2022

Bayesian Methods for the Design and Analysis of Cluster Randomised Controlled Trials

Jones, Benjamin Gary

http://hdl.handle.net/10026.1/19529

http://dx.doi.org/10.24382/687
University of Plymouth

UNIVERSITY OF PLYMOUTH

Bayesian Methods for the Design and Analysis of Cluster Randomised Controlled Trials

by

Benjamin Gary Jones

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

Peninsula Medical School

April 2022

Acknowledgements

WITH huge thanks to my supervisory team Dr Amy Baker, Professor Siobhan Creanor, Dr Rana Moyeed and Dr Adam Streeter, for their unending support, advice and patience throughout this project. Thanks also to Professor Obioha Ukoumunne and Dr Nick Axford for their advice and feedback during the transfer viva. And to all of my colleagues past and present at the Universities of Plymouth and Exeter, for interesting and helpful discussions around the topics contained within this thesis and beyond.

In heartfelt appreciation of my wife, who's support, both emotional and practical, has been instrumental in my completion of this project.

And finally, in memory of my Dad, who's advice and support continues to guide me to this day.

Authors declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

Word count for the main body of this thesis: **55,537**

Signed: ________________________

Date: 04 August 2022

Publications:

Jones, B.G., Streeter, A.J., Baker, A., Moyeed, R. & Creanor, S. *Bayesian Statistics in the design and analysis of cluster randomised controlled trials and their reporting quality. A methodological systematic review.* BMC Systematic Reviews 2021, 10:91 doi: 10.1186/s13643-021-01637-1

Posters and conference presentations:

Current Developments in Cluster Randomised Trials and Stepped-Wedge Designs, Pragmatic Clinical Trials Unit, Queen Mary University London. Oral presentation: *A Bayesian Power Prior Approach for Incorporating Pilot Data into Cluster Randomised Controlled Trial Analysis and Design.* 2021.

Young Statisticians' Meeting, University of Manchester. Oral presentation: *A Bayesian Power Prior approach for incorporating pilot data into Cluster Randomised Controlled Trial analysis and design.* 2020.

5th International Clinical Trials Methodology Conference, Brighton. Poster presentation: *Bayesian Statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review.* 2019.

Young Statisticians' Meeting, University of Leeds. Oral presentation: *Using Bayesian Methods to Account for Uncertainty in Parameter Estimations Used in the Design of Cluster Randomised Controlled Trials.* 2019.

Abstract

Bayesian Methods for the Design and Analysis of Cluster Randomised Controlled Trials

Benjamin Gary Jones

CLUSTER Randomised Controlled Trials involve randomising groups of participants, rather than the individual participants themselves, whilst the outcomes are measured on the participants. Whilst there are a number of practical and methodological advantages to such a design, there are also statistical implications, both in terms of study design and sample size calculation, and in analysis. The methodology underpinning the cluster randomised design is now well-established in the statistical literature. However, the overwhelming majority of methodological developments to date have been within the frequentist paradigm, and as such, there is an opportunity to explore methodological developments in the context of Bayesian approaches to the design and analysis of Cluster Randomised Controlled Trials, which is the focus of this thesis.

This thesis begins by identifying and quantifying the practical application of Bayesian methods to such cluster randomised trial designs, as well as existing methodological developments in the area, through a methodological systematic review. The review highlights that whilst there have been some efforts to develop Bayesian methodology for Cluster Randomised Controlled Trials, the practical uptake of such methods remains low.

Next, a novel application of an informative class of prior distribution, the power prior, is proposed whereby information is borrowed from continuous, clustered, historical data, such as that from a pilot or feasibility study. The performance of this approach is evaluated, and superiority, in comparison to established methods, is demonstrated for certain performance metrics. The novel application of the power prior methodology is then explored in the context of study design and sample size calculation for a Cluster Randomised Controlled Trial, whereby the impact of the use of these new methods is quantified in the context of the impact on type I error and statistical power. It is demonstrated that the adoption of these methods has the potential to reduce sample size requirements, thereby facilitating more efficient trial design and reducing research waste. However, it is also shown that, under the traditional frequentist interpretation, inflated type I error rates can be expected as a result of borrowing information through the power prior.

In order to address the limitation of inflated type I error, an approach is presented in which the degree of information borrowing through the power prior is determined in order to control Bayesian type I error at some nominal level. It is shown that by adopting a Bayesian interpretation of design operating characteristics, information borrowing methods can be used whilst maintaining type I error control.

Finally, a newly developed R package, `PPCRCT` is described which allows for straightforward implementation of the methodology presented within this thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike's Information Criterion
BF	Bayes Factor
BMA	Bayes model averaging
BMI SDS	Body Mass Index Standard Deviation Score
BUGS	Bayesian inference Using Gibbs Sampling
CA	Clipped Alternative
CENTRAL	Cochrane Central Register of Controlled Trials
CI	Confidence Interval
COiC	Childhood Obesity in China
CONSORT	Consolidated Standards of Reporting Trials
COVID-19	Coronavirus Disease 2019
CRAN	Comprehensive R Archive Network
CRCT	Cluster Randomised Controlled Trial
CrI	Credible Interval
CRO	Clinical Research Organisation
CTU	Clinical Trials Unit
DA	Default Alternative
DN	Default Null
FA	Fixed Alternative
FDPP	Fixed Discounting Power Prior
FN	Fixed Null
FnHiM	Fun 'n healthy in Moreland
GAM	Generalised Additive Model

GCV	Generalised Cross Validation
GEE	Generalised Estimating Equation
GP	General Practitioner
HEIA	Health in Adolescents
HeLP	Healthy Lifestyle Programme
HMC	Hamiltonian Monte Carlo
HPDI	Highest Posterior Density Interval
ICC	Intracluster Correlation Coefficient
ICPP	Ibrahim Chen Power Prior
JAGS	Just Another Gibbs Sampler
LOESS line	Locally weighted smoothing line
MCID	Minimum Clinically Important Difference
MCMC	Markov Chain Monte Carlo
MSE	Mean Squared Error
NPP	Normalised Power Prior
PBPP	Partial Borrowing Power Prior
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QTN	Quintile Truncated Null
RCT	Randomised Controlled Trial
RWM	Random Walk Metropolis
SD	Standard Deviation
TA	Truncated Alternative
TTN	Tertile Truncated Null

Chapter 1

Introduction

The key concepts underpinning this thesis are introduced in this chapter. Specifically, the concept of cluster randomisation, along with its rationale, and the associated methodological implications are discussed. The project from which the exemplar data used throughout this thesis was obtained, the Health Lifestyles Programme, is introduced. This chapter also provides an introduction to Bayesian statistics and outlines some of the most commonly used Markov Chain Monte Carlo methods that make modern Bayesian inference possible. The literature on information borrowing methods in randomised controlled trials is reviewed.

***

1.1 Cluster Randomised Controlled Trials

THE Randomised Controlled Trial (RCT) is widely accepted as the "gold standard" of evidence for determining the efficacy or effectiveness of an experimental intervention. By randomly allocating participants to receive the intervention of interest, or to receive a control treatment, any observed between-group differences can subsequently be attributed to the intervention itself, ensuring balance across (measured or unmeasured) confounding variables, and so reliable, scientifically robust conclusions can be drawn.

Whilst the RCT is a well-established experimental design, more recent methodological developments have been made in the area of Cluster Randomised Controlled Trials (CRCTs), an increasingly common type of RCT in which randomisation occurs at a group ("cluster") level rather than at an individual level, but with outcomes still measured on individuals. Examples of such clusters include General Practitioner (GP) practices, schools or geographical regions, within which multiple individuals may provide outcome data.

The modern CRCT design is perhaps still considered to be a relatively new approach, yet the underpinning methodology is becoming increasingly well-established in the literature. Prior to the 1980s, the use of the CRCT design was sparse [Bland, 2004]. However, reported use of the design within the medical literature has been increasing substantially, from just one report per year in the 1960s, to seven in 1990, and over 120 in 2008 [Donner et al., 1990, Moberg and Kramer, 2015].

Cluster level randomisation adds an additional level of complexity (logistically, methodologically, and ethically) to study design, conduct, analysis, and reporting. As a result, careful consideration of the appropriateness of using the CRCT study design is always necessary. Despite the additional complexity, there is often a sound scientific rationale for adopting this study design. Eldridge and Kerry [Eldridge and Kerry, 2012] outlined the common justifications for adopting a cluster randomised trial design:

When the intervention is implemented at the cluster level: This occurs when the intervention is delivered to entire clusters, and individual-level implementation would not be possible. Examples include educational interventions targeted on an entire population, or a policy intervention affecting an entire community.

When there exists practical or ethical difficulties in implementing individual-level randomisation: Examples of where this is applicable may include large trials in low income countries, where it becomes more straightforward for fieldworkers to consistently deliver the same intervention.

When there is risk of contamination amongst those delivering the intervention: For example, if the intervention is a new form of advice to be delivered by GPs to patients, potentially contradictory to standard care, contamination may occur if the GP is expected to give out both forms of advice to different patients. In such a scenario, randomisation at the level of those delivering the intervention can avoid this potential contamination.

When there is risk of contamination between individuals within a cluster: This can occur when there is risk of the control group participants being partially exposed to the intervention (or vice versa) through, for example, interactions occurring amongst participants within clusters.

When there are logistical or cost benefits: For example, it would be cheaper to only equip clusters randomised to the intervention with expensive equipment, rather than all clusters.

When there is easy access to routinely collected cluster-level data: If interest lies in an outcome at the cluster-level, rather than at the individual-level, the cluster-level data may be routinely available without formal individual-level consent.

1.1.1 Statistical Analysis

Cluster randomisation has the potential to solve many practical and logistical issues encountered in trial design and delivery. However, it also introduces an additional level of statistical complexity which must be accounted for during the analysis of the trial data, as well as in the study design. When randomisation is undertaken at the cluster level, it is often the case that measurements within the same cluster are more correlated to one another than they are to measurements from other clusters. Failure to appropriately account for this correlation during statistical modelling can result in erroneously narrow confidence intervals, an inflated type I error and therefore an increased risk of reaching a spurious conclusion of efficacy or effectiveness. Typical analysis methods for handling this correlation structure include cluster level analysis; hierarchical modelling (also known as random effects or mixed effects models); and marginal models using Generalised Estimating Equations (GEEs). Such methods can be applied to different types of outcome data. However, the focus within this thesis is on CRCTs with continuous outcomes, and the analysis methods for such data are outlined in more detail below.

Cluster Level Analysis

The simplest method of accounting for clustering within statistical analysis is to model some aggregate measure for each cluster, rather than modelling data for each participant within a cluster. It is often also appropriate to introduce weighting to account for any differences between cluster sizes. A disadvantage of a cluster level analysis is that it is not straightforward to include adjustments for individual-level covariates, such as a baseline outcome measure, which are often recommended in order to improve statistical efficiency and improve the precision of estimates of treatment effects.

Hierarchical Models

Hierarchical models are often interchangeably referred to as random effects models, mixed effects models or multilevel models. Throughout this thesis, such models will be referred to as hierarchical models, whilst acknowledging the variation in terminology that exists within the literature.

Perhaps the most common approach to account for the correlated structure of data obtained from CRCTs is through the use of hierarchical models, where an additional random effect (residual) term is included in the statistical model to allow for, and capture, the mean effect of an individual being in a given cluster. Given a two-arm CRCT with a continuous outcome $Y_{i,j}$ for participant j in cluster i, a linear hierarchical model can be expressed as

$$Y_{i,j} = \beta + \theta x_{i,j} + \mu_i + \varepsilon_{i,j} \tag{1.1}$$

where

- β is the constant (intercept) term;
- θ is the treatment effect;
- $x_{i,j}$ is an indicator variable for allocation to the intervention arm, where $x_{i,j} = 1$ when participant j in cluster i is allocated to the intervention arm, and $x_{i,j} = 0$ otherwise;
- $\varepsilon_{i,j} \sim \mathrm{N}(0, \sigma^2)$ is the residual term for participant j in cluster i, and
- $\mu_i \sim \mathrm{N}(0, \sigma_c^2)$ is the random effect term, representing the mean effect on $Y_{i,j}$ of being in cluster i.

Generalised Estimating Equations

An alternative to using a hierarchical model detailed above to allow for the correlation structure amongst participants within the same cluster is to use a GEE [Liang and Zeger, 1986]. Again, given a two-arm CRCT, $Y_{i,j}$ is the continuous outcome for participant j in cluster i and is modelled using a linear relationship as

$$Y_{i,j} = \beta + \theta x_{i,j} + \varepsilon_{i,j} \tag{1.2}$$

As in the hierarchical model shown in Equation (1.1), β is the intercept term, θ is the treatment effect, and $x_{i,j}$ is the indicator variable for allocation to the intervention arm. However, the model shown in Equation (1.2) includes no cluster-level random effect term (μ_i in Equation (1.1)). As a result, the correlation present between the values of $Y_{i,j}$ within each cluster will similarly occur amongst the residuals, $\varepsilon_{i,j}$, within each cluster. GEEs are a method through which, instead of directly modelling the correlation structure as in a hierarchical model, it is treated as a nuisance and estimated from the residuals, and instead the mean response is modelled. The process of estimating the treatment effect is undertaken separately to the estimation of its precision. Different correlation structures can be assumed in the modelling. In the analysis of CRCTs, the usual correlation assumption is that of exchangeability, which assumes equal correlation between cluster members. In the absence of any additional evidence, this is an appropriate assumption [Eldridge and Kerry, 2012].

1.1.2 Design Considerations

The correlation structure introduced as a result of cluster level randomisation has implications not only for the choice of statistical analysis, but must also be considered during the study design phase, and particularly in the context of sample size determination.

Typically, an appropriate adjustment for clustering within a sample size calculation involves first calculating the required sample size assuming an individually randomised study design, before inflating this sample size according to some measure of the degree of clustering, known as the *Design Effect (Deff)*.

For an individually randomised two-arm trial with a continuous outcome, the total number of participants required per arm in order to detect a minimum treatment effect size of δ, at the two-sided α% significance level, with $(1-\beta)$% power, and denoting the standard deviation of the outcome as σ, is

$$N = 2\frac{(\Phi^{-1}(1-\frac{\alpha}{2}) + \Phi^{-1}(1-\beta))^2}{\delta^2}\sigma^2 \tag{1.3}$$

where $\Phi^{-1}(p)$ denotes the p^{th} quantile of the standard normal distribution.

The design effect, which is used to inflate the required sample size according to the expected degree of clustering in the data, can be expressed as

$$Deff = 1 + \rho(m-1) \tag{1.4}$$

where ρ denotes the Intracluster Correlation Coefficient (ICC), and m is the cluster size.

The (inflated) required sample size for a CRCT can be obtained either through increasing the cluster size, m, or the number of recruited clusters, and is often determined according to the practical or logistical constraints of the study. In scenarios where the number of participants per cluster, m, is fixed or constrained, it will likely be necessary to increase the number of recruited clusters, k, in order to reach the required sample size. In such scenarios, the required number of clusters per arm, k, can be obtained by calculating

$$k = N(1 + \rho(m-1))/m \tag{1.5}$$

where N is the number of participants per arm, before inflation to allow for clustering.

Alternatively, the number of clusters that can be feasibly included in a CRCT may be constrained. In such cases, the cluster size, m, must instead be increased in order to ensure a sufficiently large number of participants are included in the study to achieve a desired level of power. The required cluster size, m, can be calculated by rearranging Equation (1.5) so that

$$m = \frac{N(1-\rho)}{k - N\rho} \tag{1.6}$$

As in individually randomised trials, it is important to consider and mitigate against the risk of loss to follow-up by inflating the calculated sample size according to an estimate of the attrition rate. However, in CRCTs, it is often necessary to not only inflate for attrition at the individual-level, but also to account for the possibility of entire clusters dropping out.

Variability in Cluster Size

The formula for the design effect shown in Equation (1.4) assumes a fixed cluster size, m. In practice, this is rarely the case, with naturally occurring variability in cluster size often inevitable. Furthermore, variability in cluster sizes can result in a loss of power relative to equally sized clusters with the same total number of individuals, and the importance of appropriately accounting for this during study design and sample size estimation has been shown [Eldridge et al., 2006, Lauer et al., 2015].

When variability in cluster size is small, a simple solution is to substitute m in Equation (1.4) with $\bar{m}$, where $\bar{m}$ is the expected mean cluster size [Eldridge and Kerry, 2012].

In some cases, the size of each cluster may be known in advance. In such cases, the value of m in Equation (1.4) can be substituted with m_a, where

$$m_a = \frac{\sum m_i^2}{\sum m_i}$$

and m_i is the number of participants in cluster i [Donner et al., 1981].

However, it is not common to know the size of each cluster to be included in advance, particularly at the study design stage. It may be feasible, however, to estimate the mean and standard deviation of the expected cluster sizes through existing published research or other prior knowledge of the nature of the clusters. Denoting the mean and standard deviation of the cluster sizes as $\bar{m}$ and s_c respectively, the coefficient of variation of cluster sizes, cv, can be calculated as

$$cv = \frac{s_c}{\bar{m}} \tag{1.7}$$

A design effect that accounts for the estimated variability in cluster size (for subsequent use in the sample size calculation) [Eldridge et al., 2006] can then be calculated as

$$Deff = 1 + ((cv^2 + 1)\bar{m} - 1)\rho \tag{1.8}$$

1.1.3 The Intracluster Correlation Coefficient

The ICC for an outcome can be thought of as a measure of the degree of correlation for that outcome amongst participants within clusters, or as the proportion of variability for that outcome that is between clusters. In a CRCT, the variance components can be split into the variance *between* the clusters, and the variance *within* the clusters, which sum to give the *total* variance. The importance of the ICC in CRCT study design and sample size calculation is evident in the formulation of the design effect shown in Equation (1.4).

A common interpretation of the ICC is as the proportion of the overall variability that can be explained by the variability between clusters. For continuous data and assuming that the within-cluster variance, σ^2, is the same in each cluster, using the notation from Equation (1.1) (i.e. that σ denotes the within-cluster SD, and σ_c^2 denotes the between-cluster SD), the ICC can be expressed as

$$\rho = \frac{\sigma_c^2}{\sigma_c^2 + \sigma^2} \tag{1.9}$$

The Consolidated Standards of Reporting Trials (CONSORT) guidelines provide guidance for researchers in order to facilitate the complete and transparent reporting of clinical trials [Moher et al., 2010]. A number of extensions to these guidelines, specific to various non-standard trial methodologies, have also been developed, including an extension to CRCTs [Campbell et al., 2012]. As a result of the importance of the ICC in informing future study design, this extension recommends that the ICC is reported for all primary and secondary outcome measures in trial results publications. Therefore, published trial reports are a common source of ICC estimates.

