the UK, identifying UK respondents, the literature review, and the revision of the draft paper. All authors have read and approved the final manuscript.

### Additional material

#### Additional file 1

*The impact of privacy on public health practice: Public health professional questionnaire, CANADA. This is the paper version of the Canadian-English survey that was adapted for the web to collect public health professional perspectives and requirements on location-privacy in Canada.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2458-8-156-S1.pdf>]

#### Additional file 2

*L'impact de la protection des renseignements personnels sur la pratique en santé publique, CANADA. This is the paper version of the Canadian-French survey that was adapted for the web to collect public health professional perspectives and requirements on location-privacy in Canada.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2458-8-156-S2.pdf>]

#### Additional file 3

*The impact of privacy on public health practice: Public health professional questionnaire, UNITED KINGDOM. This is the paper version of the UK survey that was adapted for the web to collect public health professional perspectives and requirements on location-privacy in Canada.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2458-8-156-S3.pdf>]

### Acknowledgements

This research was funded and supported by the Office of Public Health Practice, the Public Health Agency of Canada. Many thanks to Professors Gerard Rushton [54] and Markku Löytönen [55] for providing insightful comments on the survey during the piloting phase. An early and much less detailed report on this study was presented at the 19<sup>th</sup> Annual IUHPE World Conference on Health Promotion and Health Education, held in Vancouver, British Columbia, Canada, June 10–15, 2007 [56].

### References

1. Last JM: *A dictionary of public health* Edited by: Last JM. New York, Oxford University Press, Inc.; 2007.
2. Last JM: *A dictionary of epidemiology* Fourth edition. Edited by: Spasoff RA, Harris SS and Thuriaux MC. New York, Oxford University Press, Inc.; 2001.
3. **National Library of Medicine - Medical Subject Headings 2007** [<http://www.nlm.nih.gov/cgi/mesh/2007/MB.cgi?mode=&index=21704&field=call&HM=&I=&PA=&form=&input=>]. National Institutes of Health
4. Boulos MNK: **Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom.** *International Journal of Health Geographics* 2004, **3**:1 [<http://www.ij-healthgeographics.com/content/3/1/1>].
5. Warren SD, Brandeis LD: **The right to privacy.** *Harvard Law Review* 1890, **4**:193-220.
6. Naylor D, Basrur SH, Bergeron MG, Brunham RC, Butler-Jones D, Dafoe G, Ferguson-Paré M, Lussing F, McGeer A, Neufeld KR, Plummer F: **Learning from SARS: renewal of public health in Canada. A report of the National Advisory Committee on SARS and Public Health.** *Volume Publication Number: 1210.* Ottawa, Ontario, Canada, Health Canada; 2003.
7. Hulton L, Brandon G, McAlister S, Sage J: **Better local information. From breastfeeding to badgers - Brighton & Hove's one stop interactive statistics and mapping service.** 2004 [<http://www.citystats.org>]. United Kingdom, South East England Public Health Observatory, NHS
8. Robinson B: **Privacy, funding doubts shutter Calif. RHIO.** 2007 [<http://www.govhealthit.com/online/news/97855-1.html>].
9. **E-Health Insider: Researchers underline need for access to records** 2007 [[http://www.e-health-insider.com/news/2766/researchers\\_underline\\_need\\_for\\_access\\_to\\_records](http://www.e-health-insider.com/news/2766/researchers_underline_need_for_access_to_records)].
10. Munro M: **Our privacy rules 'block health research': Important studies held back, scientist says.** *The Vancouver Sun* 2004:A5.
11. Khamsi R: **Strict data protection may stifle health research.** *NewScientist.com* [<http://www.newscientist.com/article/dn8595.html>]. 2006

