2014-07

Mining textual data from primary healthcare records: Automatic identification of patient phenotype cohorts

Zhou, Shang-Ming

10.1109/ijcnn.2014.6889494
2014 International Joint Conference on Neural Networks (IJCNN)
IEEE

Mining Textual Data from Primary Healthcare Records
- Automatic Identification of Patient Phenotype Cohorts

Shang-Ming Zhou, Muhammad A Rahman, Mark Atkinson, Sinead Brophy

Abstract— **Due to advances of the "omics" technologies, rich sources of clinical, biomedical, contextual, and environmental data about each patient have been available in medical and health sciences. However, an enormous amount of electronic health records is actually generated as textual data, such as descriptive terms/concepts. No doubt, efficiently harnessing these valuable textual data would allow doctors and nurses to identify the most appropriate treatments and the predicted outcomes for a given patient in real time. We used textual data to identify patient phenotypes from UK primary care records that were managed by Read codes (a clinical classification system). The fine granularity level of Read codes leads to a huge number of clinical terms to be handled. Unfortunately, traditional medical statistics methods have struggled to process this sort of data effectively. In this paper, we described how the problem of patient phenotype identification can be transformed into document classification task, a text mining scheme is addressed to integrate feature ranking methods and genetic algorithm to identify the most parsimonious subset of features that still holds the capacity of characterizing the distinction of patient phenotypes. The experimental results have demonstrated that compact feature sets with 2 or 3 important terms describing clinical events were effectively identified from 16852 Read codes while their classification accuracy remained high level of agreements with specialists from secondary care in classifying testing samples.**

I. Introduction

ROUTINELY collected primary care data provides an enormous opportunity for clinical and translational research, such as clinical trial recruitment, outcome prediction, survival analysis, and other kinds of retrospective studies [1-3]. The patient records held by the general practices in primary care offer patients a picture of morbidity from cradle to grave. During routine clinic visits, if a primary care physician has suspicion of a disease, such as ankylosing spondylitis, in a patient, the patient will be referred to a specialist, such as a rheumatologist, who will order further diagnostic tests in secondary care until a diagnosis result is reached. Often this is a very time consuming and expensive process [4, 5]. So any approaches and techniques which can speed up the identification and validation of disease diagnosis are of particular interest for improving the efficiency of healthcare.

Specific validation studies have suggested high validity of diagnoses recorded in the primary care data [6-9]. However, generally speaking, primary care data has been underused for discovery research due to the difficulty of extracting highly accurate clinical data and lack of appropriate medical statistical methods for analyzing complex types of data[10, 11]. This paper addresses a novel way of utilizing primary care records to automatically identify early predictors of disease progression about ankylosing spondilitis (AS) illness phenotypes via integration of machine learning, statistical feature selection and genetic algorithm[12].

The AS is one of the most common forms of inflammatory arthritis. As a type of progressive (long-term) inflammatory disease, AS mainly affects the spine and the sacroiliac joints (in the pelvis), including bones, muscles and ligaments. It often causes quality of life impairment and functional limitations [13, 14], similar to or worse than cancer, congestive heart failure, diabetes or depression [15], with around one in 10 AS patients at risk of long-term disability[16]. The AS patients are usually diagnosed many years after onset of symptoms, and there is no cure for AS. So long-term outcome in AS patients is highly dependent upon an early recognition, aggressive control and therapy of inflammation. When properly treated, AS patients could lead fairly normal lives.

However, within the current fragmented health care system, it is extremely time-consuming to identify and recruit AS patients for large cohort study in genetic validation studies and clinical trials of new therapies. Fortunately, patient records collected routinely from primary care can provide a rich source of data which offers a way of tackling this issue. But extracting meaningful pieces of information from primary care database for identification of patients who satisfy predefined criteria remains a challenging task, because an enormous amount of primary care records is actually generated as textual data, such as descriptive terms/concepts, whilst these criteria are buried within the textual data across multiple data points in the electronic healthcare record of a patient, moreover clinical text is the most difficult data type to analyse computationally[17]. Primary care informatics is emerging as a discipline of academic scientific study about how to harness these data [18, 19].