Furthermore, a number of studies have been undertaken with the aim of collating ICCs from a range of settings relevant to a range of outcomes, and exploring patterns in their estimates. An example of such a study was undertaken by Ukoumunne et al. [Ukoumunne et al., 1999], who collated ICCs calculated using data from the Health Survey for England pertaining to a range of lifestyle, cardiovascular and other outcome measures for varying cluster types, including households, towns, postcode sectors and district and regional health authorities.

Eldridge et al. [Eldridge and Kerry, 2012] have discussed at length the potential sources of ICC estimates which can help to inform the assumptions made during sample size calculation and study design. Despite the increasing emphasis placed on the importance of reporting ICCs, it often remains challenging to identify relevant and appropriate ICC estimates to inform study design.

An alternative means of estimating an ICC for a sample size calculation is to first undertake a pilot or feasibility study, after which the data collected can be used to directly calculate the ICC. However, whilst a pilot or feasibility study may be useful for many reasons (for example to test the practical or logistical elements of running a larger scale study), it is not recommended that the results of such studies are used to calculate ICCs to inform sample size calculation because the typically small sample sizes in such studies often result in substantial imprecision. It is, however, acknowledged that estimates from pilot or feasibility studies could be used to contribute to a wider evidence base for ICC patterns [Eldridge et al., 2015].

It is recommended that, where possible, multiple sources of ICC estimates are sought when gathering evidence to justify the choice of ICC in a sample size calculation [Eldridge and Kerry, 2012, Eldridge et al., 2015]. If multiple relevant ICC estimates are obtained, a simple approach to aggregating may be to take an average or, as a conservative approach, a maximum. Furthermore, Turner et al. [Turner et al., 2004, Turner et al., 2005] proposed a Bayesian meta-analytic approach through which multiple ICC estimates can be combined, taking cognisance of study size and relevance and accounting for the overall uncertainty in the subsequent sample size or power calculations. Appropriate sensitivity analyses to explore the implications of the choice of ICC on expected power or sample size requirements should always be undertaken.

***

1.2 An Exemplar Dataset: The Healthy Lifestyles Programme Cluster Randomised Controlled Trial

In order to test the methodology developed as part of this PhD project, an exemplar dataset from a high quality CRCT was sought. Ethical approval was obtained from the Plymouth University Faculty of Health and Human Sciences Research Ethics Committee to utilise the high quality data already collected as part of the Healthy Lifestyles Programme (HeLP) CRCT and associated pilot study, which has provided two complementary, exemplar data sets for use within this project. A brief introduction to the HeLP study is given below.

Obesity in childhood is considered to be one of the most serious public health challenges of the 21st century [World Health Organisation, 2020], and can have serious health consequences in later life, including type 2 diabetes, increased risk of metabolic syndrome in youths and adults, and obesity in adulthood [Biro and Wien, 2010], as well as wider social and economic implications [Public Health England, 2015]. More than 1 in 5 children in England are now overweight or obese when they begin school, and almost 1 in 3 by the time they leave primary school [The NHS Information Centre, 2011].

The HeLP trial [Wyatt et al., 2013] was a pragmatic, superiority CRCT designed and successfully delivered to assess the effectiveness of a school-based intervention to prevent obesity in children. Participating schools ("clusters") were randomised to intervention or control groups, with baseline measures taken prior to randomisation. The intervention, implemented over three school terms, from the spring and summer term of Year 5 until the autumn term of Year 6, aimed to promote a healthy, balanced diet alongside a more active lifestyle through a combination of education, support and activity aimed at children, parents and teachers. The primary outcome was Body Mass Index Standard Deviation Score (BMI SDS) at 24 months post randomisation, the final time point in the study. Secondary outcomes included waist circumference, percentage body fat SDs, the proportion of children classified as overweight or obese at 24 months post baseline and objectively measured physical activity and food intake at 18 months.

The HeLP trial aimed to obtain primary outcome data from 762 pupils, in order to achieve 90% power, with a 5% two-sided significance level, to detect a between-group difference in the primary outcome of BMI SDS at 24 months of 0.25 units. The sample size calculation assumed a standard deviation (SD) of 1.3 units, an ICC of 0.02, a within-child correlation between baseline and follow-up assessments of 0.8, and an average cluster size of 35, with a coefficient of variation of cluster sizes of 0.5. Allowing for loss to follow-up of up to 20%, the study aimed to recruit 28 schools and a minimum of 952 children.

Prior to the design and delivery of the HeLP study, an external pilot study was undertaken, recruiting and randomising (1:1) four schools, comprising a total of 202 children, to the HeLP intervention, or to control [Lloyd et al., 2012]. Whilst pilot studies are not designed or undertaken in order to address questions surrounding intervention effectiveness or efficacy, they can provide preliminary signals of effectiveness which may help to justify the conduct of a subsequent fully-powered, definitive trial, which was the case in the HeLP pilot study.

The definitive study exceeded recruitment targets, randomising 32 schools, with a total of 1324 pupils enrolled into the trial. No schools dropped out of the study, and 1244 pupils provided primary outcome data at baseline and 24 month follow-up and were therefore included in the primary analysis of the primary outcome [Lloyd et al., 2018]. No statistically significant difference was found in BMI SDS at 24 months between the treatment groups. Sensitivity analyses produced similar conclusions. Furthermore, there was no significant difference in treatment groups in terms of waist circumference, percentage body fat SDs, or physical activity levels. Despite this, there was evidence to suggest a significant reduction in the consumption of energy dense snacks, and fewer negative food markers, in the intervention group compared to the control group. However, as secondary outcomes, these results should be interpreted as exploratory.

***

1.3 Bayesian Statistics

1.3.1 An Introduction

In Bayesian statistics, one first seeks to specify a *prior* belief, before examining forthcoming evidence (i.e. data) and updating that *prior* belief accordingly to arrive at a *posterior* belief. Conversely, a frequentist statistician would interpret a single estimate as if it were one of many (hypothetical) repeated experiments. A Bayesian statistician would treat sample data as fixed and would describe unknown parameters probabilistically, whereas a frequentist would treat data as a random sample, and their parameters as fixed (but unknown) across each random sample.

Bayes' theorem, named after the 18th century mathematician and Reverend Thomas Bayes, is a mathematical statement of conditional probability. Given two events, A and B, Bayes' theorem states that the probability of event A occurring, given the occurrence of another event B, is

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{1.10}$$

In the context of Bayesian inference, Equation (1.10) is simply applied to data through a likelihood, incorporating parameters to be estimated. So, for some parameter(s) θ, and data D, Bayesian inference is concerned with obtaining the *posterior distribution* of θ, given the data. Mathematically, this can be expressed as

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)} \tag{1.11}$$

where $P(D|\theta)$ is the likelihood of the data given the parameters and $P(\theta)$ is the prior distribution of θ, which is a representation of the current state of knowledge or belief about θ *before* contemplating the evidence contained within the data, D. $P(D)$ is the marginal distribution of the data, often known as the normalising constant, and its purpose is to standardise the posterior distribution to ensure that it integrates to 1, as is required for a proper distribution. $P(D)$ is the integral of the product of the likelihood and the prior with respect to θ, evaluated over the range of possible values of θ. Formally

$$P(D) = \int_{\Theta} P(D|\theta) \times P(\theta) d\theta \tag{1.12}$$

Because the parameters have been integrated out, the value of $P(D)$ does not depend on θ, and so the following adaptation of Bayes' rule in Equation (1.11) is often invoked:

$$P(\theta|D) \propto P(D|\theta) \times P(\theta) \tag{1.13}$$

As the likelihood and the prior distribution(s) are often relatively easy to specify, the posterior distribution can be obtained up to a constant of proportionality. That is to say

$$P(\theta|D) = K \times P(D|\theta) \times P(\theta) \tag{1.14}$$

where $K = 1/P(D)$ is some value to be computed. In practice, K is often intractable or very difficult to calculate. For this reason, the practical utility of the Bayesian approach to statistical inference was limited for many years, as its use was restricted to special cases of likelihood and prior specification which resulted in mathematically tractable posterior distributions from which posterior samples can be directly obtained. Fortunately, modern Markov Chain Monte Carlo (MCMC) methods can be applied directly to Equation (1.13), making it possible to sample from posterior distributions without the need for explicit computation of K. Some of the most common MCMC methods are outlined in §1.3.2 below.

The "Bayesian versus frequentist" debate has been ongoing for many years, taking cognisance of a range of practical, theoretical and philosophical viewpoints. Whilst contribution to this wider debate is beyond the scope of this thesis, it is useful to highlight in particular some of the potential benefits of adopting a Bayesian approach to statistical inference. Firstly, whilst incorporation of prior information can be subject to criticism as a result of its potential to introduce bias through subjective beliefs or opinions, when properly specified it can provide an intuitive and robust mechanism for incorporating existing evidence into an analysis. Philosophically, it is difficult to argue that Bayesian reasoning does not mirror the broader scientific process, where evidence is accrued over time and best practice and scientific opinion regularly revised according to new and emerging evidence. Secondly, in Bayesian inference, the aim is to obtain (samples from) a posterior distribution. In contrast to a frequentist analysis, where interest lies in obtaining a point estimate and associated confidence interval, obtaining a full posterior distribution for a parameter of interest allows for more flexible and thorough visualisation, exploration and interpretation of statistical results. These results can be communicated in the context of probability in the usual sense, as opposed to within the somewhat unintuitive frequentist context, pertaining to a large number of hypothetical (but unobserved) repetitions of an experiment.

1.3.2 Markov Chain Monte Carlo Methods

The computational challenge of deriving the value of K in Equation (1.14) meant that for many years, Bayesian inference was predominantly a theoretical rather than an applied endeavour. 1953 saw the publication, in the field of physics, of the first research paper on MCMC methods [Metropolis et al., 1953], with these methods later generalised by Hastings in 1970 [Hastings, 1970]. However, it was not until the late 1980s that MCMC methods began to truly influence mainstream applied statistics, with the development of the Gibbs Sampling approach and the resulting paper by Gelfand and Smith [Gelfand and Smith, 1990] providing a genuinely practical and accessible way of obtaining samples from Bayesian posterior distributions, circumventing the need for explicit computation of the normalising constant. This development, described as an "Epiphany in the World of Statistics" [Robert and Casella, 2011], was the moment after which Bayesian statistics became a practical and viable method of statistical inference. Throughout the 1990s, methodological research into the use of MCMC methods in Bayesian inference accelerated, including (but certainly not limited to) application to linear [Wang et al., 1993, Wang et al., 1994] and generalised linear mixed effects models [Zeger and Karim, 1991], changepoint analysis [Carlin et al., 1992] and variable selection methods in regression [George and McCulloch, 1993]. Further encouraging the increased uptake of Bayesian inference using MCMC was the development of BUGS (Bayesian inference Using Gibbs Sampling) software [Lunn et al., 2000], which was first presented at a conference in Valencia in 1991. A more detailed description of the history of the development of MCMC methods is provided by Roberts and Casella [Robert and Casella, 2011].

Today, the use of MCMC methods in Bayesian statistics is a mature, yet active and evolving field. A range of methods are now available and application is relatively straightforward for applied researchers through various software options, including WinBUGS [Lunn et al., 2000], JAGS (Just Another Gibbs Sampler) [Plummer, 2003] and Stan [Carpenter et al., 2017], all of which can be implemented through the statistical programming language R [R Core Team, 2019].

The Gibbs Sampler

The Gibbs sampler works by sampling iteratively from the marginal conditional distribution of each parameter of interest sequentially, updating the current value at each iteration. It is particularly useful in cases when there is conditional conjugacy. That is to say, when the conditional distribution of each parameter, given the other parameters, has a standard statistical distribution from which samples can be directly drawn.

Assume some data, denoted $\mathbf{x}$, and some parameters of interest $\theta = (\theta_1, \theta_2, \theta_3, \ldots, \theta_r)$.

Then, by Bayes' Theorem, the multivariate posterior distribution of the parameters given the data can be written as

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{\pi(\mathbf{x}|\theta)\pi(\theta)}{\pi(\mathbf{x})}\\ &\propto \pi(\mathbf{x}|\theta)\pi(\theta)\end{aligned}$$

Then the conditional distribution of each θ_j, $\pi(\theta_j|\theta_{-j},\mathbf{x})$, where θ_{-j} denotes the vector of parameters θ excluding θ_j, can be expressed by focusing only on terms in the overall joint posterior distribution that involve θ_j. Provided each of the marginal posteriors can be expressed as a proper distribution, Gibbs sampling can be applied as follows:

1. Initialise the algorithm with initial values, $\theta^{(0)}$
2. For $i = 1,\ldots,N$, obtain a sample from $\theta^{(i)}$:
   - 2.1. Sample $\theta_1^{(i)}$ from $\pi(\theta_1^{(i-1)}|\theta_2^{(i-1)},\ldots,\theta_r^{(i-1)},\mathbf{x})$
   - 2.2. Sample $\theta_2^{(i)}$ from $\pi(\theta_2^{(i-1)}|\theta_1^{(i)},\theta_3^{(i-1)},\ldots,\theta_r^{(i-1)},\mathbf{x})$
   - 2.3. ...
   - 2.4. Sample $\theta_r^{(i)}$ from $\pi(\theta_r^{(i-1)}|\theta_1^{(i)},\ldots,\theta_{r-1}^{(i)},\mathbf{x})$
3. Discard the first z iterations. Provided a sufficiently large z and N, the remaining samples, $\theta^{(z+1)},\ldots,\theta^{(N)}$, can be treated as realisations from the target posterior distribution $\pi(\theta|\mathbf{x})$.

The Metropolis Hastings Algorithm

The Gibbs sampling algorithm works well when the conditional distribution of each parameter of interest can be expressed as a proper statistical distribution. In certain circumstances, the correct choice of prior distribution can ensure that is the case. However, it is often either not possible to ensure proper conditional distributions, or undesirable to specify prior distributions on the basis of mathematical or computational convenience alone. In such cases, the Metropolis-Hastings algorithm provides a framework to obtain samples from a joint posterior distribution even when the conditional distribution of the parameters are not proper statistical distributions and therefore cannot be sampled from directly.

The Metropolis-Hastings algorithm differs from the Gibbs Sampler in that, rather than sampling each parameter directly from the tractable distribution from which its marginal posterior belongs, a candidate updated value is proposed based on the current value, and accepted with some probability. Formally, the algorithm can be constructed as follows.

1. Initialise the algorithm with initial values, $\theta^{(0)}$, and define some proposal distribution, q.

2. For $i = 1, \ldots, N$, obtain a sample for $\theta^{(i)}$

2.1. Obtain a sample for $\theta_1^{(i)}$:

2.1.1. Propose a candidate value, $\xi_1^\star$, for $\theta_1^{(i)}$ by sampling from $q(\cdot|\theta_1^{(i-1)})$

2.1.2. Calculate the acceptance probability for the proposed value,

$$\alpha(\xi_1^\star, \theta_1^{(i-1)}) = \min\left(1, \frac{\pi(\xi_1^\star|\theta_2^{(i-1)}, \ldots, \theta_r^{(i-1)}, \mathbf{x}) \cdot q(\theta_1^{(i-1)}|\xi_1^\star)}{\pi(\theta_1^{(i-1)}|\theta_2^{(i-1)}, \ldots, \theta_r^{(i-1)}, \mathbf{x}) \cdot q(\xi_1^\star|\theta_1^{(i-1)})}\right)$$

2.1.3. Accept the candidate value $\xi_1^\star$ with probability $\alpha(\xi_1^\star, \theta_1^{(i-1)})$ and set $\theta_1^{(i)} = \xi_1^\star$. Otherwise, set $\theta_1^{(i)} = \theta_1^{(i-1)}$

2.2. Obtain a sample for $\theta_2^{(i)}$:

2.2.1. Propose a candidate value, $\xi_2^\star$, for $\theta_2^{(i)}$ by sampling from $q(\cdot|\theta_2^{(i-1)})$

2.2.2. Calculate the acceptance probability for the proposed value,

$$\alpha(\xi_2^\star, \theta_2^{(i-1)}) = \min\left(1, \frac{\pi(\xi_2^\star|\theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_r^{(i-1)}, \mathbf{x}) \cdot q(\theta_2^{(i-1)}|\xi_2^\star)}{\pi(\theta_2^{(i-1)}|\theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_r^{(i-1)}, \mathbf{x}) \cdot q(\xi_2^\star|\theta_2^{(i-1)})}\right)$$

2.2.3. Accept the candidate value $\xi_2^\star$ with probability $\alpha(\xi_2^\star, \theta_2^{(i-1)})$ and set $\theta_2^{(i)} = \xi_2^\star$. Otherwise, set $\theta_2^{(i)} = \theta_2^{(i-1)}$

2.3. ...

2.4. Obtain a sample for $\theta_r^{(i)}$:

2.4.1. Propose a candidate value, $\xi_r^\star$, for $\theta_r^{(i)}$ by sampling from $q(\cdot|\theta_r^{(i-1)})$

2.4.2. Calculate the acceptance probability for the proposed value,

$$\alpha(\theta_r^{(i)}, \theta_r^{(i-1)}) = \min\left(1, \frac{\pi(\xi_r^\star|\theta_1^{(i)}, \ldots, \theta_{r-1}^{(i)}, \mathbf{x}) \cdot q(\theta_r^{(i-1)}|\xi_r^\star)}{\pi(\theta_r^{(i-1)}|\theta_1^{(i)}, \ldots, \theta_{r-1}^{(i)}, \mathbf{x}) \cdot q(\xi_r^{(i)}|\theta_r^{(i-1)})}\right)$$

2.4.3. Accept the candidate value $\xi_r^\star$ with probability $\alpha(\xi_r^\star, \theta_r^{(i-1)})$ and set $\theta_r^{(i)} = \xi_r^\star$. Otherwise, set $\theta_r^{(i)} = \theta_r^{(i-1)}$

3. Discard the first z iterations. Provided a sufficiently large z and N, the remaining samples, $\theta^{(z+1)}, \ldots, \theta^{(N)}$, can be treated as realisations from the target posterior distribution $\pi(\theta|\mathbf{x})$.

Furthermore, the Gibbs Sampler outlined above is a special case of the Metropolis-Hastings algorithm in which the acceptance probability is equal to 1.