12. Annas GJ: **Book Review: Is privacy the enemy of public health.** *Health Affairs* 1999, **18**:197-198.
13. Bayer R, Colgrove J: **Public health vs. civil liberties.** *Science* 2002, **297**:1811 [<http://www.sciencemag.org>].
14. *Law in Public Health Practice* Second edition. Edited by: Hoffman RE, Lopez WV, Matthews GW, Rothstein MA and Foster KL. New York, Oxford University Press; 2007.
15. Council NR: *Putting people on the map: protecting confidentiality with linked social-spatial data* Edited by: Gutmann MP and Stern PC. Washington, DC, The National Academies Press; 2007.
16. **Health and Social Care Act 2001** 2001 [[http://www.opsi.gov.uk/acts/acts2001/ukpga\\_20010015\\_en\\_1](http://www.opsi.gov.uk/acts/acts2001/ukpga_20010015_en_1)]. c. 15 (UK)
17. **The Privacy Act R.S.** 1985 [[http://laws.justice.gc.ca/en/P-21/section-\[section-no\].html](http://laws.justice.gc.ca/en/P-21/section-[section-no].html)]. c. P-21 (Canada)
18. **Personal Information Protection and Electronic Documents Act** 2000 [[http://laws.justice.gc.ca/en/P-8.6/section-\[section-no\].html](http://laws.justice.gc.ca/en/P-8.6/section-[section-no].html)]. c. 5 P-8.6 (Canada)
19. **Data Protection Act 1998** 1998 [<http://www.opsi.gov.uk/ACTS/acts/1998/19980029.htm>]. c. 29 (UK)
20. **Directive 95/46/EC of the European Parliament and of the council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data** 1995:0031-0050 [<http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML>]. OJL 281
21. POST: **Data protection & medical research.** 2005. POSTnote 235
22. Pickle LW, Szczyr M, Lewis DR, Stinchcomb DG: **The crossroads of GIS and health information: a workshop on developing a research agenda to improve cancer control.** *International Journal of Health Geographics* 2006, **5**: [<http://www.ij-healthgeographics.com/content/5/1/51>].
23. Boulos MNK, Cai Q, Padgett JA, Rushton G: **Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses.** *Journal of Biomedical Informatics* 2006, **39**:160-170.
24. Jeffery C, Ozonoff A, Forsberg L, Nuño M, Pagano M: **The cost of obfuscation when reporting locations of cases in syndromic surveillance systems.** *Advances in Disease Surveillance* 2006, **1**:36.
25. Olson KL, Grannis SJ, Mandl KD: **Privacy protection versus cluster detection in spatial epidemiology.** *American Journal of Public Health* 2006, **96**:2002-2008.
26. Mei-Po K, Casas I, Schmitz BC: **Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks?** *Cartographica* 2004, **39**:15-28.
27. Curtis AJ, Mills JW, Leitner M: **Spatial Confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina.** *Int J Health Geogr* 2006, **5**:44.
28. Brownstein JS, Cassa CA, Kohane IS, Mandl KD: **An unsupervised classification method for inferring original case locations from low-resolution disease maps.** *International Journal of Health Geographics* 2006, **5**.
29. Brownstein JS, Cassa CA, Mandl KD: **No Place to Hide — Reverse Identification of Patients from Published Maps.** *New England Journal of Medicine* 2006, **355**:1741-1742.
30. Barrett G, Cassell JA, Peacock JL, Coleman MP: **National survey of British public's views on use of identifiable medical data by the National Cancer Registry.** *British Medical Journal* 2006, **332**:1068-1072.
31. **Fact Sheet: The Privacy Act: Not an excuse to promote secrecy** 2006 [[http://www.privcom.gc.ca/fs-fi/02\\_05\\_d\\_29\\_e.asp](http://www.privcom.gc.ca/fs-fi/02_05_d_29_e.asp)]. Office of the Privacy Commissioner of Canada
32. Marchildon GP: **Health Systems in Transition - Canada.** 2005, **7(3)**: [<http://www.euro.who.int/document/e87954.pdf>]. World Health Organization
33. Esmail N: **The black hole that is Canada's medicare.** *Times Colonist* 2005 [[http://www.fraserinstitute.org/Commerce.web/article\\_details.aspx?pubID=3526](http://www.fraserinstitute.org/Commerce.web/article_details.aspx?pubID=3526)].