The objectives of this paper are to develop a decision support model ("*Virtual Rheumatologist*") that can automatically identify AS illness phenotype based on clinical events from primary care records; identify early risk factors, efficient measures for assessments, treatments and diagnosis of AS; and speed up the collection of AS patient

Shang-Ming Zhou, Public Health Informatics Group, College of Medicine, Swansea University SA2 8PP, UK. (e-mail: smzhou@ieee.org; s.zhou@swansea.ac.uk).

Muhammad A Rahman, Public Health Informatics Group, College of Medicine, Swansea University SA2 8PP, UK. (email: m.a.Rahman@swansea.ac.uk)

Mark Atkinson, Public Health Informatics Group, College of Medicine, Swansea University SA2 8PP, UK.(email: m.atkinson@Swansea.ac.uk)

Sinead Brophy, Public Health Informatics Group, College of Medicine, Swansea University SA2 8PP, UK. (e-mail: s.brophy@swansea.ac.uk).

cohorts from the electronic health records for further genetic studies and clinical trials of new interventions. To fulfil these tasks, we proposed a framework of mining primary care textual data for automatic extraction of relevant clinical information to identify early predictors of disease progression, in which the Naïve Bayes classifier [20] was used to classify the outcome of a patient as AS patient or non-AS (NAS) patient, whilst the feature ranking based feature selection (FS) methods and genetic algorithm (GA) were integrated to selects the most relevant feature subsets from primary care records. The motivation of integrating feature ranking based FS and GA was driven by the fact that a huge number of coded terms (acting as features) are involved in automatic identification of AS illness phenotypes, but feature subsets selected by either feature ranking based FS methods or GA alone are not so parsimonious as we expected in this study for clinical decision supports.

The organisation of this paper is as follows. Section 2 presents how the problem of disease phenotype identification can be treated as a document classification task. Then the scheme of selecting the most relevant clinical Read codes is addressed in Section 3. Section 4 presents the experimental results. And Section 5 concludes this paper.

II. Treating Disease Phenotype Identification as Document Classification Task

At the first sight, identifying patient cohorts and document classification are different problems. In this section, we address how the two problems can be innovatively linked together, so primary care informatics could emerge as a wide field to which text mining techniques can be applied.

A. Read Code Briefing

The RCT system with 5-bytes provides around 83000 clinical descriptive terms in hierarchical structure comprising five levels of detail, whilst each successive level offers more detail to a concept. There are 3 top level of hierarchy which refer to *Process of Medicine* starting with a number (0 to 9), *Diagnoses* starting with a capital letter (A to Z), and *Medication and Appliances* starting with a lower case letter (a-y) respectively. Table1 depicts examples of artificial primary care records.

TABLE I. EXAMPLES OF PRIMARY CARE RECORDS

Encrypted Patient ID	Date	Event Value	Read Code
P1001	01/11/1988	0	N32..
P1001	03/11/1990		S5...
P1001	03/11/1990	20	136..
P1001	…		…
P1001	03/11/1990	0	S5...
P1001	03/03/2002		4266.
P1001	03/03/2002	0	N100.
P1001	03/03/2002		jA52.
P1002	13/11/1988		912C.
P1002	14/12/1988	20	137P.
P1002	…		…
P1002	18/06/1990	14.3	42Z7.
P1002	10/06/1998	0	S5...
P1002	28/11/1998	0	F4J0.

Fig. 1. The procedure of AS patient healthcare

In the RCT based primary care database, although there may be a field like "*Event Value*" that records the values of corresponding Read codes, such as laboratory test results, number of cigarettes, amount of alcohol etc, our analysis demonstrated that this field has too much inaccurate, inconsistent and missing information to be used for research on AS illness phenotype identification. So the only available information for this study is the occurrences of Read codes indicating clinical events. For example, the patient – *P1001* had the profile described by the sequence of Read codes – [*"N32..", "S5..." , "136..", ..., "S5...", "4266.", "N100.", "jA52."*]. One event may occur multiple times during different dates of clinic visits. It is noted that the code – "*N100.*" in primary care records indicated the case where the general practice doctor had a suspicion of AS in the patient. Normally the patient with suspected disease in general practice will be referred to a specialist, such as a rheumatologist, who will perform further diagnostic tests until a confirmation is reached as shown in Figure 1. We cast the task of identifying AS illness phenotypes as a text classification problem which treats all distinct clinical terms from Read codes as the vector of feature and each profile of patient composing a series of Read codes as a type of document.