The Random Walk Metropolis

The Random Walk Metropolis (RWM) is a special case of the Metropolis Hastings algorithm in which the proposal distribution used to generate the new candidate value is centred on the current value. Letting θ denote the current state of the chain, and letting $\xi^\star$ denote the candidate value,

$$\xi^\star = \theta + \varepsilon$$

where ε is some symmetric random variable with $\mathbb{E}(\varepsilon) = 0$. Under such conditions, the acceptance probability simplifies to

$$\alpha(\xi^\star, \theta) = \min\left(\frac{\pi(\xi^\star)}{\pi(\theta)}, 1\right)$$

One of the key advantages of the RWM is that it is simple to implement in practice. All that is required is specification of ε. A common choice is a normal distribution, so that $\xi^\star \sim \mathrm{N}(\theta, \sigma)$. The choice of σ is an important issue in RWM; too small a value can result in large acceptance rates, where the chain moves often, but in very small increments, resulting in slow exploration of the posterior distribution. Conversely, a large value of σ can result in a low acceptance rate, where the chain moves infrequently, but in large increments, again resulting in inefficient exploration of the posterior distribution. Much research has been undertaken into the choice of σ, including on optimal acceptance probabilities for univariate [Roberts and Rosenthal, 2001] and multivariate [Roberts et al., 1997] updating procedures, and adaptive MCMC methods which update the value of σ as the chain progresses [Andrieu and Thoms, 2008], although further discussion of these topics is beyond the scope of this thesis.

Hamiltonian Monte Carlo

Whilst the RWM often works well in practice, the random walk element (i.e. proposing the next value, randomly, based on the current value) of the procedure does suffer with inherent inefficiencies which can often make thorough exploration of the target posterior distribution a long and computationally laborious task. Hamiltonian Monte Carlo (HMC) methods, also know as Hybrid Monte Carlo methods, attempt to address this inefficiency by introducing an additional “momentum” term to facilitate more rapid exploration of posterior distributions. A discussion within the context of Bayesian inference is provided by Gelman et al. [Gelman et al., 2013]; and more conceptual introductions with stronger ties to the physics literature from which the methods are motivated (although still with a statistical reader in mind) are provided by Betancourt [Betancourt, 2018] and Neal [Neal, 2012].

Suppose an r-dimensional vector of parameters of interest, θ. Then HMC involves introducing an additional, associated r-dimensional vector of momentum variables, ϕ. θ and ϕ are each updated using a Metropolis-Hastings algorithm, with ϕ used to inform the proposal distribution for θ.

Suppose that, as previously, interest lies in obtaining samples from the posterior distribution of the parameters of interest, θ, given some data, $\mathbf{x}$, denoted $\pi(\theta|\mathbf{x})$. After introducing the momentum variables, ϕ, a new joint distribution of interest can be defined, $\pi(\theta,\phi|\mathbf{x}) = \pi(\phi)\pi(\theta|\mathbf{x})$, although ϕ are simply auxiliary variables and so obtaining samples from θ remains the primary aim.

What differentiates HMC methods from the Metropolis-Hastings methods introduced above is the way in which new candidate values for each of the components of θ are proposed. Specifically, in HMC, Hamiltonian dynamics are utilised, and the leapfrog method used in order to approximate the Hamiltonian equations. In order to implement the leapfrog method, the gradient of the log of the posterior density must be calculated, either numerically or analytically. Specifically,

$$\frac{d\log\pi(\theta|\mathbf{x})}{d\theta} = \left(\frac{\partial\log\pi(\theta|\mathbf{x})}{\partial\theta_1},\dots,\frac{\partial\log\pi(\theta|\mathbf{x})}{\partial\theta_r}\right)$$

Furthermore, for the distribution of the momentum term, $\pi(\phi)$, an independent multivariate Normal distribution is typically assumed, with mean $\mathbf{0}$ and diagonal covariance matrix $\mathbf{M}$, so that $\phi_i \sim \mathrm{N}(0,m_i)$. Then, for each iteration, i, within the wider MCMC procedure, and after specifying some number of leapfrog steps, L and small ε, such that $\varepsilon L = 1$, the leapfrog procedure can proceed as follows:

1. Make a “half-step” update of the momentum vector, ϕ, using

$$\phi^{(i-1+\varepsilon/2)} = \phi^{(i-1)} + \frac{1}{2}\varepsilon\frac{d\log\pi(\theta^{(i-1)}|\mathbf{x})}{d\theta^{(i-1)}}$$

2. Make a “full-step” update of the parameter, θ, using

$$\theta^{(i-1+\varepsilon)} = \theta^{(i-1)} + \varepsilon\frac{\phi^{(i-1+\varepsilon/2)}}{\mathbf{M}}$$

3. Make a second “half-step” update of the momentum vector, ϕ, using

$$\phi^{(i-1+\varepsilon)} = \phi^{(i-1+\varepsilon/2)} + \frac{1}{2}\varepsilon\frac{d\log\pi(\theta^{(i-1+\varepsilon)}|\mathbf{x})}{d\theta^{(i-1+\varepsilon)}}$$

After L steps, the leapfrog procedure provides an updated set of values for the momentum term, $\phi^{(i)}$ and the parameters of interest, $\theta^{(i)}$.

The overall HMC algorithm can then proceed as follows:

1. Specify the number of leapfrog steps, L, step size, ε and diagonal covariance matrix, $\mathbf{M}$, for the momentum vector ϕ. Initialise the algorithm with initial values $\theta^{(0)}$.

2. For i in $1,\ldots,N$, generate a set of values of the momentum vector, $\phi^{(i-1)} \sim \mathrm{N}(0,\mathbf{M})$. Generate a sample of parameters, $\theta^{(i)}$ as follows:

   2.1. Use the leapfrog procedure to propose candidate values, $\phi^{\star}$ and $\theta^{\star}$, the values stored after L leapfrog steps.

   2.2. Calculate the acceptance probability for the candidate value,

   $$\alpha(\theta^{(i-1)},\theta^{\star}) = \min\left(1, \frac{\pi(\theta^{\star}|\mathbf{x})\pi(\phi^{\star})}{\pi(\theta^{(i-1)}|\mathbf{x})\pi(\phi^{(i-1)})}\right)$$

   2.3. Accept the candidate value θ^{*} with probability $\alpha(\theta^{(i-1)},\theta^{\star})$ and set $\theta^{(i)} = \theta^{\star}$.

3. Discard the first k iterations. Provided a sufficiently large k and N, the remaining samples, $\theta^{(k+1)},\ldots,\theta^{(N)}$, can be treated as realisations from the target posterior distribution $\pi(\theta|\mathbf{x})$.

The use of HMC methods result in substantial efficiency gains compared to random-walk based MCMC methods. HMC methods can now be implemented through the probabilistic programming language Stan, which allows both the specification of complex, bespoke statistical models, as well as more standard regression and multilevel models which can be fitted with ease using the `R` packages `brms` [Bürkner, 2018] and `rstanarm` [Goodrich et al., 2020]. HMC methods implemented using Stan are used to perform the majority of the Bayesian inference contained within this thesis.

***

1.4 Bayesian Statistics in Randomised Controlled Trials

Although most traditional RCTs are designed and analysed within the frequentist paradigm, the use of Bayesian methods as an alternative approach has been widely discussed in the literature. The use of such methods can both provide opportunities to design and conduct more efficient trials and produce more interpretable results, but can also present a number of challenges. A thorough overview of Bayesian methods in clinical trial and healthcare evaluation is provided by Spiegelhalter, Abrams and Myles [Spiegelhalter et al., 2004], although much methodological work in this field has been undertaken since the publication of this book.

The use of Bayesian methods in trial design and analysis, by its nature, can facilitate the incorporation of existing evidence or belief through the specification of a prior distribution. When these prior distributions are empirically justified based on scientific evidence, informative prior distributions tend to be less controversial than those which are simply reflective of prior belief or opinion. The inclusion of well-justified, informative prior information has the potential to improve a statistical analysis by generating more robust, thoroughly informed conclusions reached in a more efficient manner.

Bayesian methods can also be used within the context of trial analysis to facilitate more naturally interpretable analytical results. For example, as explained by Bland and Altman [Bland and Altman, 1998], a Bayesian analysis would allow one to conclude that there is a 95% probability that the true value lies within the interval, given the observed data. Similarly, in rejecting a null hypothesis, a Bayesian statistician could state that the probability that the null hypothesis is true is less than 5%. A frequentist statistician, on the other hand, would conclude that, if the experiment were to be repeated many times, 95% of the intervals would contain the true value. In rejecting the null hypothesis, a frequentist statistician would conclude a probability of less than 5% of observing a value at least as extreme as the one observed in the data, given the null hypothesis is true. The Bayesian interpretation of such results is arguably more intuitive and easily interpreted by non-statistical or clinical audiences in comparison to the frequentist interpretation.

In the ongoing pursuit for more efficient trial designs that are able to answer multiple research questions more rapidly and in a less resource-intensive manner, Bayesian methods are being increasingly employed in the field of adaptive trial designs, which allow for pre-specified interim examinations of accruing trial data to inform adaptations in trial design whilst the study is ongoing. Bayesian adaptive designs have been considered in the context of constructing early stopping rules for intervention effectiveness or futility [Ryan et al., 2019], response-adaptive randomisation [Brown et al., 2016] and early phase dose-finding studies [Wheeler et al., 2019].

The use of Bayesian methods in RCTs is widely discussed in the literature and their application is becoming increasingly common. However, a brief scoping review suggested that, whilst there have been some methodological developments in the field of Bayesian statistics within the context of CRCTs, application is rare. As a result, a methodological systematic review seeking to identify both practical implementation of Bayesian methods within CRCTs, and methodological developments in the field is presented in Chapter 2.

***

1.5 Specification of Prior Distributions

An important element of a Bayesian analysis is the specification of prior distributions, required for each parameter in the Bayesian model. Often, prior distributions are chosen to be non-informative, reflecting an absence of information about a parameter of interest before analysis of the data. Results of Bayesian analyses with non-informative prior distributions are usually similar to those from a corresponding frequentist analysis. Alternatively, informative prior distributions can be specified, which express the level of current evidence or belief about a parameter of interest, such as a treatment effect, which in turn has the potential to add value to a Bayesian statistical analysis. However, informative prior distributions are often difficult to specify and justify in practice. One approach may be to attempt to elicit expert opinion from individuals or groups of individuals, although this can be subject to criticism due to the potential to introduce a degree of subjectivity. An alternative approach is to attempt to utilise information from existing data to construct informative prior distributions. Regardless of the approach taken, expressing such evidence, either from expert opinion or from existing data, as a statistical (prior) distribution remains challenging. One approach that seeks to address this challenge in the context of using existing data to construct informative prior distributions is the power prior, which is outlined in more detail in §1.6 below.

***

1.6 Power Priors

1.6.1 An Introduction

An important and often controversial topic within the context of Bayesian inference is in the elicitation of informative prior information. The power prior [Ibrahim and Chen, 2000] is a useful class of informative prior distributions which can provide a systematic, data driven framework to facilitate the direct incorporation of historical data within a Bayesian analysis. A thorough overview of the theoretical properties of the power prior, as well as a range of applications, is provided by Ibrahim and Chen [Ibrahim et al., 2015].

The basic formulation of the power prior involves parametrisation of the prior distribution according to the likelihood of the historical data, and first using this to update some initial (usually uninformative) prior specification. The likelihood for this historical data is then discounted according to some discounting factor, a_0. Formally, given a historical dataset, D_0, and a set of parameters θ, where $L(\theta|D_0)$ denotes the likelihood of the historical data, and $\pi_0(\theta)$ represents an initial (usually non-informative) prior distribution, the power prior can be expressed as

$$\begin{aligned}\pi(\theta|D_0,a_0) &= \frac{L(\theta|D_0)^{a_0}\pi_0(\theta)}{\int_\Theta L(\theta|D_0)^{a_0}\pi_0(\theta)d\theta} \\ &\propto L(\theta|D_0)^{a_0}\pi_0(\theta)\end{aligned} \tag{1.15}$$

Typically a_0 will be constrained such that $a_0 \in [0,1]$, with $a_0 = 1$ fully incorporating the historical data, $a_0 = 0$ incorporating none of the historical data and values of a_0 between 0 and 1 representing partial discounting of the historical evidence. The formulation in Equation (1.15) assumes that a_0 is some fixed constant to be elicited and specified in advance by expert opinion or existing evidence, and is therefore hereafter referred to as the fixed discount power prior (FDPP).

However, it is often not straightforward to specify a well-justified, fixed value for a_0. As an alternative, a fully Bayesian approach can be adopted whereby a_0 is treated as an additional parameter, assigned a prior distribution and estimated as part of the inference process. Such an approach also has the advantage of avoiding the introduction of subjectivity through the need to specify a fixed value of a_0. For this approach, the following power prior formulation has been proposed [Ibrahim and Chen, 2000], which is hereafter referred to as the Ibrahim-Chen power prior (ICPP):

$$\pi(\theta,a_0|D_0) \propto L(\theta|D_0)^{a_0}\pi_0(\theta)\pi_0(a_0) \tag{1.16}$$

However, there are problems with the ICPP approach shown in Equation (1.16), as have been highlighted previously [Duan et al., 2006, Neuenschwander et al., 2009]. Firstly, once a_0 is introduced within the power prior formulation as a parameter, the normalising constant (i.e. the denominator of Equation (1.15)) now depends on a_0 (which is no longer fixed) and therefore can no longer be ignored through proportionality. Secondly, this formulation violates the Likelihood Principle as explained by Duan et al. [Duan et al., 2006]. This formulation tends to result in near-complete discounting of the historical data (i.e. values of a_0 close to 0), as has been noted in the literature on a number of occasions [Duan et al., 2006, Neuenschwander et al., 2009, Neelon and

O'Malley, 2010], and in fact near-complete discounting even occurs when the historical and current data are the same (i.e. $D = D_0$).

As a result, a modification to the ICPP has been proposed by both Duan et al. [Duan et al., 2006] and Neuenschwander et al. [Neuenschwander et al., 2009] in order to account for the missing normalising constant, which is hereafter referred to as the normalised power prior (NPP):

$$\pi(\theta, a_0|D_0) = C(a_0)L(\theta|D_0)^{a_0}\pi_0(\theta)\pi_0(a_0) \tag{1.17}$$

where

$$C(a_0) = \frac{1}{\int_\Theta L(\theta|D_0, a_0)^{a_0}\pi_0(\theta)d\theta} \tag{1.18}$$

In cases where $L(\theta|D_0)$ and $\pi_0(\theta)$ are conjugate, meaning that their product can be expressed as a proper distribution, $C(a_0)$ can be obtained mathematically as the kernel of the density function of the posterior distribution. An example of such a case is where the outcome of interest is binomial with probability of success given by θ, with a beta prior distribution assigned to θ. However, for more complex modelling approaches, including hierarchical models such as those typically used to analyse CRCT data, $C(a_0)$ is typically formed by a mathematically intractable, high-dimensional integral and therefore must be calculated using numerical approximation methods, which are discussed in more detail in §3.3.

1.6.2 Using Historical Data in Randomised Controlled Trials

The concept of information borrowing from historical evidence or preceding trials to inform analysis of current trial data is not new. An early example of such a concept is proposed by Pocock [Pocock, 1976], who discussed incorporating historical controls, alongside randomised controls, into a clinical trial analysis in order to improve the estimation of the treatment effect. Pocock outlined six conditions which should be satisfied in order to justify the suitability of a historical group for incorporation into a subsequent trial analysis: (i) the historical group must have received a precisely defined standard treatment which should be the same as the treatment received by the randomised control group; (ii) the group must have been part of a study with the same patient inclusion criteria; (iii) the methods of treatment evaluation must be the same; (iv) the distributions of important patient characteristics in the group should be comparable with those in the new trial; (v) the previous study must have been performed in the same organisation with largely the same clinical investigators and (vi) there must be no other indications leading one to expect differing results between the randomised control participants and the historical controls. Pocock proposed a Bayesian method of statistical analysis to incorporate historical controls within a trial analysis, involving specifying a Normal prior distribution for the control group centered upon the historical control group mean with variance specified according to the variability in the historical data and an additional (subjectively defined) variance term. Pocock also explored the implications of incorporating historical controls on study design, but focused on the optimum choice of allocation ratio given a predetermined overall sample size, rather than on the determination of the overall sample size itself.

Neuenschwander et al. [Neuenschwander et al., 2010] proposed a meta-analytic predictive approach (under both the Bayesian and frequentist frameworks) to incorporate evidence from historical controls from earlier studies whilst accounting for heterogeneity between studies, alongside methods for calculation of a "prior effective sample size" which quantifies the amount of information incorporated from the historical data in relation to the actual sample size (i.e. the number of control participants in the historical trials). This methodology was extended by Schmidli et al. [Schmidli et al., 2014], who proposed a robust meta-analytic predictive prior to account for potential conflict between the historical trials and the current trial. These methods can be implemented with relative ease using the R package `RBesT` (R Bayesian Evidence Synthesis Tools) [Weber et al., 2019]. Hobbs et al. [Hobbs et al., 2012] proposed a novel method of prior specification (the *commensurate prior*) for generalised linear models, where the parameters for the current and historical datasets are distinct, but with the prior distribution for the parameter in the current dataset informed by, and centered upon, the corresponding parameter from the historical dataset. Zheng and Wason [Zheng and Wason, 2022] have also considered the idea of information borrowing in the context of basket trials, proposing methodology for partially pooling information from across patient subgroups using the Hellinger Distance and a "spike and slab" prior [Mitchell and Beauchamp, 1988] in order to quantify the degree of borrowing.

It is natural to consider the problem of the incorporation of external evidence within the Bayesian paradigm, and as a result, the majority of the methodology described here is framed as such.

Using Power Priors to Incorporate Historical Information within Randomised Controlled Trials

The use of power priors within the context of RCTs has been discussed fairly widely in the literature. Hobbs et al. [Hobbs et al., 2011] proposed an adaptation of the power prior for use in RCTs, in which the prior distribution of the discounting factor, a_0, is parametrised directly by a measure of the commensurability between the current and historical datasets, termed the *location commensurate power prior*. De Santis [De Santis, 2006] discussed the use of power priors within RCTs, presenting a geometric prior distribution approach involving splitting the historical data into training samples, and describing the similarities to the power prior approach in the context of the choice of discounting factor. Gravestock and Held [Gravestock and Held, 2017], in addressing criticisms that the automatic calibration of a_0 in the NPP method does not appropriately reflect differences between historical and current datasets [Neelon and O'Malley, 2010], proposed an empirical Bayes approach for eliciting the value of the discounting factor a_0. Golchi [Golchi, 2020] proposed a novel method whereby the discounting of the historical controls is undertaken at the individual level rather than at the study level, by using the Mahalanobis distance between each participant and the posterior predictive distribution as the means by which to quantify the dissimilarity between each historical individual and the current trial data.