34. Irvine B, Ferguson S, Cackett B: **Background briefing: the Canadian health care system.** 2005 [[http://www.cne.org/pub\\_pdf/2002\\_08\\_health\\_care\\_in\\_canada.pdf](http://www.cne.org/pub_pdf/2002_08_health_care_in_canada.pdf)]. Centre for the New Europe
35. **PersonPlaceTime.org** 2008 [<http://www.personplacetime.org>].
36. Turner C, Bishay H, Peng B, Merifield A: **The ALPHA project: an architecture for leveraging public health applications.** *International Journal of Medical Informatics* 2006, **75**:741-754.
37. NHS: **NHS Health Informatics Community.** 2007 [<http://www.espace.connectingforhealth.nhs.uk/>].
38. **The Map & Data Exchange (Public Health Agency of Canada)** 2007 [<https://php-psp.phac-aspc.gc.ca/>].
39. Eysenbach G: **Improving the quality of web surveys: the checklist for reporting results of internet e-surveys (CHERRIES).** *Journal of Medical Internet Research* 2004, **6(3e34)** [<http://www.jmir.org/2004/3/e34>]. doi:10.2196/jmir.6.3.e34
40. Souhami R: **Governance of research that uses identifiable personal data.** *British Medical Journal* 2006, **333**:315-316.
41. Singleton P, Wadsworth M: **Consent for the use of personal medical data in research.** *British Medical Journal* 2006, **333**:255-258.
42. Hewison J, Haines A: **Overcoming barriers to recruitment in health research.** *British Medical Journal* 2006, **333**:300-302.
43. **Canadian Institutes of Health Research** 2007 [<http://www.cihr-irsc.gc.ca/e/23019.html>].
44. **Canadians and the privacy landscape.** 2007 [[http://www.privcom.gc.ca/information/survey/2007/ekos\\_2007\\_02\\_e.asp](http://www.privcom.gc.ca/information/survey/2007/ekos_2007_02_e.asp)]. Office of the Privacy Commissioner of Canada
45. Cassa CA, Grannis SJ, Overhage JM, Mandl KD: **A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection.** *J Am Med Inform Assoc* 2006, **13(2)**:160-165.
46. **Universal Declaration of Human Rights** 1948 [<http://www.un.org/Overview/rights.html>]. United Nations General Assembly Res. 217 A (III)
47. **Convention for the Protection of Human Rights and Fundamental Freedoms** 1950 [<http://conventions.coe.int/Treaty/en/Treaties/Htm/005.htm>]. Council of Europe C.E.T.S No. 005
48. Sims H: **Public confidence in government, and government service delivery.** 2001 [[http://www.cspc-efpc.gc.ca/Research/publications/pdfs/HarveySimms\\_e.pdf](http://www.cspc-efpc.gc.ca/Research/publications/pdfs/HarveySimms_e.pdf)]. Canadian Centre for Management Development Report # P105B
49. Nye JS: **In government we don't trust.** *Foreign Policy* 1997:99-111 [<http://www.foreignpolicy.com/Ning/archive/archive/108/ingovwedonttrust.pdf>].
50. Green JM, Draper AK, Dowler EA, Fele G, Hagenhoff V, Rusanen M, Rusanen T: **Public understanding of food risks in four European countries: a qualitative study.** *European Journal of Public Health* 2005, **15**:523-527.
51. Newton K: **Political support: social capital, civil society and political and economic performance.** *Political Studies* 2006, **54**:846-864.
52. Dalton RJ: **The social transformation of trust in government.** *International Review of Sociology* 2005, **15**:133-154.
53. Hudson J: **Institutional trust and subjective well-being across the EU.** *Kyklos* 2006, **59**:43-62.
54. **Geography at the University of Iowa - Gerard Rushton** 2008 [<http://www.uiowa.edu/~geog/faculty/rushton.htm>].
55. **University of Helsinki - Department of Geography - Markku Löytönen** 2008 [[http://www.helsinki.fi/geography/Markku\\_Loytonen\\_eng.html](http://www.helsinki.fi/geography/Markku_Loytonen_eng.html)].
56. **The 19th IUHPE World Conference on Health Promotion & Health Education** 2008 [[http://www.iuhpeconference.org/en/conference/abstract\\_view.php?ID=329688](http://www.iuhpeconference.org/en/conference/abstract_view.php?ID=329688)].

### Pre-publication history

The pre-publication history for this paper can be accessed here:

<http://www.biomedcentral.com/1471-2458/8/156/prepub>