B. Naive Classifier

In the document classification for AS patient identification, given the patient conditions described by a group of Read codes (treated as a document), a classifier is to be developed to assess whether the patient is AS diseased or not. In this study, we used the Naive Bayesian classifier [20] to perform this task (see Figure 2).

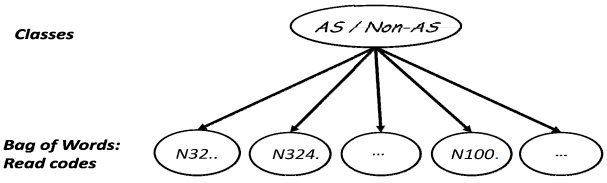

Fig. 2. Naïve Bayesian classifier for classifying AS and Non-AS patients

Imagine that documents are drawn from two classes of AS patients (ASP) and non-AS patients (NAP) which are modelled as sets of Read codes (terms). The (independent) probability that the *i-th* Read code r_i of a given patient occurs in a document from class $c \in \{ASP, NAP\}$ is written as $p(r_i|c)$. We assume that the Read codes are randomly distributed in the document. Often patients had multiple

clinic visits for different or same types of clinical events. The order of these events is assumed not to carry useful information for discriminating the AS and non-AS patients, which is concerned with positional independence of Read codes.

The $p(t_i|c)$ can be interpreted as a measure of how much evidence the *i-th* Read code brings given the class c is correct. Then for a given patient r, the Naive Bayesian classifier is a probabilistic model to calculate the posterior probability of the patient being in the class c,

$$p(c|r) = \frac{p(c)p(r|c)}{p(r)}$$

Because the denominator does not depend on c and the values of r are given, so that the denominator is effectively constant across the classes. Then given a set of training samples, to build a Naive Bayesian classifier, we only need to estimate the prior $p(c)$ and the likelihood $p(r|c)$ from the training data. The evidence of a Read code contributing to the identity of a patient for a class would depend on its occurrence. More frequent terms (Read codes) are likely to make stronger contribution to classifying the patient than infrequent terms. The prior of the class c is estimated as

$$p(c) = \frac{N_c}{N}$$

where N_c is the number of patients in the class c. And with the conditional independence assumption, the probability of a patient r containing all of the Read codes given a class c, is

$$p(r|c) = \prod_{1 \le i \le N_d} p(r_i|c)$$

where $\langle r_1, r_2, \cdots, r_{N_d} \rangle$ are the tokens of Read codes in the document r and N_d is the number of the tokens in r, whilst the $p(r_i|c)$ is estimated as the relative frequency of the Read code r_i in documents of the class c:

$$p(r_i|c) = \frac{O_{r_i,c}}{\sum_{i=1}^{N_R} O_{r_i,c}}$$

where $O_{r_i,c}$ is the count of occurrences of the term r_i in training documents from class c, and N_R is the number of all distinctive Read codes available in the training documents.

To make prediction for a given test patient r, the *maximum a posteriori principle* is used to assign the patient to a class as

$$c_r = \arg\max_c \left[p(c) \prod_{1 \le i \le N_d} p(r_i|c) \right]$$

III. Selecting the Most Relevant Read Codes

A huge number of Read code terms often need to be handled when identifying the AS illness phenotypes from the UK primary care records. This brings a big analytical and computational challenge.

In order to find the parsimonious set of Read codes that are most relevant to classification of patient illness phenotype, a scheme of integrating feature ranking based FS and GA is used as illustrated in Figure 3. First the available data is split into training samples, validation samples and testing samples. The training samples are used for training the Naïve Bayesian (NB) classifier, the validation samples for testing performance of selected Read codes during feature selection, and the testing samples for evaluating the generalization performance of final selected feature set.

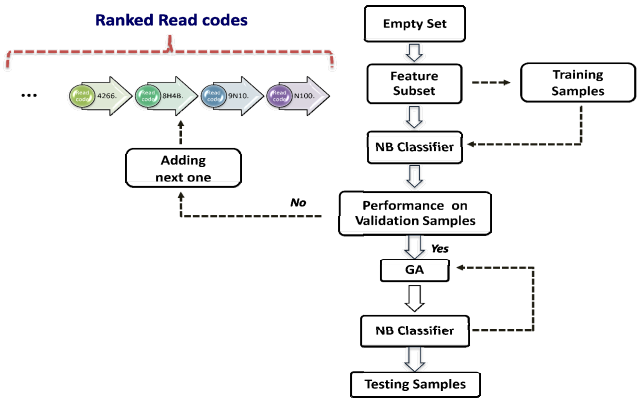

Fig.3. Integration feature selection scheme for identifying the most relevant Read codes.