Van Rosmalen et al. [van Rosmalen et al., 2018] compared various methods of incorporating historical data, including a test-then-pool approach, the FDPP with $a_0 = 0.5$, the NPP, the meta-analytic and robust meta-analytic approaches and Pocock's approach. Comparison was firstly through an exemplar dataset in the area of oncology, where the outcome was overall survival and a number of historical datasets with comparable control groups were available. A simulation study based on a survival outcome was also presented. The authors concluded that the meta-analytic predictive approach performed best, resulting in consistently increased power and precision alongside well-controlled type I error. They also noted that the NPP approach performed well when the heterogeneity between trials was low, but experienced inflated type I error when a higher degree of heterogeneity was introduced. Similarly, Viele et al. [Viele et al., 2014] explored and compared various information borrowing methods to incorporate historical controls in the context of RCTs with a binary outcome, including a test-then-pool approach, the FDPP with $a_0 = 0.4$ and hierarchical modelling, noting the potential of such methods to increase power whilst controlling the inflation of type I error.

The incorporation of historical control data via the power prior approach has also been considered in the context of study design and sample size calculation. Hees and Kaiser [Hees and Kieser, 2017] presented a method of incorporating historical data into a blinded sample size recalculation in a trial with a binary outcome, demonstrating an increase in power and therefore a reduction in required sample size, provided the historical and current data are not conflicting. Psioda and Ibrahim [Psioda and Ibrahim, 2019] explored the use of the FDPP in sample size determination where historical data are used to inform the treatment effect. Power priors have also been considered in the design of non-inferiority trials [Chen et al., 2011], sequential meta-analytic designs [Chen et al., 2014a, Ibrahim et al., 2012] and trials with recurrent events data [Chen et al., 2014b].

To date, all of the research on the incorporation of historical data in to the analysis of an RCT has focused on individually randomised trials rather than CRCTs. The only exception to this is a recent PhD thesis, in which power prior methodology has been considered by Xiao [Xiao, 2017] in the context of CRCTs. Xiao proposed a novel method for incorporating historical, non-clustered data in to the analysis of CRCT data with a binary outcome. Two approaches for the elicitation of fixed values of a_0 were proposed, according to the symmetric and asymmetric Kullback-Liebler divergence measures quantifying the distance between the posterior distributions of the treatment effect according to the historical and current datasets when analysed separately. The proposed methods were extended to multi-arm cluster randomised trials and generalised to outcomes from the exponential family of distributions. An R package was also presented for convenient implementation in analysis and study design.

***

1.7 The Opportunity for Novel Contribution within this Thesis

Whilst the literature regarding the incorporation of historical data within the design and analysis of RCTs is fairly substantial, it remains an emerging area of research with significant opportunity to contribute new knowledge. In particular, the majority of the research published to date focuses on the incorporation of historical control data, rather than data pertaining to both intervention and control arms from previous studies such as pilot or feasibility studies. The exception to this is Psioda and Ibrahim [Psioda and Ibrahim, 2019], who discussed using the power prior to incorporate historical data pertaining to both intervention and control participants in order to inform the estimated treatment effect. Furthermore, the majority of the research to date has focused on the development of information borrowing methods for use within individually randomised trials. To date, the single exception to this is the recent PhD thesis [Xiao, 2017] outlined above. However, the research presented within that thesis proposed power prior methodology in which the discounting factor was fixed, and only considered the incorporation of non-clustered historical data. There is opportunity to explore methods in which the discounting factor is treated as a parameter, and to consider how these approaches can be applied to the incorporation of clustered historical trial data (e.g. obtained through a pilot CRCT). Combined with the paucity of methodological research focusing on Bayesian methods within CRCTs more generally in recent years, there therefore remains significant opportunity to contribute novel research to an emerging field of trial methodology.

***

1.8 Overview of the Thesis

This thesis presents a thorough exploration of the use of Bayesian methods in the design and analysis of CRCTs. In Chapter 2 a systematic review is presented, which identifies and quantifies both methodological research in the field of Bayesian CRCTs, and the practical application of Bayesian methods in the design and analysis of CRCTs. Chapter 3 proposes novel methodology to facilitate the construction of informative prior distributions based on historical data, such as data from a pilot or feasibility study, through the use of power priors. The impact of using these informative power priors is explored, through simulation, in the context of study design and sample size calculation in Chapter 4. Chapter 5 proposes an approach to maximise the amount of information borrowed through an informative power prior, whilst also controlling Bayesian type I error at some nominal level. Finally, in Chapter 6, an R package is presented which facilitates straightforward implementation of the novel methodology developed within this thesis.

Chapter 2

Bayesian Statistics in the Design and Analysis of Cluster Randomised Controlled Trials and their Reporting Quality: a Methodological Systematic Review

A systematic review exploring and quantifying both methodological research into, and application of, Bayesian methodology in cluster randomised controlled trials, is presented within this chapter. The quality of trials reporting the use of Bayesian methods is summarised, measured using metrics from the CONSORT extension to cluster randomised controlled trials. Both the main review (2018) and an updated review (2021) are presented.

***

2.1 Introduction

THE methodological systematic review presented in this chapter has been previously published [Jones et al., 2021], and has been updated and refined for inclusion in this thesis.

CRCTs are a relatively novel study design, but the methodology is now well established in the literature. There have been rapid increases both in the practical application of the CRCT design, and in the development of the underpinning statistical methodology. This is illustrated in Figure 2.1, which shows a year-on-year increase in the number of PubMed results for a search of CRCTs. Alongside such a rapid increase in the use of the CRCT design, there have been some attempts to develop new Bayesian methodology for the design and analysis of such trials. This ranges from utilising well-established Bayesian hierarchical modelling approaches to account for the clustered nature of the data [Spiegelhalter, 2001], through to more novel approaches to study design and sample size calculation [Turner et al., 2004, Turner et al., 2005]. The Bayesian approach to analysis in particular may offer a number of advantages over the frequentist approach in CRCTs. When fitting hierarchical models, as is often applicable in the analysis of a CRCT, the hierarchical Bayesian framework provides a flexible, intuitive approach to statistical inference. Furthermore, Bayesian analysis facilitates a more natural, probabilistic interpretation of results, and has the potential to facilitate a transition away from frequentist hypothesis testing and p-values, an approach which has faced increasing criticism in recent years [Wasserstein et al., 2019]. Whilst often condemned for its potential for introducing subjectivity, the use of Bayesian methods can utilise prior information from other sources as opposed to basing inference solely on a single dataset at hand. The incorporation of existing evidence or known properties about the possible distribution of the parameters of interest can lead to more reliable and robust statistical inferences and conclusions. In many cases, the rationale for the inclusion of informative priors is well justified, for example, results from previous research, expert opinion or even existing data (such as from pilot or feasibility studies). Whilst the potential advantages of the Bayesian approach to both the analysis of clinical trials [Lewis and Wears, 1993] and hierarchical data [Gelman et al., 2013] have been previously discussed and documented in the literature, it is unclear whether such methods are being regularly utilised within the context of CRCTs.

With the increased use of CRCTs, the need for consistent, high-quality reporting is crucial. In response to this recognised need, the CONSORT extension to Cluster Randomised Trials was first published in 2004 [Campbell et al., 2004] and updated in 2012 [Campbell et al., 2012]. The CONSORT statement provides recommendations for reporting of randomised trials and, whilst there is no extension for trials using Bayesian methodology, nor was it developed exclusively for frequentist methods. A recent review of the methodological quality of sample size calculations in a sample of 300 CRCTs published between 2000 and 2008, found that only 55.3% (N = 166) presented a sample size calculation, of which only 61.4% (N = 102) accounted appropriately for clustering [Rutterford et al., 2015]. A separate published review of the same sample of CRCTs examined the impact of the 2004 CONSORT CRCT extension on more general methodological quality and concluded that adherence to published reporting guidelines and quality remained low [Ivers et al., 2011]. Similar reviews of reporting quality in more recent CRCTs (up to 2015) have been conducted and produced comparable conclusions [Diaz-Ordaz et al., 2013, Tokolahi et al., 2016]. However, to the author's knowledge, no previous review has focussed specifically on CRCTs which incorporated Bayesian methods, and so both the quantity and methodological quality of these are unknown.

This review aimed to:

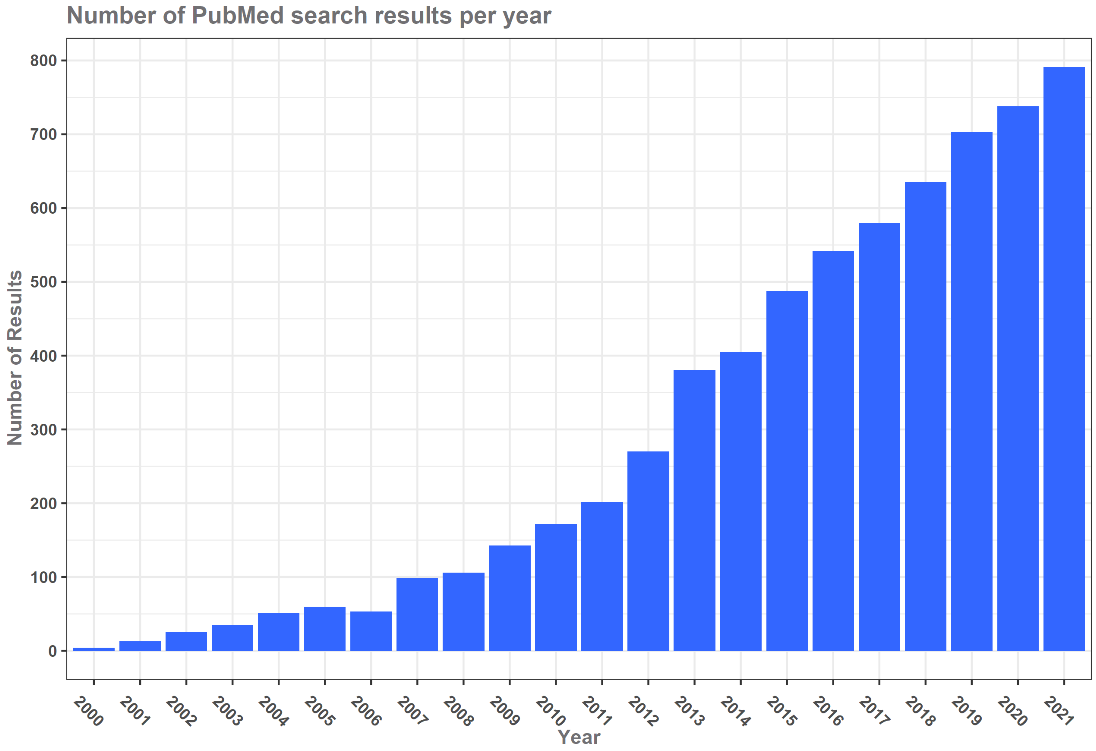


Figure 2.1: Number of PubMed search results per year for a search of “cluster randomised controlled trial” OR “cluster randomized controlled trial” NOT “stepped”.

1. Quantify and explore the use of Bayesian methodology in the design and/or analysis of CRCTs;

2. Appraise the quality of reporting of CRCTs conducted in a Bayesian framework against the current relevant CONSORT guidelines and identify whether the reporting quality differs from previous reviews assessing reporting quality in CRCTs more generally (most of which pertain to frequentist trials);

3. Identify relevant methodological research in the field of Bayesian methods for CRCTs.

The impact of the introduction of the CONSORT extension for CRCTs in 2004, and update in 2012, on reporting quality was also appraised. The initial searches for this systematic review were run in 2018 during the early stages of the PhD programme. However, as the wider PhD studies were ongoing until early 2022, a brief update, with more constrained aims, was undertaken in September 2021. Specifically, the aims of the update were to: (i) identify any additional CRCTs using Bayesian methodology in their design or analysis; and (ii) to identify any further methodological developments in this area.

***

2.2 Methods

The full protocol for this methodological systematic review was developed prospectively and made publicly available online [Jones, 2018] (Appendix A) before commencing the literature searches. The review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [Moher et al., 2009].

2.2.1 Inclusion and Exclusion Criteria

All published parallel group CRCTs in which Bayesian methodology was used in either the study design (including sample size calculation) or statistical analysis were sought. In addition, any papers in which Bayesian methodology was discussed or considered, even if such methods were not implemented in the study, were also eligible for inclusion, whilst recognising that such a scenario would be unlikely. No restriction to the search strategy or inclusion criteria was made on the basis of publication date, location, intervention type or population in any way, provided the relevant paper was published in the English language.

In order to be included in this review, it had to be evident that randomisation in the study occurred at a cluster level, in which multiple units (participants) were randomised together, as per the definition of a CRCT.

References were not excluded on the basis of type (category) of published paper. Specifically, not only were primary reports of efficacy or effectiveness included, but also protocol papers, papers reporting secondary analyses and publications reporting results of pilot/feasibility studies. Studies reporting Bayesian methodological developments in the area of CRCTs were also identified and included. At the data extraction stage, supplementary literature related to the same study was sought, if indicated, to obtain the required information, but such examples were only included as a single entry. It was anticipated, for example, that this might include obtaining additional detail, from a published protocol or monograph, that had been omitted in the corresponding primary results paper.

Published papers reporting only cost-effectiveness analyses, results or methodology were excluded. Studies implementing a stepped wedge or other longitudinal cluster randomised design were also excluded, as the methodological considerations are different and the reporting quality metrics presented in the CONSORT extension to CRCTs [Campbell et al., 2012] are not always valid for such designs. Since commencement of this systematic review, separate CONSORT guidelines for stepped-wedge designs have been published [Hemming et al., 2018]. Conference proceedings and masters and PhD dissertations were not included, although a relevant PhD thesis is outlined in Chapter 1.

2.2.2 Data Sources and Search Methods

The searches for relevant publications were run on 24 July 2018, through both Medline and Embase using Ovid, as well as the Cochrane Central Register of Controlled Trials (CENTRAL). The full electronic search strategy was an extension of a previously proposed strategy to identify CRCTs [Taljaard et al., 2010], adapted to identify only studies including the word (or variations of) “Bayes” in the title, abstract or text. The full electronic search strategy used to search Medline and Embase is shown in Table 2.1, with minor syntactic adaptations required in order to run the search in CENTRAL, as shown in Table 2.2.

Additional literature was included where appropriate through additional *ad hoc* hand searching of personal reference collections, in particular to identify relevant methodological publications.

Table 2.1: Search Strategy used to search Medline and Embase within Ovid.

#	Search
	Existing published strategy for randomised controlled trials
1	(article OR randomized controlled trials).pt.
2	Animals/
3	Humans/
4	# 2 NOT (#2 AND # 3)
5	# 1 NOT # 4
	Cluster-design related terms
6	(cluster$ adj2 randomi$).tw.
7	((communit$ adj2 intervention$) or (communit$ adj2 randomi$)).tw.
8	group$ randomi$.tw.
9	#6 OR #7 OR #8
10	intervention?.tw.
11	Cluster Analysis/
12	Health Promotion/
13	Program Evaluation/
14	Health Education/
15	#10 OR #11 OR #12 OR #13 OR #14
16	#9 OR #15
	Bayesian search terms
17	bayes$.af.
18	#16 AND #17
	Final search
19	#18 AND #5
20	limit #19 to (randomized controlled trial)

Table 2.2: Search Strategy used to search CENTRAL.

#	Search
	Existing published strategy for randomised controlled trials
1	“randomized controlled trial”:pt
2	MeSH descriptor: [Animals] explode all trees
3	MeSH descriptor: [Humans] explode all trees
4	#2 NOT (#2 AND #3)
5	# 1 NOT # 4
	Cluster-design related terms
6	cluster* near/2 randomi*:ti,ab,kw
7	((communit* near/2 intervention*) OR (communit* near/2 randomi*)):ti,ab,kw
8	group* randomi*:ti,ab,kw
9	#6 OR #7 OR #8
10	intervention?:ti,ab,kw
11	MeSH descriptor: [Cluster Analysis] explode all trees
12	MeSH descriptor: [Health Promotion] explode all trees
13	MeSH descriptor: [Program Evaluation] explode all trees
14	MeSH descriptor: [Health Education] explode all trees
15	#10 OR #11 OR #12 OR #13 OR #14
16	#9 OR #15
	Bayesian search terms
17	bayes*
18	#16 AND #17
	Final search
19	#18 AND #5

2.2.3 Reference Sifting and Quality Control

After conducting electronic searches, all references were downloaded and imported to Mendeley [Elsevier, 2020] for electronic deduplication. Following this, remaining references were exported and uploaded to Rayyan [Ouzzani et al., 2016]. Each reference was independently reviewed by two reviewers and a decision made to include or exclude on the basis of the information available from the title and the abstract, assessed against the pre-specified inclusion/exclusion criteria [Jones, 2018] (Appendix A). Rayyan includes a blinding feature, which was switched on during the independent sifts and then disabled. Any disagreements were resolved through discussion and, where required, a final decision was made by a third, independent reviewer.

After the initial sift, full text articles were obtained for all remaining references. The full texts were reviewed and once again inclusion/exclusion decisions were captured using Rayyan. Two independent reviewers then re-examined approximately half each of all full texts and independently made inclusion or exclusion decisions. Any disagreements were once again resolved through further discussion.

2.2.4 Data Extraction

For the primary and secondary published reports of trial results, a range of data was collected, including demographic data, technical detail regarding design and analysis methodology with relation to Bayesian techniques, and information regarding statistician involvement with the study and their respective affiliations. For papers reporting primary results, a selection of reporting quality metrics, taken from the 2012 CONSORT extension to CRCTs [Campbell et al., 2012], was also collected. In addition, whether or not p-values were reported for comparison of baseline demographics was recorded, as has been collected in previous systematic reviews of CRCTs [Diaz-Ordaz et al., 2013, Froud et al., 2012] as well as Clinical Trial Unit (CTU) involvement in the study, and journal endorsement of the CONSORT guidelines.

Based on a previously used criterion [Diaz-Ordaz et al., 2013, Delgado-Rodriguez et al., 2001, Dechartres et al., 2011], a paper was considered to have had statistician involvement if there was a clearly designated statistician, or if at least one of the co-authors belonged to a department of epidemiology or biostatistics. If it was not possible to obtain this information from the authorship list on the paper, online searching was undertaken to determine this from the qualification or affiliation of the authors. In any cases where it was not possible to obtain the required information, statistician involvement was recorded as “no”. The statistician’s affiliation to a CTU, an academic statistical department, a commercial pharmaceutical company, a Clinical Research Organisation (CRO) or “other” was collected. CTU involvement in the study was determined if at least one author had a listed affiliation to a CTU. If author affiliations were not available in the paper or online, this was recorded as “no”.

Journal endorsement of the CONSORT statement was assessed using previously defined criteria [Diaz-Ordaz et al., 2013]. Specifically, a journal’s strength of endorsement was classified as: (i) high if the words “required”, “must”, “should” or “strongly recommended” were used in their author instructions; (ii) medium if the words “encouraged”, “recommended”, “advised” or “please” were used; and (iii) low if “may wish to consider” or “see CONSORT” was used. A fourth category, “none”, was recorded if the journal made no mention of the CONSORT statement in its guidelines to authors.

Separate data extraction forms were developed for primary (Appendix B) and secondary (Appendix C) results papers to ensure that all the required information was obtained independently, consistently and without bias. The forms were piloted prior to data extraction. Data extraction was undertaken by two people, independently, and all discrepancies discussed and finalised. Classification of each paper as primary or secondary was also undertaken by two people, independently. Any disagreements were resolved through discussion. All data recorded in the data extraction forms were double-entered in to separate excel spreadsheets for primary and secondary papers.

Formal data extraction was not undertaken for the methodological papers, but rather these papers were examined for the purpose of qualitative reporting and descriptive summaries of the methods developed, to gain an understanding of the extent of methodological developments in this area. The methodological papers were examined once, by the same single reviewer.

2.2.5 Analysis

Descriptive statistics of frequencies and percentages or means and standard deviations are presented, as appropriate, for demographic data relating to each of the results publications, including trial location, number of participants recruited and type of primary outcome, by category of published results (primary or secondary). For the reporting quality measures, the number of primary results papers satisfying each criterion are presented overall, by year (categorised as being published before or after the publication of the 2012 extension to the CONSORT guidelines for CRCTs [Campbell et al., 2012]), by journal endorsement of the CONSORT guidelines (high or medium versus low or none) and by statistician involvement in the trial. The use or consideration of Bayesian methods in the design and/or sample size calculation and/or analysis are also quantified and presented, as well as the level of information incorporated into the prior distributions specified. The parameters for which the prior distributions were specified is also reported, if this information was available. Finally, a qualitative synthesis of the methodological papers was undertaken to summarise the areas of focus in the development of new methods.

2.2.6 Systematic Review Update September 2021

The search strategies outlined in Table 2.1 and Table 2.2 were re-run on 30 September 2021, constrained to the years 2018 – 2021. In order to avoid any duplication of references identified in both searches, all references identified for 2018 were manually cross-checked against the initial reference list from the first search, and any which appeared in both were removed from the 2021 update.

After removing items already identified in the original search, the remaining references were deduplicated in Mendeley, before being uploaded into Rayyan for sifting. As previously, the first sift was undertaken on the basis of titles and abstracts alone, and the second on the basis of the full texts. In the update, each stage of the sifting was only undertaken once.

Whilst the references identified in the update were categorised into primary or secondary results papers, or methodological papers, no formal data extraction was undertaken. Rather, the papers were collated for the purposes of a qualitative summary to quantify, assess and report any noteworthy uses of Bayesian methods in CRCTs, or any methodological developments, since the initial searches in 2018.

***

2.3 Results from the 2018 Search

Running the electronic search strategy in 2018 identified 325 records, of which 48 were identified as duplicates and removed. The remaining 277 records were screened on the basis of the detail available within the title and abstract, of which 219 were excluded: 51 were the wrong study design (such as non-randomised studies, stepped wedge designs or meta-analyses), 160 were individually randomised trials and eight were papers reporting cost-effectiveness only. Full texts were obtained for the remaining 58 papers. At this final stage, following independent review of the full texts, a further 37 were removed (25 were individually randomised, five did not include any mention of Bayesian methodology, six were the wrong study design and one paper reported only cost-effectiveness results), leaving 21 papers from the electronic search. A further six papers, all of which were methodological papers, were added through additional hand searches, resulting in a total of 27 included items (Figure 2.2). The full list of references for the included papers is detailed in Table 2.3. Eleven (41%) were reports of CRCT results, of which seven (64%, R1 – R7) were primary results papers and four (36%, R8 – R11) reported secondary analyses. Thirteen papers (48%, M1 – M13) reported methodological developments and the remaining three (11%, C1 – C3) reported comparisons of methods, assessing the performance of various existing methodologies.

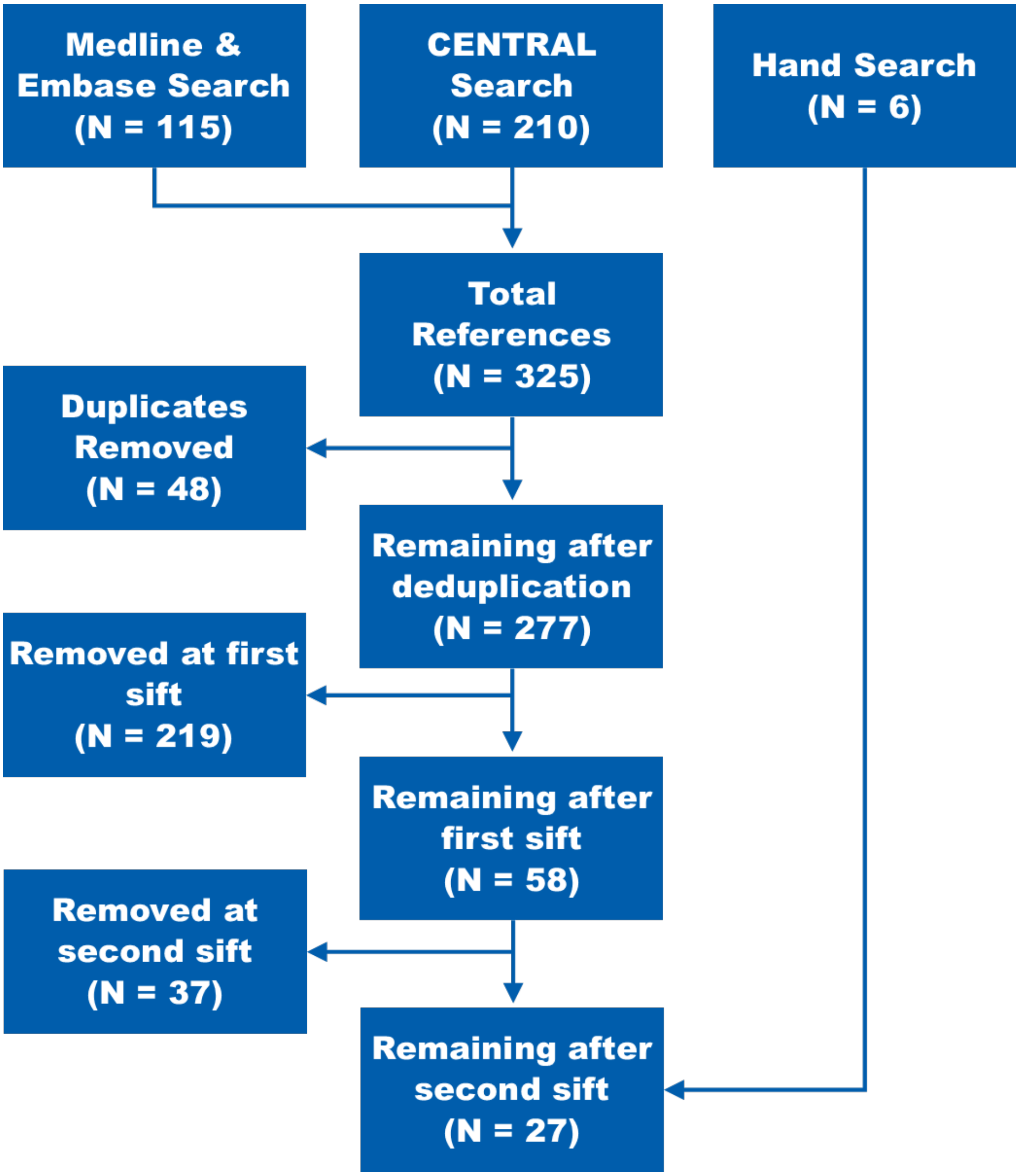

Medline & Embase Search (N = 115)
CENTRAL Search (N = 210)
Hand Search (N = 6)
Total References (N = 325)
Duplicates Removed (N = 48)
Remaining after deduplication (N = 277)
Removed at first sift (N = 219)
Remaining after first sift (N = 58)
Removed at second sift (N = 37)
Remaining after second sift (N = 27)

Figure 2.2: Flow diagram of the identification process for the 27 publications included in the systematic review.

Table 2.3: References included in the review. Prefix "R" refers to results papers, "M" to methodological papers and "C" to comparison of methods papers.

R1	Carabin H, Millogo A, Ngowi HA, et al. Effectiveness of a community-based educational programme in reducing the cumulative incidence and prevalence of human Taenia solium cysticercosis in Burkina Faso in 2011–14 (EFECAB): a cluster-randomised controlled trial. *Lancet Glob Heal.* 2018;6(4):e411-e425. doi:10.1016/S2214-109X(18)30027-5
R2	Foxcroft DR, Callen H, Davies EL, Okulicz-Kozaryn K. Effectiveness of the strengthening families programme 10-14 in Poland: Cluster randomized controlled trial. *Eur J Public Health.* 2017;27(3):494-500. doi:10.1093/eurpub/ckw195
R3	Levy BT, Hartz A, Woodworth G, Xu Y, Sinift S. Interventions to Improving Osteoporosis Screening: An Iowa Research Network (IRENE) Study. *J Am Board Fam Med.* 2009;22(4):360-367. doi:10.3122/jabfm.2009.04.080071
R4	Ngowi HA, Carabin H, Kassuku AA, Mlozi MRS, Mlangwa JED, Willingham AL. A health-education intervention trial to reduce porcine cysticercosis in Mbulu District, Tanzania. *Prev Vet Med.* 2008;85(1-2):52-67. doi:10.1016/j.prevetmed.2007.12.014
R5	Rahme E, Choquette D, Beaulieu M, et al. Impact of a general practitioner educational intervention on osteoarthritis treatment in an elderly population. *Am J Med.* 2005;118(11):1262-1270. doi:10.1016/j.amjmed.2005.03.026
R6	Swanson KM, Chen H-T, Graham JC, Wojnar DM, Petras A. Resolution of Depression and Grief during the First Year after Miscarriage: A Randomized Controlled Clinical Trial of Couples-Focused Interventions. *J Women's Heal.* 2009;18(8):1245-1257. doi:10.1089/jwh.2008.1202
R7	Van Deurssen E, Meijster T, Oude Hengel KM, et al. Effectiveness of a Multidimensional Randomized Control Intervention to Reduce Quartz Exposure among Construction Workers. *Ann Occup Hyg.* 2015;59(8):959-971. doi:10.1093/annhyg/mev037
R8	Amza A, Kadri B, Nassirou B, et al. Community risk factors for ocular chlamydia infection in Niger: Pre-treatment results from a cluster-randomized trachoma trial. *PLoS Negl Trop Dis.* 2012;6(4). doi:10.1371/journal.pntd.0001586

R9 Hovi T, Ollgren J, Savolainen-Kopra C, T. H, J. O. Intensified hand-hygiene campaign including soap-and-water wash may prevent acute infections in office workers, as shown by a recognized-exposure - adjusted analysis of a randomized trial. *BMC Infect Dis*. 2017;17(1):47. doi:http://dx.doi.org/10.1186/s12879-016-2157-z

R10 Barlis P, Regar E, Serruys PW, et al. An optical coherence tomography study of a biodegradable vs. durable polymer-coated limus-eluting stent: A LEADERS trial sub-study. *Eur Heart J*. 2010;31(2):165-176. doi:10.1093/eurheartj/ehp480

R11 See CW, O'Brien KS, Keenan JD, et al. The effect of mass azithromycin distribution on childhood mortality: Beliefs and estimates of efficacy. *Am J Trop Med Hyg.* 2015;93(5):1106-1109. doi:10.1111/sjos.12316

M1 Alexander N, Emerson P. Analysis of incidence rates in cluster-randomized trials of interventions against recurrent infections, with an application to trachoma. *Stat Med*. 2005;24(17):2637-2647. doi:10.1002/sim.2138

M2 Clark AB, Bachmann MO. Bayesian methods of analysis for cluster randomized trials with count outcome data. *Stat Med*. 2010;29(2):199-209. doi:10.1002/sim.3747

M3 Nixon RM, Duffy SW, Fender GR. Imputation of a true endpoint from a surrogate: Application to a cluster randomized controlled trial with partial information on the true endpoint. *BMC Med Res Methodol*. 2003;3:1-11. doi:10.1186/1471-2288-3-17

M4 Olsen MK, DeLong ER, Oddone EZ, Bosworth HB. Strategies for analyzing multilevel cluster-randomized studies with binary outcomes collected at varying intervals of time. *Stat Med*. 2008;27(29):6055-6071. doi:10.1002/sim.3446

M5 Thompson SG, Warn DE, Turner RM. Bayesian methods for analysis of binary outcome data in cluster randomized trials on the absolute risk scale. *Stat Med*. 2004;23(3):389-410. doi:10.1002/sim.1567

M6 Turner RM, Prevost AT, Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med.* 2004;23(8):1195-1214. doi:10.1002/sim.1721

M7 Turner RM, Omar RZ, Thompson SG. Modelling multivariate outcomes in hierarchical data, with application to cluster randomised trials. *Biometrical J*. 2006;48(3):333-345. doi:10.1002/bimj.200310147

M8	Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous responses. *Stat Med.* 2001;20(3):435-452. doi:10.1002/1097-0258(20010215)20:3<435::AID-SIM804>3.0.CO;2-E
M9	Kikuchi T, Gittins J. A behavioural Bayes approach for sample size determination in cluster randomized clinical trials. *J R Stat Soc Ser C Appl Stat.* 2010;59(5):875-888. doi:10.1111/j.1467-9876.2010.00732.x
M10	Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clin Trials*. 2005;2(2):108-118. doi:10.1191/1740774505cn072oa
M11	Turner RM, Omar RZ, Thompson SG. Constructing intervals for the intracluster correlation coefficient using Bayesian modelling, and application in cluster randomized trials. *Stat Med*. 2006;25(9):1443-1456. doi:10.1002/sim.2304
M12	Uhlmann L, Jensen K, Kieser M. Bayesian network meta-analysis for cluster randomized trials with binary outcomes. *Res Synth Methods.* 2016;8(October 2015):236-250. doi:10.1002/jrsm.1210
M13	Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med*. 2001;20(3):453-472. doi:10.1002/1097-0258(20010215)20:3<453::AID-SIM803>3.0.CO;2-L
C1	Peters TJ, Richards SH, Bankhead CR, Ades AE, Sterne JAC. Comparison of methods for analysing cluster randomized trials: An example involving a factorial design. *Int J Epidemiol*. 2003;32(5):840-846. doi:10.1093/ije/dyg228
C2	Pacheco GD, Hattendorf J, Colford JM, Mäusezahl D, Smith T. Performance of analytical methods for overdispersed counts in cluster randomized trials: Sample size, degree of clustering and imbalance. *Stat Med*. 2009;28(24):2989-3011. doi:10.1002/sim.3681
C3	Ma J, Thabane L, Kaczorowski J, et al. Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: The community hypertension assessment trial (CHAT). *BMC Med Res Methodol*. 2009;9(1). doi:10.1186/1471-2288-9-37

2.3.1 Demographics

Descriptions of demographics in the 11 results papers are displayed in Table 2.4. Target sample sizes and numbers of clusters were only collected for primary results papers. It was deemed necessary to distinguish "numbers approached" from target sample sizes, as the numbers approached seemed likely driven by logistical rather than statistical considerations, and so were not included in the summary statistics of the target sample sizes. Clear statistician involvement was identified in eight (73%) of the 11 results papers, and in one (13%) of those eight the statistician had a clear association with a CTU. It was not possible to identify more general CTU involvement with trial or data management in any instance.

Table 2.4: Demographic characteristics for the eleven results papers.

N (%) unless otherwise stated	Total $(N = 11)$	Primary $(N = 7)$	Secondary $(N = 4)$
Year of Publication			
Pre 2005	0 (0)	0 (0)	0 (0)
2005 - 2012	6 (55)	4 (57)	2 (50)
Post 2012	5 (46)	3 (43)	2 (50)
Location of First Author[a]			
UK	2 (18)	1 (14)	1 (25)
US/Canada	5 (46)	4 (57)	1 (25)
Europe excluding UK	3 (27)	1 (14)	2 (50)
Australia/New Zealand	0 (0)	0 (0)	0 (0)
Africa	2 (18)	1 (14)	1 (25)
Asia	0 (0)	0 (0)	0 (0)
Other	0 (0)	0 (0)	0 (0)
Location of Study[a]			
UK	1 (9)	0 (0)	1 (25)
US/Canada	3 (27)	3 (43)	0 (0)
Europe excl. UK	4 (36)	2 (29)	2 (50)
Australia/New Zealand	0 (0)	0 (0)	0 (0)
Africa	4 (36)	2 (29)	2 (50)
Asia	0 (0)	0 (0)	0 (0)
Other	0 (0)	0 (0)	0 (0)

Target Sample Size; mean (SD) [range]	N/A	N = 3[b] 1466.7 (1868.6) [120, 3600]	N/A
Target Number of Clusters; mean (SD) [range]	N/A	N = 2[c] 200.0 (198.0) [60, 340]	N/A
Recruited Sample Size; mean (SD) [range]	N = 11 10898.5 (19816.1) [116, 66204]	N = 7 2484.6 (3700.1) [116, 9928]	N = 4 25662.8 (28762.5) [683, 66204]
Recruited Number of Clusters; mean (SD) [range]	N = 11 58.8 (95.6) [5, 341]	N = 7 69.1 (121.6) [5, 341]	N = 4 40.8 (13.2) [21,48]
Randomisation Unit			
Medical Facility	1 (9)	1 (14)	0 (0)
Village/Community/District	6 (55)	4 (57)	2 (50)
Organisation	1 (9)	1 (14)	0 (0)
Couple	1 (9)	1 (14)	0 (0)
Individual	1 (9)	0 (0)	1 (25)
Working Unit (office)	1 (9)	0 (0)	1 (25)
Primary Outcome Type			
Binary	9 (82)	5 (71)	4 (100)
Continuous	2 (18)	2 (29)	0 (0)
Statistician Involvement	8 (73)	5 (71)	3 (75)
Statistician Association[d]			
Clinical Trials Unit	1 (13)	0 (0.0)	1 (33)
Academic Statistical Department	7 (88)	5 (100)	2 (67)
Pharmaceutical Company	0 (0)	0 (0)	0 (0)
Clinical Research Organisation	0 (0)	0 (0)	0 (0)
Other	0 (0)	0 (0)	0 (0)

Journal Endorsement of the CONSORT Guidelines			
High	N/A	3 (43)	N/A
Medium	N/A	1 (14)	N/A
Low	N/A	0 (0)	N/A
None	N/A	3 (43)	N/A

[a] One author was associated with an institution in both Europe and the UK, and the associated study was run across both locations. The denominator used for the calculations is based on the number of papers

[b] Two studies specified the number of participants approached but these were not explicitly stated/justified recruitment targets and so were excluded

[c] Four studies specified the number of clusters approached but these were not explicitly stated/justified recruitment targets and so were excluded

[d] Denominators used to calculate percentages are based on the number of studies with statistician involvement

2.3.2 Reporting Quality

The reporting quality of the seven primary results papers was mixed (Table 2.5). Four papers (57%) included a description of the sample size calculation, but none of these clearly accounted for clustering, provided the ICC used in the sample size calculation or took into consideration potential variability in cluster size or accounted for this in the sample size calculation. Similarly, none of these seven papers reported estimated ICCs for any of the primary or secondary outcomes, despite the potential value of such estimates in informing the design of future studies. However, it was clear in six (86%) of the primary results papers how clustering was accounted for in the statistical analysis.