In this scheme, before conducting the feature selection, one crucial step is to use a feature ranking index to rank the Read codes based on training samples, for example from the most significant term to the least significant term. This is independent to classification task. The feature ranking index measures the capability of each term to distinguish the AS patient from non-AS patient. Then a greedy search starts from empty set, and add one term at a time from the ranked Read codes. A NB classifier is constructed using the training samples with only the selected feature subset. Then the classification performance is evaluated on the validation samples. This process continues until a satisfied performance of selected feature subset is reached. This feature ranking procedure works well in most situations for real world applications. But for the task of identifying AS patients from primary care records based on Read codes, because a huge number of Read codes are involved in the raw data initially, the feature subset selected by feature ranking procedure may still contain too many terms to be interpreted by human experts. One room for further improving the parsimony of selected feature subset is to consider the different combinations of terms which have not been focused in the feature ranking procedure. So in this study, the GA is further used to identify the efficient combination of Read codes from the selected feature subset by feature ranking procedure.

In this study, the feature ranking indices used are based on relative frequency, hypothesis testing statistics and information theory.

A. Feature Ranking via Frequency

Frequency based index is the simplest and straightforward approach to feature ranking. The frequency of a term refers to the number of documents in which the term occurs. So frequency based index aims to select the commonest terms. The rationale of selecting these terms lies in they are often well-estimated and most often available as evidence. However, this sort of feature selecting is based on an unstated assumption that rare terms are either non-informative for classification, or not influential in global prediction performance. The reality is that a low frequency term can be relatively informative about presence or absence of the term contributing to the correct categorization decision, so its contribution to document classification can be significant. Actually in this study many Read codes are infrequent, but demonstrate strong capability to distinguish AS patients from non-AS patients.

B. Feature Ranking via Chi-squared Statistic

In medical statistics, the Chi-squared (χ^2) test of contingency of variables is a widely used statistic to measure the differences between the observed values and those that would be expected if the variables were independent [21]. Small differences indicate little dependence between the variables, whereas large differences show dependence. In text classification, the two events become occurrence of a term and occurrence of a class. The Chi-squared statistic measures the differences between the observed counts and expected counts if two events (the term and the class) were independent. The rationale of using Chi-squared statistic for feature selection lies in that if there is dependence between the two events (a term and a class), then the occurrence of the term leads to the occurrence of the class more likely (or less likely). In our study, using the contingency table of a Read code *r* and a class *c* (see Table 2), the Chi-squared statistic is calculated by

$$\chi^2(r,c) = \frac{N(N_{11}N_{00} - N_{10}N_{01})^2}{N_{.1}N_{.0}N_{1.}N_{0.}}$$

N_{11} represents the counts of the Read code occurring in the documents belonging to the class *c*, N_{10} the counts of the Read code occurring in the documents belonging to other class(es) except c, N_{01} the counts of the Read code not occurring in the documents belonging to the class c, N_{00} the counts of the Read code not occurring in the documents belonging to other class(es), while $N_{.1} = N_{11} + N_{01}$, $N_{.0} = N_{10} + N_{00}$, $N_{1.} = N_{11} + N_{10}$, $N_{0.} = N_{01} + N_{00}$, and $N = N_{.1}+N_{.0}+N_{1.}+N_{0.}$.

TABLE II. CONTINGENCY TABLE OF A READ CODE *R* AND A CLASS *C* (¬*C* REPRESENTS THE OTHER CLASS(ES) EXCEPT C, ¬*R* REPRESENTS THE ABSENCE OF THE TERM *R*)

	c	$\neg c$	
r	N_{11}	N_{10}	$N_{1.}$
$\neg r$	N_{01}	N_{00}	$N_{0.}$
	$N_{.1}$	$N_{.0}$	

Then the importance of the term is evaluated as the largest Chi-squared value of the term across the classes.