Reporting quality metrics have also been summarised by: i) publication date before or after the publication of the CONSORT extension to CRCTs in 2012; ii) journal endorsement of the CONSORT guidelines; and iii) involvement of a statistician in the study (Table 2.5). Due to the small number of available papers, journal endorsement of the CONSORT guidelines was dichotomised in to “High” or “Medium” versus “Low” or “None”. The intention was to summarise these results by three time periods (pre-2005, 2005 – 2012 and 2012 – 2018) to assess any effect of the publication of the CONSORT extensions for CRCTs in 2004 and 2012 on reporting quality. However, no CRCTs using Bayesian methodology were identified that were published before 2005. Summary statistics pertaining to the pre-specified reporting quality metrics are detailed in Table 2.5. However, due to the small number of primary results papers identified (seven in total), no meaningful comparisons of reporting quality between the subgroups (publication date, endorsement of CONSORT guidelines and statistician involvement) can be made.

One of the papers retrieved was a pre-specified sub-study and so was classified as a secondary results paper (Table 2.3, R10). Despite not being a primary results paper and therefore not obligated to follow CONSORT guidelines, the reporting quality of this paper was high: a sample size calculation was presented and appropriately accounted for clustering, including specification of the assumed ICC; the flow of clusters and individuals through the study was well documented; and all levels of clustering were accounted for within a hierarchical modelling framework.

Table 2.5: Reporting quality metrics for the seven primary results papers.

Reporting Quality Criteria N (%)		**Year of Publication**		**Journal Endorsement of CONSORT Guidelines**		**Statistician Involvement**	
	Total (N = 7)	*2012 or Earlier (N = 4)*	*2013 Onwards (N = 3)*	*High/ Medium (N = 4)*	*Low/ None (N = 3)*	*Yes (N = 5)*	*No (N = 2)*
Description of sample size method	4 (57)	2 (50)	2 (67)	2 (50)	2 (67)	2 (40)	2 (100)
Clustering clearly accounted for in sample size calculation[a]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Specification of required number of clusters[a]	2 (50.0)	1 (50)	1 (50)	1 (50)	1 (50)	1 (50)	1 (50)
Specification of assumed cluster size[a]	2 (50)	1 (50)	1 (50)	1 (50)	1 (50)	1 (50)	1 (50)
Specification of whether equal or unequal cluster sizes assumed[a]	1 (25)	1 (50)	0 (0)	0 (0)	1 (50)	0 (0)	1 (50)
Variability in cluster size accounted for[a]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Specification of the ICC used for the sample size calculation[a]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Indication of the uncertainty in the ICC[a]	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Accounted for the uncertainty in the ICC[a]	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Reporting Quality Criteria N (%)		Year of Publication		Journal Endorsement of CONSORT Guidelines		Statistician Involvement	
	Total (N = 7)	*2012 or Earlier (N = 4)*	*2013 Onwards (N = 3)*	*High/ Medium (N = 4)*	*Low/ None (N = 3)*	*Yes (N = 5)*	*No (N = 2)*
Other CONSORT metrics							
Details of how clustering was accounted for in the analysis	6 (86)	4 (100)	2 (67)	4 (100)	2 (67)	5 (100)	1 (50)
Specified the number of clusters randomised	7 (100)	4 (100)	3 (100)	4 (100)	3 (100)	5 (100)	2 (100)
Specified the number of clusters receiving intended treatment							
Explicit	5 (71)	3 (75)	2 (67)	4 (100)	1 (33)	4 (80)	1 (50)
Implied	2 (29)	1 (25)	1 (33)	0 (0)	2 (67)	1 (20)	1 (50)
Specified the number of clusters in primary outcome analysis							
Explicit	2 (29)	1 (25)	1 (33)	2 (50)	0 (0)	2 (40)	0 (0)
Implied	5 (71)	3 (75)	2 (67)	2 (50)	3 (100)	3 (60)	2 (100)
Details of cluster-level losses and exclusions							
Explicit	3 (43)	2 (50)	1 (33)	2 (50)	1 (33)	2 (40)	1 (50)
Implied	4 (57)	2 (50)	2 (67)	2 (50)	2 (67)	3 (60)	1 (50)
Details of individual-level losses and exclusions	4 (57)	2 (50)	2 (67)	2 (50)	2 (67)	2 (40)	2 (100)

Reporting Quality Criteria N (%)		**Year of Publication**		**Journal Endorsement of CONSORT Guidelines**		**Statistician Involvement**	
	Total (N = 7)	*2012 or Earlier (N = 4)*	*2013 Onwards (N = 3)*	*High/ Medium (N = 4)*	*Low/ None (N = 3)*	*Yes (N = 5)*	*No (N = 2)*
Individual-level baseline characteristics	7 (100)	4 (100)	3 (100)	4 (100)	3 (100)	5 (100)	2 (100)
Cluster-level baseline characteristics	2 (29)	2 (50)	0 (0)	1 (25)	1 (33)	1 (20)	1 (50)
ICCs provided for primary outcomes	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ICCs provided for secondary outcomes[b]	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
P-values provided for baseline comparisons	5 (71)	3 (75)	2 (67)	3 (75)	2 (67)	3 (60)	2 (100)
Clustering clearly accounted for in calculation of p-values[c]	1 (20)	1 (33)	0 (0)	1 (33)	0 (0)	1 (33)	0 (0)
Unclear if clustering accounted for in calculation of p-values[c]	1 (20)	1 (33)	0 (0)	1 (33)	0 (0)	1 (33)	0 (0)

[a] The denominator used to calculate the percentages is based on the number of studies which described the method of sample size calculation

[b] One study did not report any secondary outcomes

[c] The denominator used to calculate the percentages is based on the number of studies that provided p-values for baseline comparison

2.3.3 Use of Bayesian Methodology

No results papers were identified, which followed, or even discussed, a Bayesian approach to study design or sample size calculation. One secondary paper did, however, specify that the design effect used to inflate the sample size calculation was derived from the results of a Bayesian hierarchical model (R10).

Of the eleven results papers included in the review, all adopted some form of Bayesian approach to statistical analysis (Table 2.6). In nine (82%; R1-R7, R9, R10) of the 11 papers, Bayesian hierarchical modelling techniques were employed to account for the clustered structure of the data. Another study (R8) employed Bayes Model Averaging to conduct multiple regression, citing the risk of overfitting that can be associated with stepwise regression in model-fitting as the reason for adopting this approach. One study conducted a literature search of Cochrane Reviews and extracted the key summary statistics (mortality) before converting each in to a log-odds ratio. These statistics were combined in to a single arithmetic mean in order to construct an empirical prior. This prior was then combined with the likelihood from the CRCT to obtain a Bayesian posterior distribution of the relative risk of mortality in the intervention group versus the control group (R11).

Table 2.6: Summary of Bayesian methods used in primary and secondary results papers.

N (%) unless otherwise stated	Total	Primary	Secondary
	$(N = 11)$	$(N = 7)$	$(N = 4)$
Sample Size (Used)	0 (0.0)	0 (0.0)	0 (0.0)
Sample Size (Discussed)	0 (0.0)	0 (0.0)	0 (0.0)
Analysis (Used)	11 (100.0)	7 (100.0)	4 (100.0)
Priors used			
Informative	2 (18.2)[a]	1 (14.3)[a]	1 (25.0)
Weakly Informative	1 (9.1)	1 (14.3)	0 (0.0)
Non-informative	5 (45.5)[a]	3 (42.9)[a]	2 (50.0)
Unspecified	4 (36.4)	3 (42.9)	1 (25.0)
Analysis (Discussed)	N/A	N/A	N/A

[a] One paper reported the use of two Bayesian models - the first model implementing a non-informative prior and the second model utilising "collateral" information.

In these results papers, prior distributions used in the analyses were informative in two (18%; R3, R11) papers. In one, (R3) "collateral" information from a previous study was used to construct a prior distribution for the variation in practice effects (specifically, the standard deviation for practice-level rates). In the other (R11) an informative prior distribution for the treatment effect parameter within a negative binomial regression was constructed based on a meta-analysis of relevant reviews obtained from the Cochrane library, and used to inform the estimation of the outcome of interest (the relative risk of childhood mortality). No information was provided on the prior distributions placed on the variance components. Weakly informative prior distributions were used in one (9%; R2) study, by placing Student-t distributed priors centred at 0 on the treatment effect parameter and other fixed logistic regression coefficients, which the authors acknowledged would only affect inference if the data provided little information about the parameters. No detail was provided on the prior distributions specified for the variance components in this paper. Five (46%; R1, R3, R5, R9, R10) papers specified the use of non-informative prior distributions, although only one of these (R5) provided more specific detail, stating normal prior distributions for the treatment effect and each of the fixed logistic regression coefficients, and uniform prior distributions for the variance components. Four studies (36%; R4, R6, R7, R8) did not specify their choice of prior distribution. One paper fitted two Bayesian models (R3) - one model implementing a non-informative prior and the other utilising "collateral" information, so the use of both an informative and a non-informative prior was recorded.

2.3.4 Bayesian Methodological Developments

Thirteen (48%) of the 27 papers included in the review were categorised as methodological papers, where the focus was on the development of Bayesian methods for use in the design or analysis of CRCTs, as opposed to applying existing methods to data from CRCTs. Of these 13 papers, 11 (85%) were defined as "pure" methods papers, in which Bayesian methodological developments are reported independently of an applied scenario (although study data may have been used to demonstrate the method). Two (15%) of the 13 papers were categorised as being methodological but with the developments being driven by a specific statistical problem encountered in a CRCT, in which the method is presented and subsequently used to analyse the data of interest. Finally, three (11%) of the 27 papers were categorised as comparison of methods papers, in which existing methodology (both Bayesian and frequentist) were applied to the same data for comparative purposes.

Of the 11 "pure" methodological papers, seven (64%; M2, M4, M5, M7, M11, M12, M13) presented analysis methods, two (18%; M6, M9) presented methods for design/sample size calculation and two (18%; M8, M10) presented elements of both. Both papers driven by a specific application presented analysis methods (M1, M3).

The analysis methods papers predominantly presented Bayesian hierarchical modelling methodology applied to dealing with a range of data types, such as incidence rates (M1), count data (M2) and binary data (M4, M5, M13), in a Bayesian setting, citing flexibility of modelling and the ability to incorporate prior information and account for the complex variance structures as key advantages. One paper reported Bayesian methods for modelling multivariate outcomes (M7), which allow for multiple outcomes without concern for multiplicity whilst accommodating complex correlation structures. Another paper presented Bayesian network meta-analysis methods for CRCTs (M12), allowing for comparison of multiple treatment arms whilst accounting for the complex correlation structure inherent in clustered data.

A number of methodological papers identified within the review focused on the ICC. One such paper centred on analysis only, presenting methods for constructing credible intervals for the ICC and suggesting prior distributions for use in modelling (M11). The two papers in which both design and analysis were discussed focus heavily on the ICC; one provided a range of options for choice of prior distribution alongside recommendations, before discussing briefly how the uncertainty in the ICC can be accounted for in sample size calculations (M8). The other paper presented methods for formulating prior distributions for use in sample size calculations and statistical analysis on the basis of multiple previous estimates, whilst incorporating the relevance of the studies from which they were obtained (M10). One of the papers presenting only study design methodology also focused on ICCs, and developed methods to formulate prior distributions from single and multiple previous ICC estimates for use in sample size calculations (M6).

The remaining study design paper presented a behavioural Bayes approach (M9), extending existing methodology [Pezeshk and Gittins, 2002, Gittins and Pezeshk, 2002, Gittins and Pezeshk, 2000b, Gittins and Pezeshk, 2000a] for sample size determination in individually randomised trials to CRCTs. The method incorporates estimated financial costs and benefits of the intervention to produce a net benefit, rather than being based on detecting the more usual target difference in primary outcome alone.

***

2.4 Results from the September 2021 Update

After removing items that had already been identified from the 2018 search, a total of 150 additional references were identified through searching the CENTRAL database, and 136 through Embase and Medline. Deduplication removed 57 of these references, leaving a total of 229 at the first sift. After sifting the remaining publications on the basis of titles and abstracts, 198 references were excluded: 175 due to employing individual-level, rather than cluster-level, randomisation, and 23 as a result of using an ineligible study design, such as non-randomised studies or stepped wedge designs. As a result, 31 papers remained and were assessed again at second sift based on the full-text articles. At second sift, 21 of the 31 remaining papers were excluded: 17 due to using individual-level randomisation, three due to being an ineligible study design and one because no Bayesian methods were used or discussed. This left ten papers for inclusion in the update. In addition, a further three papers were identified through informal searching, and so in total an additional thirteen papers were identified and included in the update. One of the papers identified from informal searching was published in 2014, and so was missed in the original search.

Details of the newly identified papers are shown in Table 2.7. Of the thirteen papers, two are primary results papers (R*1 – R*2), six are secondary results papers (R*3 – R*8), one is a statistical analysis plan (S*1) and the remaining four are methodological papers (M*1 – M*4).

Table 2.7: References included in the review. Prefix “R” refers to results papers, “S” to statistical analysis plans and “C” to comparison of methods papers.

R*1	Nooijen C, Blom V, Ekblom Ö, et al. The effectiveness of multi-component interventions targeting physical activity or sedentary behaviour amongst office workers: A three-arm cluster randomised controlled trial. *BMC Public Health.* 2020;20(1), 1329. https://doi.org/10.1186/s12889-020-09433-7
R*2	Sanchez Z, Valente J, Galvão P, et al. A cluster randomized controlled trial evaluating the effectiveness of the school-based drug prevention program #Tamojunto2.0. *Addiction.* 2021;116(6), 1580–1592. https://doi.org/10.1111/add.15358
R*3	Blom V, Drake E, Kallings L., Ekblom M, & Nooijen C. The effects on self-efficacy, motivation and perceived barriers of an intervention targeting physical activity and sedentary behaviours in office workers: a cluster randomized control trial. *BMC Public Health.* 2021;21(1), 1048. https://doi.org/10.1186/s12889-021-11083-2
R*4	Gladstone R, Bojang E, Hart J, et al. Mass drug administration with azithromycin for trachoma elimination and the population structure of Streptococcus pneumoniae in the nasopharynx. *Clinical Microbiology and Infection.* 2021;27(6), 864–870. https://doi.org/10.1016/j.cmi.2020.07.039

R*5	MacPherson P, Lebina L, Motsomi K, et al. Prevalence and risk factors for latent tuberculosis infection among household contacts of index cases in two South African provinces: Analysis of baseline data from a cluster-randomised trial. *PLoS ONE*, 2020;15(3), e0230376. https://doi.org/10.1371/journal.pone.0230376
R*6	Newton N, Teesson M, Mather M, et al. Universal cannabis outcomes from the Climate and Preventure (CAP) study: A cluster randomised controlled trial. *Substance Abuse: Treatment, Prevention, and Policy*. 2018;13(1), 34. https://doi.org/10.1186/s13011-018-0171-4
R*7	Sahlu I, Bauer C, Ganaba R, et al. The impact of imperfect screening tools on measuring the prevalence of epilepsy and headaches in Burkina Faso. *PLoS Neglected Tropical Diseases*, 2019;13(1), e0007109. https://doi.org/10.1371/journal.pntd.0007109
R*8	Shogren K, Hicks T, Burke K, et al. Examining the impact of the SDLMI and whose future is it? Over a two-year period with students with intellectual disability. *American Journal on Intellectual and Developmental Disabilities*, 2020;125(4), 217–229. https://doi.org/10.1352/1944-7558-125.3.217
S*1	Dixon S, Sontrop J, Al-Jaishi A, et al. MyTEMP: Statistical Analysis Plan of a Registry-Based, Cluster-Randomized Clinical Trial. *Canadian Journal of Kidney Health and Disease*, 2021;8. https://doi.org/10.1177/20543581211041182
M*1	Dienes Z, Coulton S, & Heather N. Using Bayes factors to evaluate evidence for no effect: examples from the SIPS project. *Addiction*, 2018;113(2), 240–246. https://doi.org/10.1111/add.14002
M*2	Hox J, Moerbeek M, Kluytmans A, & van de Schoot R. Analyzing indirect effects in cluster randomized trials. The effect of estimation method, number of groups and group sizes on accuracy and power. *Frontiers in Psychology*, 2014;5(FEB), 78. https://doi.org/10.3389/fpsyg.2014.00078
M*3	Moerbeek M. Bayesian evaluation of informative hypotheses in cluster-randomized trials. *Behavior Research Methods*, 2019;51(1), 126–137. https://doi.org/10.3758/s13428-018-1149-x
M*4	Wilson D, Wason J, Brown J, et al. Bayesian design and analysis of external pilot trials for complex interventions. *Statistics in Medicine*, 2021;40(12), 2877–2892. https://doi.org/10.1002/SIM.8941

Amongst the two primary papers, one (R*1) employed Bayesian hierarchical models after encountering problems with fitting frequentist mixed effects models due to singularities (variance components close to zero), and specified non-informative prior distributions. The other (R*2) used Bayes factors to extend the usual frequentist hypothesis testing approach to not only examine the evidence for rejecting the null hypothesis, but also in favour of the null hypothesis itself. In the published statistical analysis plan (S*1), secondary Bayesian analysis was pre-specified to complement the primary frequentist analysis. Specifically, the authors acknowledged that the study was powered to detect a pre-specified effect size, but proposed to run complementary Bayesian analyses to quantify the probability that smaller, potentially clinically meaningful, effect sizes were present. They also outlined plans to explore a range of prior distributions.

A range of Bayesian modelling approaches were employed within the secondary results papers, with Bayesian hierarchical models commonly being used to handle the correlation structure induced by cluster-level randomisation. Two papers (R*3, R*5) used Bayesian hierarchical regression models instead of the more usual frequentist models. A third secondary results paper (R*6) also employed Bayesian hierarchical regression in order to handle the clustering within the data, and used a region of practical equivalence testing framework to extend the typical frequentist hypothesis testing approach to not only assess the evidence against the null, but also in favour of it. Another (R*8) fitted hierarchical models with splines in order to allow for non-linear growth trajectories over time. Adopting a Bayesian approach also allowed the authors to present probabilities for superiority between trial arms. One paper (R*4) used Bayesian methods in order to analyse pneumococcal population structure from data collected from three cross-sectional surveys in one arm of a CRCT. Finally, one paper (R*7) presented a secondary analysis of data collected as part of a trial, which was initially identified and discussed in the first round of literature searching (R1). At baseline, a screening questionnaire for epilepsy and severe headaches was collected during the CRCT, and data from those screening positive were compared to a random sample of those screening negative, as well as from a subsequent physician/neurologist diagnosis. Bayesian latent class models were then fitted to the data to estimate the prevalence of epilepsy and severe headaches, with probabilities of subsequent diagnoses modelled in order to control verification bias.

Weakly informative or non-informative prior distributions were specified in all secondary analysis papers, except R*7 which used informative prior distributions where evidence to inform this was available, and R*4 which did not specify the prior distributions used.

One of the methodological papers discussed Bayes factors (M*1). Although not specifically in the context of CRCTs, the paper used three CRCTs as exemplars for studies in which the use of Bayes factors can add value to frequentist hypothesis testing by not only evaluating evidence against the null hypothesis, but also by evaluating the evidence in favour of the null hypothesis) [Dienes, 2014]. Another methodological paper (M*3) also discussed Bayes factors, but in this case considered them specifically in the context of CRCTs, examining their behaviour under simulation for varying cluster sizes, study sizes and ICCs. A third methodological paper (M*2) explored the use of Bayesian multilevel structural equation models for mediation analysis (i.e. the estimation of indirect treatment effects) in CRCTs. Through a simulation study, the authors concluded that the Bayesian approach performs better than frequentist maximum likelihood estimation, particularly when sample sizes (either number of clusters or number of individuals) are small. The final methodological paper (M*4) considered Bayesian methods to aid in the design and analysis of pilot trials for complex interventions. Specifically, the authors proposed formalising the decision to proceed (or otherwise) from an external pilot study to a definitive trial through Bayesian modelling, by quantifying the probability of design parameters of interest (e.g. adherence) falling into one of three pre-specified decision criteria (green, amber or red). Whilst the methods are not developed or presented specifically for CRCTs, they are nonetheless applicable to such designs, and indeed one of the examples presented is a CRCT.

***

2.5 Discussion

This is the first methodological systematic review of the use, or consideration, of Bayesian methods in CRCTs.

As the number of included papers in the initial review is small, drawing robust conclusions regarding overall reporting quality between the pre-specified subgroups (Table 2.5) is not possible. However, in 2013, Diaz-Ordaz presented a summary of reviews of CRCT quality, in which the percentage of studies accounting for clustering in the sample size calculation and statistical analysis ranged from 0% to 71% and 37% to 92%, respectively [Diaz-Ordaz et al., 2013]. An additional review of reporting and methodological quality of CRCTs was published in 2016 [Tokolahi et al., 2016]. Including the data from the more recent review together with Diaz-Ordaz's summary, the mean (SD) percentage of studies accounting for clustering in the sample size calculation and analysis was 35% (24) and 64% (16), respectively. For comparison, this review identified no primary results papers that clearly accounted for clustering in the sample size calculation, and six (86%) papers that clearly accounted for clustering in the analysis. Although this review included only a small number of papers, reporting quality according to these key metrics may differ somewhat between studies using

Bayesian methodology and the wider pool of CRCTs, as none of the papers identified clearly accounted for clustering in sample size calculation. As such, there may be a need to further improve the reporting of CRCTs utilising Bayesian methodology. Conversely, Bayesian CRCTs seem to more often account for clustering in analysis. This is likely due to the popularity of Bayesian hierarchical modelling within the set of included papers, which is a natural way to account for the clustering induced by cluster-level randomisation. However, it could also be due to the fact that the reports of Bayesian analyses of CRCTs are more recent, once the methodological implications of cluster randomisation were more widely understood.

Evidently, the use of Bayesian methods in the design or analysis of CRCTs remains uncommon relative to the use of frequentist methods, with only eleven primary or secondary results papers reporting doing so up to July 2018, and an additional eight identified up until September 2021. This is despite the increasing use of CRCT designs, with over 120 reported in 2008 alone [Moberg and Kramer, 2015] and the number of PubMed search results rising almost year-on-year since 2006 (Figure 2.2).

Neither this methodological systematic review, nor the subsequent update, identified a single reported CRCT which utilised a Bayesian approach to inform study design or sample size calculation. This is despite some efforts to develop methodology in this area, as highlighted in the methodological aspect of the review. Explaining the reason for this lack of uptake of Bayesian methodology in the design of CRCTs would be little more than speculation. However, possibilities include fundamental disagreements with the approach, still limited development of methodology, inaccessibility of software to implement the methods or a lack of sufficient knowledge or understanding, which may extend beyond the researchers themselves to regulatory bodies, prospective journals and reviewers. Whilst there have been Bayesian methodological developments in both design and analysis of CRCTs, these have been limited in comparison to the development of frequentist methods, which are now well-established in the literature. None of the thirteen published methodological papers initially identified, nor the four identified in the update, appear to have developed publicly available software in order to aid implementation (although some papers reported that code is available from the authors on request). On the other hand, frequentist analysis and sample size calculations for CRCTs can be conducted with relative ease in standard statistical software such as Stata. As such, there is need to increase the availability and accessibility of these Bayesian methods, which have the potential to offer advantages over the frequentist approach within the context of CRCTs.

A common criticism of the Bayesian approach in general, and in particular within the analysis of clinical trial data, is the subjective nature of the choice of prior distribution, although it is recommended that sensitivity analyses be performed in order to assess the strength of the effect of the prior [Spiegelhalter et al., 2004]. Interestingly, however, only two of the 11 results papers that were initially identified, and two of those identified in the update (including the published statistical analysis plan) utilised an informative prior distribution. Six of the initial results papers specified non-informative or weakly informative prior distributions (of which one employed two models), as did five of those identified in the update. One paper identified in the updated search calculated Bayes factors, and so informative priors were required in order to make the calculation. The remaining papers (four from the initial review, and one from the updated review) did not report their choice of prior and therefore likely used an uncontroversial, uninformative prior formulation by default. As such, the majority of the papers identified circumvented the common criticism of the introduction of subjectivity through the specification of prior distributions. Despite this, the use of a well-justified, informative prior distribution has the potential to add significant value to a statistical analysis, and indeed could facilitate more efficient CRCT study design. As a result, there is opportunity for methodological development to specify informative yet rigorous and non-subjective prior specifications for CRCTs, which may enhance the uptake of Bayesian methods in this area, whilst improving the efficiency of trial design and robustness of conclusions from statistical analysis.

Following the outbreak of the Coronavirus Disease 2019 (COVID-19) pandemic in early 2020, trials of interventions of vaccines and treatments for COVID-19 were designed and opened to recruitment at an unprecedented pace. Many of these involved the use of Bayesian methods in both their design and analysis, predominantly through the use of adaptive methods in order to answer a multitude of research questions as expeditiously as possible. High profile examples include REMAP-CAP [Gordon et al., 2021] and PRINCIPLE [Yu et al., 2021], both of which rely on posterior distributions obtained through Bayesian analysis to implement (pre-specified) changes to trial design during trial delivery, such as amendments to the randomisation probabilities (response adaptive randomisation) or the early termination of trial arms due to efficacy or futility in multi-arm multi-stage platform trials [Pallmann et al., 2018]. There was a clear increase in the number of references identified during the searches for the review update compared to the initial review relative to the time periods covered, and this was likely at least in part due to the use of Bayesian adaptive methods in COVID-19 trials. However, this increase was not reflected in the overall number of papers remaining after sifting as all identified Bayesian COVID-19 trials were individually randomised, and indeed no Bayesian CRCTs addressing COVID-19 were found. Whether this is because few trials in COVID-19 use a CRCT design, or there remains hesitancy to implement Bayesian methods in CRCTs remains unclear. However, a recent example of a COVID-19 adaptive platform CRCT, PROTECT [Nanni et al., 2020], opted not to use Bayesian methods to implement the adaptations, despite the amenability of such methods to these designs. Coupled with the fact that no methodological papers focusing on adaptive designs in CRCTs were identified, there is a clear suggestion of a methodological gap in this area, which, if filled, could result in an increased uptake in such methods in the future.

2.5.1 Strengths and Limitations

A protocol for this methodological systematic review was published before commencement of the electronic search [Jones, 2018] (Appendix A) and the review was conducted according to the PRISMA guidelines [Moher et al., 2009]. The electronic search strategy to identify Bayesian approaches in CRCTs was adapted from a previously published strategy, which was demonstrated to have high precision [Taljaard et al., 2010] in identifying CRCTs. In the main part of this review, each stage of the reference sifting and data extraction process was undertaken twice, independently, to ensure accurate inclusion of references and high quality data for examination. Data extraction forms for primary (Appendix B) and secondary (Appendix C) results papers were used in order to aid in the accurate and consistent collection of data. Furthermore, the final data extraction for the main part of the review was agreed by all four members of the systematic review study team.

The reporting quality metrics collected are predominantly a subset of the CONSORT checklist for CRCTs, a well-accepted set of criteria. A small number of additional items were extracted, such as whether cluster size variability had been accounted for in the presented sample size calculation and whether p-values for baseline comparisons were provided, in order to facilitate a robust judgement of reporting quality.

Despite the rigour of the search strategy, it is important to acknowledge the possibility that some relevant publications may have been missed. In particular, the search strategy prioritised specificity, rather than sensitivity, in order to make the sifting process more manageable with limited resource. Six additional methodological papers were identified through additional informal searching in the initial review, and three in the 2021 update. One of the methodological papers identified in the update was published in 2014, and was therefore missed from the original review, highlighting the risk of missing literature, and methodological literature in particular. This is perhaps unsurprising given the search strategy was developed to identify trial results papers rather than methodological papers. No additional results papers were identified through informal searching, suggesting that the search strategy performed well in identifying these relevant papers.

Reporting quality metrics are also presented by journal endorsement of the CONSORT guidelines. However, the journals' guidelines may, in some cases, have changed since the date of the associated publications, and a journal's endorsement may have been intensified since the included papers were accepted for publication. To the best of the author's knowledge, this issue has not been raised in previous systematic reviews of trial reporting quality; archiving of journal guidelines would help researchers conducting quality assessment systematic reviews in the future. Similarly, author affiliations were sought during data collection, but again these may have changed since publication of the research, particularly for papers published some time ago.

The intention, as outlined in the review protocol (Appendix A), was to summarise the pre-specified reporting quality metrics by time periods (pre-2005, 2005 – 2012 and 2012 – 2018) according to publication date to assess the effect of the relevant CONSORT statements on reporting quality. However, it is possible that the time delay between completion of the study and submission/acceptance of the final report for publication may have resulted in some studies being categorised as published after the publication of the updated CONSORT guidance in 2012, when in fact it was designed, conducted and possibly even analysed before.

***

2.6 Conclusion

The use of Bayesian methods in the statistical analysis of CRCTs is rare, and was not used at the design stage of any of the reviewed studies or in their sample size calculations. Even during the COVID-19 pandemic, which saw a marked increase in the use of Bayesian adaptive methods in RCTs more generally, there was no evidence of any parallel increase within CRCTs, and indeed no Bayesian COVID-19 CRCTs were identified in this systematic review. However, the pandemic, and research efforts to combat it, are ongoing, and so this may yet change in the coming years.

Reporting quality may differ between CRCTs utilising Bayesian methodology compared with previous reviews of CRCT quality, although the number of papers identified in this review is too small to draw robust conclusions.

There have been some developments in Bayesian methodology for CRCTs, but comparatively little in contrast with methods developed within the frequentist paradigm. There is an opportunity and a need for further Bayesian methodological developments in the design and analysis of CRCTs if an increased uptake of such methods is to occur. Three particular areas presenting opportunity for methodological development are: (i) methods for the specification of informative prior distributions, which may facilitate more efficient study design and meaningful analyses; (ii) the development and publication of software packages which may improve accessibility and uptake of Bayesian methods in design and analysis; and (iii) the application of Bayesian adaptive designs to CRCTs.

Chapter 3

Power Priors to Facilitate Borrowing from Pilot Data in Cluster Randomised Controlled Trials

Within this chapter, a novel normalised power prior is proposed to facilitate information borrowing from historical data, when clustering is present in both the current and historical data, is proposed. A method for calculating the normalising constant is outlined. This methodology is applied to data from the Healthy Lifestyles Programme and compared to alternative power prior approaches. Finally, an extensive simulation study is presented, in order to assess the performance of the new methodology relative to more traditional analysis approaches.

***

3.1 Introduction

RECALL that in §1.6 a class of informative, data-driven prior distribution, known as the power prior, was introduced. Two types of power prior were introduced: those in which the degree of information borrowing, controlled through the discounting factor a_0, is fixed; and those in which a_0 is itself a parameter. The former is referred to as the FDPP, shown in Equation (1.15), and the latter is referred to at the NPP, shown in Equation (1.17).

CRCTs are logistically and statistically complex trials to design and deliver, and so preceding pilot or feasibility studies are often useful and therefore common. Typically, data from such pilot/feasibility studies are used only to address pre-specified feasibility objectives, such as estimation of recruitment rates or testing of trial procedures. However, there is an opportunity to consider how informative prior distributions can be constructed based on data collected from these pilot/feasibility studies for use in the design and analysis of a subsequent fully powered definitive CRCT. This chapter focuses on the novel application of the NPP in the analysis of clustered, continuous data of the type collected during execution of a CRCT. Exploration of the use of the NPP, as opposed to the FDPP, was chosen as it has the additional advantage of being entirely data driven. This is because a_0 is itself estimated during analysis as opposed to being pre-specified by the analyst as required in the FDPP method, which has the potential to introduce subjectivity and be more open to criticism. The computational challenges associated with approximating the normalising constant when fitting an NPP were described previously in §1.6. Within this chapter, an approach for handling this computational challenge is outlined and applied.

***

3.2 Calculating a Normalising Constant

From §1.3.1, recall that Bayes' Theorem can be applied in order to obtain the posterior distribution of some parameter of interest, θ, given data D. Mathematically, this can be written as

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)} \tag{3.1}$$

where $P(D|\theta)$ is the likelihood of the data given the parameters, $P(\theta)$ is the prior distribution of θ, and $P(D)$ is the marginal distribution of the data such that

$$P(D) = \int_{\Theta} P(D|\theta) \times P(\theta) d\theta \tag{3.2}$$

The value of the marginal distribution of the data, $P(D)$, is important as it ensures that the integral of the posterior distribution is equal to 1, a necessary condition for a proper probability distribution. In practice, however, $P(D)$ is often disregarded, and Bayes' Theorem is written as

$$P(\theta|D) \propto P(D|\theta) \times P(\theta) \tag{3.3}$$

meaning that the posterior distribution is proportional to the likelihood multiplied by the prior distribution, up to the value of some normalising constant $P(D)$. In practice, $P(D)$ is often a large, high-dimensional, intractable integral. However, as outlined in §1.3.2, the development of MCMC methods from the middle of the twentieth century circumvented the need to explicitly calculate $P(D)$, and instead facilitated direct sampling from posterior distributions as shown in Equation (3.3), thus paving the way to making Bayesian inference a practical, credible alternative to the frequentist approach.

Despite this, there are some scenarios in which calculation of the normalising constant is necessary, including for Bayes Factor Model Comparison [Kass and Raftery, 1995] and Bayesian Model Averaging (BMA) methods [Hoeting et al., 1999]. To meet this need, there are various numerical approximation methods. Four of these methods are outlined and compared by Gronau et al. [Gronau et al., 2017]: (i) The naive monte carlo estimator [Hammersley and Handscomb, 1964]; (ii) importance sampling; (iii) the generalised harmonic mean estimator and (iv) bridge sampling [Meng and Wong, 1996]. After introducing and comparing each approach, Gronau et al. [Gronau et al., 2017] concluded that the bridge sampling estimator is the superior approach because: (i) it has been shown to minimise mean squared error compared to the other three approaches; (ii) it is easier to choose a suitable proposal distribution; and (iii) because of the relative ease of implementation. In further work, Gronau et al. [Gronau et al., 2020] released an R package, `bridgesampling`, which facilitates straightforward implementation of bridge sampling methods. The package works seamlessly with both JAGS and Stan to obtain an approximation of the normalising constant given samples from a posterior distribution. These features are utilised in order to obtain normalising constants for the normalised power prior models outlined in more detail in §3.3 below.

***

3.3 Calculation of $C(a_0)$

In order to fit the NPP shown in Equation (1.17), an approach to determine $C(a_0)$ for any $a_0 \in [0,1]$ must be outlined. Using the bridge sampling method, it is possible to obtain $C(a_0)$ for some *fixed* value of a_0, by modelling the historical data, D_0, discounted by some fixed a_0, whilst specifying uninformative priors, $\pi(\theta)$. Specifically,

$$\pi(\theta|D_0,a_0) = \frac{L(\theta|D_0)^{a_0}\pi(\theta)}{\int_\Theta L(\theta|D_0)^{a_0}\pi(\theta)d\theta}$$

and so, for any value of a_0, bridge sampling can be used to obtain an estimate of the normalising constant, $\int_\Theta L(\theta|D_0)^{a_0}\pi(\theta)d\theta$. In light of this, a method to estimate $\int_\Theta L(\theta|D_0)^{a_0}\pi(\theta)d\theta$ for random a_0 is outlined, in line with that proposed by Carvalho and Ibrahim [Carvalho and Ibrahim, 2021].

1. Choose some Δ to partition the domain of $a_0 \in [0,1]$, in to N equally spaced values of a_0 beginning at Δ such that $N\Delta = 1$ (e.g $\Delta = 0.05$, $N = 20$).

2. For each $a_0^{(i)} = \Delta i, i = 1, \ldots, N$, obtain the posterior $\pi(\theta|D_0, a_0^{(i)})$ and subsequently $C(a_0^{(i)}) = \int_\Theta L(\theta|D_0)^{a_0^{(i)}} \pi(\theta) d\theta$ using bridge sampling.

3. Using $a_0^{(i)}$ and $C(a_0^{(i)})$, fit an appropriate model, $\mathscr{M}$ with $C(a_0)$ as the outcome variable, and a_0 as the explanatory variable.

4. Create a fine grid of possible values of $a_0 \in [0,1]$. Use $\mathscr{M}$ to predict $C(a_0)$ for each value of a_0 in the grid.

5. Fit the NPP from Equation (1.17) using an appropriate MCMC method. At each MCMC iteration, determine the two values of a_0 from the fine grid created in step 4 closest to the current value of a_0 within the MCMC algorithm, and their associated estimates of $C(a_0)$. Using simple linear interpolation between the two identified values in the grid, obtain an estimate for $C(a_0)$ given the current value of a_0 in the MCMC.