C. Feature Ranking via Information Gain

Information gain (IG) has been used for scoring features in many machine learning applications[22]. In text classification, based on information theory IG measures how much information of the presence/absence of a term contributing to the correct category prediction [23-25]. In our study, using the contingency table of a Read code *r* and a class *c* (see Table 2), the IG is calculated by

$$IG(r,c) = \frac{N_{11}}{N} log \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} log \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} log \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} log \frac{NN_{00}}{N_{0.}N_{.0}}$$

Then the importance of the term is evaluated as the largest IG value of the term across the classes which give the most information about the categories in training data.

D. Genetic Algorithm

GA is a heuristic search algorithm inspired by evolutionary ideas of natural selection and genetic[12]. As an intelligent exploitation of a random search within *a defined search space* to solve a problem, the GA was developed to mimic the process of natural system necessary for evolution, such as *inheritance, mutation, selection,* and *crossover*. In our study, the GA was integrated with feature ranking approach to identify the Read codes which are most relevant for explaining the variation of AS and non-AS patients. A binary coding is used in GA, in which 1 represents the presence of a term, 0 its absence. The fitness function is chosen to be area under curve (AUC) in receiver operating characteristic (ROC) plot generated by NB classifier. The AUC is a measure of discrimination to be used as an overall model performance given varied decision boundaries.

IV. EXPERIMENTAL RESULTS

In our study, 898763 primary care events records with 16852 distinctive Read codes were collected from a national e-health infrastructure - SAIL (Secure Anonymised Information Linkage) databank[26]. The rheumatologists from secondary care had made diagnosis on the patients from these records, among which there were 277 AS patients and 1389 non-AS patients. In order to evaluate the overall performance of these Read codes in discriminating AS and non-AS patients, we first conducted NB classification with 10-fold cross validation on these 1666 patients with 16852 Read codes. The overall classification accuracy on predicting testing patients achieved 90.2% of agreements with rheumatologists (95% confidence interval [0.89, 0.92]). However, it is impractical to expect general practitioners and health professionals to evaluate such a huge number of Read codes as risk factors to make prediction for a suspected AS

patient. So it is crucial to select parsimonious subset of Read codes revealing hidden patterns in discriminating AS patients from non-AS patients.

To fulfil this task, the data with 1666 patients is split into 3 subsets: 200 patients were used as validation samples for evaluating performance of selected features, 200 patients as testing samples for evaluating generalisation performance of final feature subset, and the remaining patients as training samples for scoring features and training NB classifiers. The classification accuracy (ACC) and AUC were used as metrics to measure the classification performances under various conditions. Figure 4 depicts the frequencies of Read codes using training samples, while Figure 5 shows the information gains of Read codes and Figure 6 the Chi-squared values of Read codes based on training samples. It can be seen that either information gains or Chi-squared values of Read codes demonstrate very different scoring patterns from frequency metric shown in Figure 4. Table 3 summarises the feature selection results and corresponding performances by the integration scheme (see Figure 3) in comparison with single feature ranking methods.

For clearness, Table 4 shows the top 20 Read codes ranked and selected according to their frequencies. It can be seen that frequency metric tends to give more weights to the *Process of Medicine* events whose Read codes start with numbers 0~9. This is because *Process of Medicine* events normally involved routine checks and tests, their Read codes naturally became frequent terms. But the problem is that frequency metric can select some frequent terms that may have no specific information for classification, for example, routine check events ("*blood pressure reading*", …) which are frequent across classes. On the contrary, *Diagnostic* events (whose Read codes start with capital letters) and *Medication* events (whose Read codes start with lower case letters) tend to occur infrequently, this is because once the general practitioner entered a diagnostic or medication Read code, often he/she would not bother to enter the same code during multiple patient clinic visits. These diagnostic or medication events may make significant contribution to patient classification even they are rare terms. However, clearly frequency based method failed to select influential diagnostic and medication Read codes.

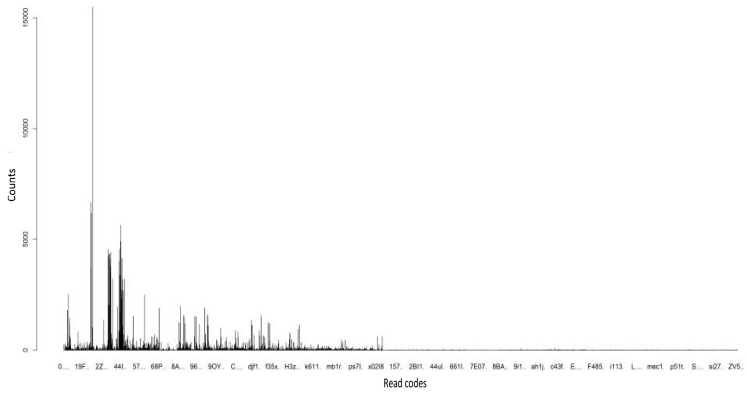

Fig.4. The frequencies of individual Read codes

In comparison, the IG based method and Chi-square based method selected much more compact and meaningful Read codes, which include events from all 3 categories - *Process of Medicine*, *Diagnosis* and *Medication*. Table 6 and Table 7 show the contingency tables of subsets of Read codes selected by IG and Chi-square metrics respectively on testing samples. Then the integration scheme selected multiple combinations of Read codes with the most parsimonious subsets consisting of only 2 or 3 terms (see IS1~IS5 in Table 3, and Tables 8~ 11). Moreover these parsimonious subsets of Read codes achieved outstanding performance in agreements with rheumatologist on classifying the testing patients.

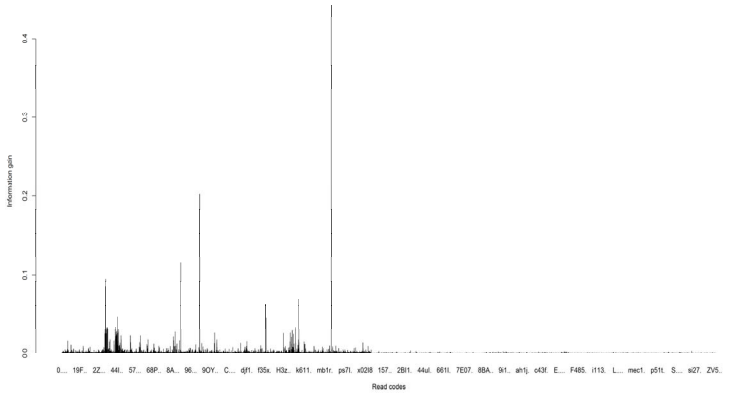

Fig.5. The information gains of individual Read codes

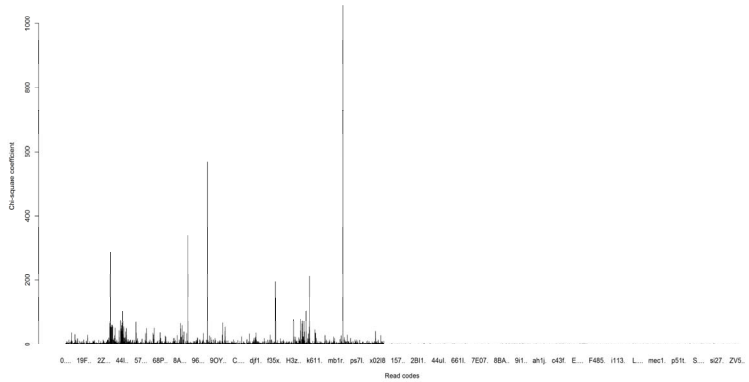

Fig. 6. The Chi-squared values of individual Read codes

TABLE III. SUMMARY OF READ CODE SELECTION BY DIFFERENT APPROACHES WITH NB CLASSIFIER

Method		Number of Read Codes	Test ACC
Raw data		16852	0.902
Frequency based method		1096	0.925
IG based method		11	0.965
Chi-square based method		17	0.965
Integration scheme (IS)	IS1	2	0.985
	IS2	2	0.965
	IS3	3	0.995
	IS4	3	0.995
	IS5	3	0.99

TABLE IV. THE TOP 20 READ CODES SELECTED BY FREQUENCY METRIC

Read code	Read code description
246..	O/E - blood pressure reading
22A..	O/E - weight
22K..	Body Mass Index
44J3.	Serum creatinine
44I4.	Serum potassium
44I5.	Serum sodium
44F..	Serum alkaline phosphatase
423..	Haemoglobin estimation
42P..	Platelet count
42H..	Total white cell count
42A..	Mean corpuscular volume (MCV)
44M4.	Serum albumin
44M3.	Serum total protein
426..	Red blood cell (RBC) count
44E..	Serum bilirubin level
42J..	Neutrophil count
428..	Mean corpusc. haemoglobin(MCH)
229..	O/E - height
42M..	Lymphocyte count
42N..	Monocyte count

TABLE V. THE CONTINGENCY TABLE OF SUBSET OF READ CODES SELECTED BY FREQUENCY BASED METHOD

		Rheumatologist	
		AS	nonAS
Model	AS	21	12
	nonAS	3	164

TABLE VI. THE CONTINGENCY TABLE OF SUBSET OF READ CODES SELECTED BY IG BASED METHOD

		Rheumatologist	
		AS	nonAS
Model	AS	22	5
	nonAS	2	171

TABLE VII. THE CONTINGENCY TABLE OF SUBSET OF READ CODES SELECTED BY CHI-SQUARE BASED METHOD

		Rheumatologist	
		AS	nonAS
Model	AS	23	6
	nonAS	1	170

TABLE VIII. THE IS1 SUBSET OF READ CODES SELECTED BY THE INTEGRATION SCHEME (LEFT), AND ITS CONTINGENCY TABLE (RIGHT)

Read code	Read code description
N100.	Ankylosing spondylitis (suspected)
44I9.	Serum inorganic phosphate

		Rheumatologist	
		AS	nonAS
Model	AS	21	0
	nonAS	3	176

TABLE IX. THE IS2 SUBSET OF READ CODES SELECTED BY THE INTEGRATION SCHEME (LEFT), AND ITS CONTINGENCY TABLE (RIGHT)

Read code	Read code description
N100.	Ankylosing spondylitis (suspected)
9N10.	Seen in rheumatology clinic

		Rheumatologist	
		AS	nonAS
Model	AS	17	0
	nonAS	7	176

TABLE X. THE IS3 SUBSET OF READ CODES SELECTED BY THE INTEGRATION SCHEME (LEFT), AND ITS CONTINGENCY TABLE (RIGHT)

Read code	Read code description
N100.	Ankylosing spondylitis (suspected)
N102.	Sacroiliitis, not elsewhere classified
42L..	Basophil count

		Rheumatologist	
		AS	nonAS
Model	AS	23	0
	nonAS	1	176

TABLE X. THE IS4 SUBSET OF READ CODES SELECTED BY THE INTEGRATION SCHEME (LEFT), AND ITS CONTINGENCY TABLE (RIGHT)

Read code	Read code description
N100.	Ankylosing spondylitis (suspected)
F4430	Anterior uveitis
42L..	Basophil count

		Rheumatologist	
		AS	nonAS
Model	AS	23	0
	nonAS	1	176

TABLE XI. THE IS5 SUBSET OF READ CODES SELECTED BY THE INTEGRATION SCHEME (LEFT), AND ITS CONTINGENCY TABLE (RIGHT)

Read code	Read code description
N100.	Ankylosing spondylitis (suspected)
jA52.	ETORICOXIB 90mg tablets
42L..	Basophil count

		Rheumatologist	
		AS	nonAS
Model	AS	23	1
	nonAS	1	175

V. CONCLUSIONS

Primary care databases provide a unique source of information for research on disease epidemiology. In this paper, we have shown how the textual Read codes data can be used to develop a document-level classifier for identifying AS illness phenotypes from electronic healthcare records. This benefited from a text mining scheme addressed in this paper that integrates feature ranking methods and GA for the sake of selecting the most compact subset of features. We believe this study will help speed up the collection of AS patient cohorts from electronic healthcare records, and identify patients for genetic studies and clinical trials of new interventions that require large sample sizes with precise definitions of disease phenotypes and response to therapies.

REFERENCES

[1] J. S. Mathias, D. Gossett, and D. W. Baker, "Use of electronic health record data to evaluate overuse of cervical cancer screening," *J Am Med Inform Assoc,* vol. 19, pp. e96-e101, Jun 2012.

[2] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, *et al.*, "A review of approaches to identifying patient phenotype cohorts using electronic health records," *J Am Med Inform Assoc,* Nov 7 2013.

[3] M. Staff, "Can data extraction from general practitioners' electronic records be used to predict clinical outcomes for patients with type 2 diabetes?," *Inform Prim Care,* vol. 20, pp. 95-102, 2012.

[4] P. D. Miller, "Claim that primary care provides in excess of 80% NHS activity is wrong," *BMJ,* vol. 347, p. f6163, 2013.

[5] D. Chicco, A. Taddio, G. Sinagra, A. Di Lenarda, F. Ferrara, M. Moretti, *et al.*, "Speeding up coeliac disease diagnosis in cardiological settings," *Arch Med Sci,* vol. 6, pp. 728-32, Oct 2010.

[6] E. Herrett, S. L. Thomas, W. M. Schoonen, L. Smeeth, and A. J. Hall, "Validation and validity of diagnoses in the General Practice Research Database: a systematic review," *Br J Clin Pharmacol,* vol. 69, pp. 4-14, Jan 2010.

[7] E. L. Herrett, S. L. Thomas, and L. Smeeth, "Validity of diagnoses in the general practice research database," *Br J Gen Pract,* vol. 61, pp. 438-9, Jul 2011.

[8] S. S. Jick, J. A. Kaye, C. Vasilakis-Scaramozza, L. A. Garcia Rodriguez, A. Ruigomez, C. R. Meier, *et al.*, "Validity of the general practice research database," *Pharmacotherapy,* vol. 23, pp. 686-9, May 2003.

[9] N. F. Khan, S. E. Harrison, and P. W. Rose, "Validity of diagnostic coding within the General Practice Research Database: a systematic review," *Br J Gen Pract,* vol. 60, pp. e128-36, Mar 2010.

[10] R. L. Tannen, M. G. Weiner, and D. Xie, "Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial Findings," *BMJ,* vol. 338, p. b81, 2009.

[11] K. P. Liao, T. Cai, V. Gainer, S. Goryachev, Q. Zeng-treitler, S. Raychaudhuri, *et al.*, "Electronic medical records for discovery research in rheumatoid arthritis," *Arthritis Care Res (Hoboken),* vol. 62, pp. 1120-7, Aug 2010.

[12] D. Goldberg, *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Norwell, MA: Kluwer Academic Publishers, 2002.

[13] S. Brophy, R. Cooksey, M. Atkinson, S. M. Zhou, M. J. Husain, S. Macey, *et al.*, "No increased rate of acute myocardial infarction or stroke among patients with ankylosing spondylitis-a retrospective cohort study using routine data," *Semin Arthritis Rheum,* vol. 42, pp. 140-5, Oct 2012.

[14] S. Brophy, R. Cooksey, H. Davies, M. S. Dennis, S. M. Zhou, and S. Siebert, "The effect of physical activity and motivation on function in ankylosing spondylitis: a cohort study," *Semin Arthritis Rheum,* vol. 42, pp. 619-26, Jun 2013.

[15] J. Braun, N. McHugh, A. Singh, J. S. Wajdula, and R. Sato, "Improvement in patient-reported outcomes for patients with ankylosing spondylitis treated with etanercept 50 mg once-weekly and 25 mg twice-weekly," *Rheumatology (Oxford),* vol. 46, pp. 999-1004, Jun 2007.

[16] NHSChoices. Ankylosing spondylitis. *(2 December 2013)*.

[17] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nat Rev Genet,* vol. 13, pp. 395-405, Jun 2012.

[18] S. de Lusignan and C. van Weel, "The use of routinely collected computer data for research in primary care: opportunities and challenges," *Fam Pract,* vol. 23, pp. 253-63, Apr 2006.

[19] S. de Lusignan, "What is primary care informatics?," *J Am Med Inform Assoc,* vol. 10, pp. 304-9, Jul-Aug 2003.

[20] W. Wei, S. Visweswaran, and G. F. Cooper, "The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data," *Journal of the American Medical Informatics Association,* vol. 18, pp. 370-375, 2011.

[21] R. H. Riffenburgh, *Statistics in Medicine, Third Edition*: Academic Press, 2012.

[22] J. R. Quinlan, *C4.5: Programs for Machine Learning*: Morgan Kaufmann, 1993.

[23] G. Uchyigit and K. Clark, "A New Feature Selection Method For Text Classification," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 21, pp. 423–438, 2007.

[24] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," presented at the Proc. of the 14th Int. Conference on Machine Learning, 1997.

[25] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*: Cambridge University Press, 2009.

[26] R. A. Lyons, K. H. Jones, G. John, C. J. Brooks, J. P. Verplancke, D. V. Ford, *et al.*, "The SAIL databank: linking multiple health and social care datasets," *BMC Med Inform Decis Mak,* vol. 9, p. 3, 2009.