Note that this approach requires that the priors used to calculate each $C(a_0^{(i)})$ must remain the same in the final step. If either the historical data, or the specification of the priors, change, steps 1-4 *must* be re-fitted before the final NPP inference can be undertaken.

***

3.4 Power Prior Analysis in Cluster Randomised Controlled Trials

In CRCTs, randomisation occurs at the group (“cluster”) level. This study design has implications for statistical analysis, in which each group must be modelled as a random effect in order to account for the potential correlation between participants within the same cluster.

Consider a two-arm CRCT with a continuous outcome of interest, and denote this outcome $Y_{i,j}$ for participant j within cluster i, $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$ where m is the number of clusters and n_i is the number of participants in cluster i. Let $\mathbf{X}$ represent an $(\sum n_i \times p)$ matrix of outcomes, including a column of 1s to represent the intercept term and $p-1$ columns of additional covariates to be included in the model. Furthermore, let β denote a vector of p covariates associated to the data within $\mathbf{X}$, and let θ denote the parameter for the average treatment effect, and let $z_{i,j}$ be a vector of length $\sum n_i$

containing an indicator for whether participant j in cluster i was allocated to the intervention group (1) or the control group (0). Let $\mathbf{b}$ represent a vector of random effects of length m. Finally, let σ^2 denote the within-cluster variance, and let σ_c^2 represent the between-cluster variance. The linear hierarchical model suitable for analysis of continuous CRCT data, which appropriately accounts for the clustered nature of the data, can be expressed as

$$\begin{aligned} Y_{i,j} &\sim \mathrm{N}(\mathbf{X}\beta + \theta z_{i,j} + b_i, \sigma^2) \\ b_i &\sim \mathrm{N}(0, \sigma_c^2) \end{aligned} \tag{3.4}$$

Denoting the data as D, the likelihood of the model in Equation (3.4) can be expressed as

$$L(\theta,\beta,\mathbf{b},\sigma^2|D) = \prod_{i=1}^{m}\prod_{j=1}^{n_i} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_{i,j} - \mathbf{X}\beta - \theta z_{i,j} - b_i)^2\right] \tag{3.5}$$

In the context of a Bayesian inference, the full posterior distribution can then be expressed as

$$\begin{aligned} &\pi(\beta,\theta,\sigma^2,\sigma_c^2,\mathbf{b}|D) = \\ &\qquad \frac{L(\theta,\beta,\mathbf{b},\sigma^2|D)\pi(b|\sigma_c^2)\pi(\beta)\pi(\theta)\pi(\sigma^2)\pi(\sigma_c^2)}{\int_{\Theta} L(\theta,\beta,\mathbf{b},\sigma^2|D)\pi(b|\sigma_c^2)\pi(\beta)\pi(\theta)\pi(\sigma^2)\pi(\sigma_c^2)d\beta d\theta d\mathbf{b} d\sigma^2 d\sigma_c^2} \\ &\qquad \propto L(\theta,\beta|D)\pi(\mathbf{b}|\sigma_c^2)\pi(\beta)\pi(\theta)\pi(\sigma^2)\pi(\sigma_c^2) \end{aligned} \tag{3.6}$$

where

$$\pi(\mathbf{b}|\sigma_c^2) = \prod_{i=1}^{m} \frac{1}{\sigma_c\sqrt{2\pi}} \exp\left(-\frac{b_i^2}{2\sigma_c^2}\right)$$

and $\pi(\beta)$,$\pi(\theta)$,$\pi(\sigma^2)$ and $\pi(\sigma_c^2)$ are the prior distributions for the covariates parameters, the treatment effect parameter, the within-cluster variance and the between-cluster variance, respectively.

Now, presume that before commencing the study in which the current data, D, was obtained, a pilot study was undertaken, in which the same or similar intervention was delivered to a sample of participants from a similar population, and in which the same data were collected. The data realised from this pilot study is denoted as D_0. Suppose that the outcome in the pilot trial is denoted as $Y_{0\tilde{i},\tilde{j}}$ for participant $\tilde{j}$ (who was not in the main trial), in cluster $\tilde{i}$ (which was not included in the main trial), $\tilde{i} = 1,\ldots,m_0$ and $\tilde{j} = 1,\ldots,n_{0\tilde{i}}$, where m_0 clusters were recruited, and cluster $\tilde{i}$ contained $n_{0\tilde{i}}$ participants. Let $\mathbf{X}_0$ be an $(\sum n_{0\tilde{i}} \times p)$ matrix of *the same* outcomes as $\mathbf{X}$, $z_{0\tilde{i},\tilde{j}}$ be a vector of length

$\sum n_{0\tilde{i}}$ containing an indicator for treatment allocation for participant $\tilde{j}$ in cluster $\tilde{i}$ and $\mathbf{b}_0$ be a vector of random effects of length m_0. Then the likelihood of the historical data is of the same form as Equation (3.5), namely

$$L(\theta,\beta,\mathbf{b}_0,\sigma^2|D_0) = \prod_{\tilde{i}=1}^{m_0}\prod_{\tilde{j}=1}^{n_{0\tilde{i}}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_{0\tilde{i},\tilde{j}} - \mathbf{X_0}\beta - \theta z_{0\tilde{i},\tilde{j}} - b_{0\tilde{i}})^2\right] \tag{3.7}$$

A NPP for a linear hierarchical model is proposed, of the form

$$\begin{aligned}&\pi(\theta,\beta,\sigma^2,\sigma_c^2,\mathbf{b}_0,a_0|D_0) = \\ &\quad \frac{\prod_{\tilde{i}=1}^{m_0}\prod_{\tilde{j}=1}^{n_{0\tilde{i}}}\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y_{0\tilde{i},\tilde{j}} - \mathbf{X_0}\beta - \theta z_{0\tilde{i},\tilde{j}} - b_{0\tilde{i}})^2\right]\right)^{a_0}\pi(\mathbf{b}_0|\sigma_c^2)}{\int_\Theta L(\theta,\beta,\mathbf{b}_0,\sigma^2|D_0)^{a_0}\pi(b_0|\sigma_c^2)\pi(\beta)\pi(\theta)\pi(\sigma^2)\pi(\sigma_c^2)d\beta d\theta d\mathbf{b}_0 d\sigma^2 d\sigma_c^2} \\ &\quad \times \pi(a_0)\pi(\beta)\pi(\theta)\pi(\sigma^2)\pi(\sigma_c^2)\end{aligned} \tag{3.8}$$

where

$$\pi(\mathbf{b}_0|\sigma_c^2) = \prod_{\tilde{i}=1}^{m_0}\frac{1}{\sigma_c\sqrt{2\pi}}\exp\left(-\frac{b_{0\tilde{i}}^2}{2\sigma_c^2}\right),$$

$\pi(\beta)$, $\pi(\theta)$, $\pi(\sigma^2)$ and $\pi(\sigma_c^2)$ are the density functions of typically non-informative priors, and $\pi(a_0)$ is the density function of a Beta$(1,1)$ distribution, which is non-informative and bounded in $[0,1]$. Then the full joint posterior distribution is given by

$$\pi(\beta,\theta,\sigma^2,\sigma_c^2,\mathbf{b},\mathbf{b}_0|D,D_0) \propto L(\theta,\beta,\mathbf{b},\sigma^2|D)\pi(\mathbf{b}|\sigma_c^2)\times\pi(\theta,\beta,\sigma^2,\sigma_c^2,\mathbf{b}_0,a_0|D_0) \tag{3.9}$$

and the denominator in Equation (3.8) is the normalising constant, $C(a_0)$, to be estimated at each iteration of the MCMC algorithm using the procedure in §3.3.

***

3.5 An Example: The Healthy Lifestyles Programme Cluster Randomised Controlled Trial

The HeLP study [Lloyd et al., 2018] was a pragmatic CRCT in which 32 schools were randomised to receive either an obesity prevention intervention, delivered within schools, or the continuation of standard education provision (i.e. “usual care”). In total, 1324 children were randomised, of which 1244 provided primary outcome data of

BMI SDS at baseline and 24 months, with the sample size chosen to provide 90% power to detect a between-group difference in BMI SDS of at least 0.25. Approximately five years prior to publication of the primary results, the results of an external feasibility study were published [Lloyd et al., 2012]. This feasibility study recruited and randomised four schools to receive either the HeLP intervention, or the continuation of standard educational provision, with 202 children taking part, of which 185 provided primary outcome data at baseline and 24 months. This feasibility study, whilst not designed or conducted in order to address the question of intervention effectiveness, reported a mean (95% CI) difference in BMI SD scores between allocated groups of -0.45 $(-1.71$ to $0.81)$ after adjustment for clustering. As the wide confidence interval indicates, there is substantial uncertainty around this estimate, but the result nonetheless signals that the clinically relevant treatment effect (-0.25) may be plausible and was worthy of further investigation through a fully-powered, definitive trial. However, the results of the subsequent fully-powered study found no evidence of a treatment effect induced by the HeLP intervention, reporting a fully-adjusted between group difference (95% CI) of -0.02 $(-0.09$ to $0.05)$. As a result, the HeLP project has provided two high-quality datasets, pertaining to the same intervention and delivered to the same population, but realising substantially different point estimates for the treatment effect (although with overlapping CIs due to substantial uncertainty in the estimation of the treatment effect from the pilot data). These exemplar datasets therefore provided an opportunity to explore the ability of the NPP to effectively incorporate historical data, and to automatically discount said historical data according to the commensurability between the two datasets, when both datasets are clustered.

3.5.1 Formulating the Model

The primary outcome of the HeLP study was BMI SDS at 24 months post randomisation. Using the terminology of §3.4, $Y_{i,j}$ is the change in BMI SDS between baseline and follow-up for participant j in cluster i, $\mathbf{X}$ is an (1244×2) matrix of which the first column contains 1s and the second contains baseline BMI SDS for each participant. β is a vector of length two, pertaining to covariates for the intercept term and baseline BMI SDS, and $z_{i,j}$ is a vector of length 1244 containing an indicator variable pertaining to each participant's allocated treatment group. Finally, θ, the primary parameter of interest, represents the mean between-treatment-group difference in change in BMI SDS between baseline and follow-up, hereafter referred to as the treatment effect. Pilot study participants are similarly represented with 0-subscripts as in §3.4. Specifically, $Y_{0\tilde{i},\tilde{j}}$ is the change in BMI SDS between baseline and follow-up, X_0 is a (185×2) matrix containing a column of 1s and a column of baseline BMI SDS for each participant and $z_{0\tilde{i},\tilde{j}}$ is a binary indicator of treatment group for participant $\tilde{j}$ in cluster $\tilde{i}$. Formulation of the NPP and joint posterior distribution for the HeLP datasets then followed exactly as in Equation (3.4) - Equation (3.9).

Most of the priors were specified as non-informative. Specifically, $\beta, \theta \sim \mathrm{N}(0,5)$ and $a_0 \sim \mathrm{Beta}(1,1)$. The prior distribution for the within-cluster variance was $\sigma \sim \mathrm{Exp}(1)$, reflecting a weakly informative prior as per recommendations [Stan Development Team., 2020]. The between-cluster variance was chosen as $\sigma_c \sim$ Half-Cauchy$(0,0.3)$ in line with Gelman's recommendations for specifying a prior for a hierarchical variance parameter when the number of clusters is small [Gelman, 2006]. Gelman noted that when the number of clusters is small, a more informative prior is necessary to restrict away from infeasibly large values of the between-cluster variance (i.e. reduce the length of the right tail of the posterior distribution of the between-cluster variance). This is of particular importance because in order to implement the procedure for numerical approximation of $C(a_0)$ (§3.3), the model must first be fitted to the pilot data alone (for a range of values of a_0) which, in this example, comprises only four clusters.

3.5.2 Approximating the Normalising Constant

The method outlined in §3.3 was used to approximate the normalising constant for fitting the NPP. To begin, suppose $\Delta = 0.05$ and therefore 20 pairs of values were obtained, $a_0^{(i)}$ and $C(a_0^{(i)}), i = 1, \ldots, 20$, by using bridge sampling to approximate $C(a_0^{(i)})$ from samples of each of the posterior distributions, $\pi(\theta, \beta, \sigma^2, \sigma_c^2, \mathbf{b}_0 | D_0, a_0^{(i)})$.

Next, two alternative approaches to specifying $\mathscr{M}$ were considered. Specifically, $\mathscr{M}_1$ was a simple linear regression, and $\mathscr{M}_2$ was a Generalised Additive Model (GAM) fitted using the R package `mgcv` [Wood, 2004, Wood, 2017] with default settings, namely with thin plate regression splines used as the smoothing basis.

Whilst Figure 3.1 implies that the relationship between a_0 and $C(a_0)$ was *almost* linear, there are subtle suggestions that $\mathscr{M}_2$ better fits the data, particularly towards smaller values of a_0. This is supported by examination of: (i) Akaike's Information Criterion (AIC), where $AIC(\mathscr{M}_1) = 53.2$ and $AIC(\mathscr{M}_2) = -38.1$ (lower values indicate a better model fit), and (ii) the Generalised Cross Validation (GCV) scores, where $GCV(\mathscr{M}_1) = 0.77$ and $GCV(\mathscr{M}_2) = 0.011$. The GCV can be interpreted as an estimate of the mean squared prediction error based on a leave-one-out cross-validation procedure [Clark, 2014], and as such, small values of the GCV indicate a better model fit. As a result, $\mathscr{M}_2$ was the preferred model choice, but implementation using $\mathscr{M}_1$ was also explored for the purposes of a sensitivity analysis.

Finally, an array of values of a_0 of length $10,000$ was constructed, alongside an associated array of predictions for $C(a_0)$, with which to estimate $C(a_0)$ using linear interpolation at each iteration in the MCMC procedure.

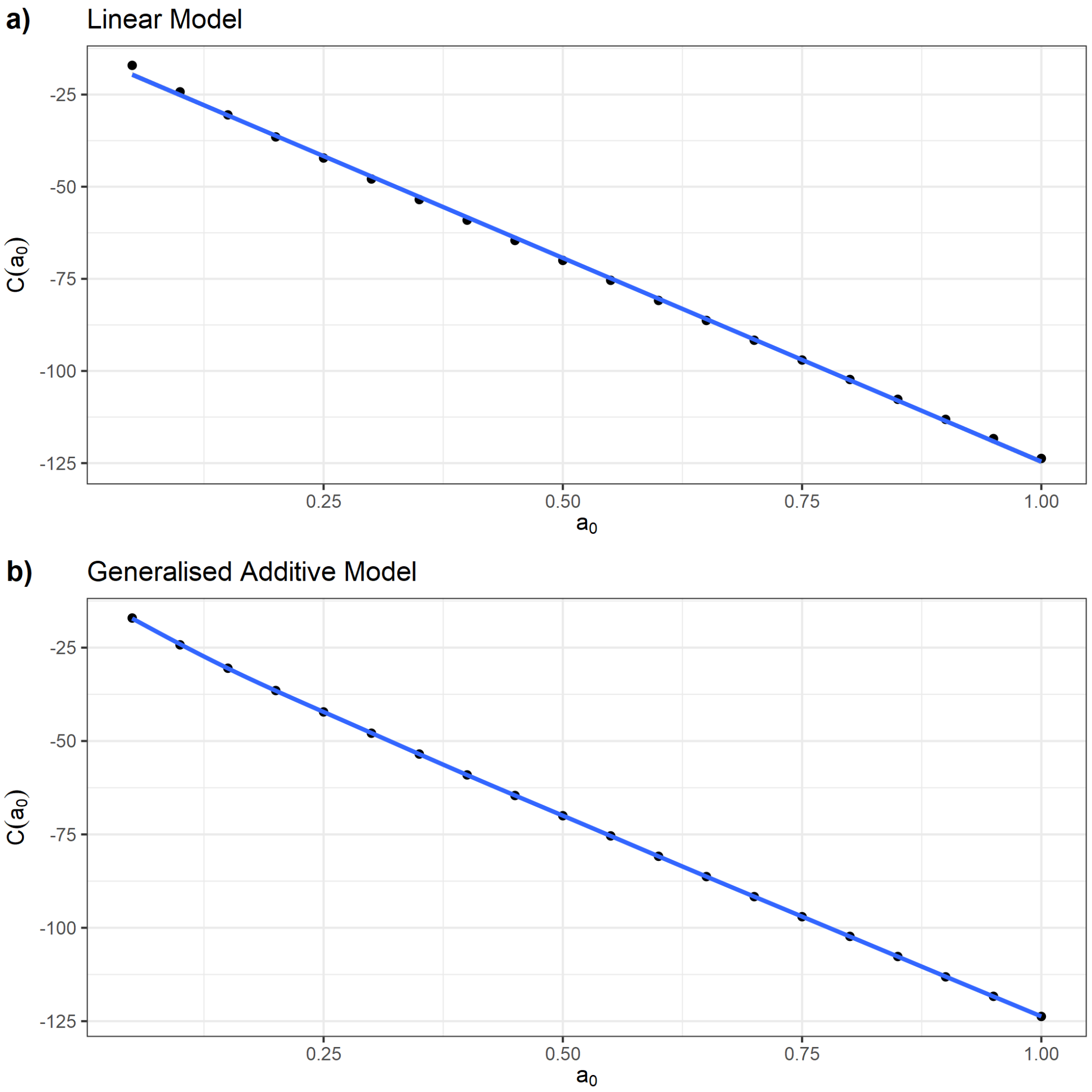


Figure 3.1: A line plot illustrating the relationship between a_0 and $C(a_0)$ for $\mathscr{M}_1$ (a) and $\mathscr{M}_2$ (b), where black dots represent the pre-specified values of a_0 and approximations of $C(a_0)$ obtained from bridge sampling.

3.5.3 Results

Estimates of the treatment effect and the ICC are presented from the following analyses:

1. Analysis of the definitive trial data alone, which is the typical approach to analysis of CRCT data;
2. Analysis of the historical and definitive trial data together, combined via simple, unweighted pooling of the data to create a single dataset;
3. The FDPP from Equation (1.15), for a range of *fixed* a_0. Specifically, $a_0 = 0, 0.2, 0.4, 0.6, 0.8, 1$;
4. The ICPP as outlined in Equation (1.16);
5. The NPP as outlined in Equation (1.17) with $\mathscr{M}_2$ used to generate the array of predictions of $C(a_0)$ ($\mathscr{M}_1$ is considered for the purpose of sensitivity analysis, discussed further in §3.5.4).

For the ICPP and the NPP approaches, a_0 was treated as random and is therefore a parameter to be estimated. Furthermore, let $\hat{a}_0$ denote the median of the posterior distribution of a_0. Under the ICPP approach, $\hat{a}_0 = 0.008$, 95% Credible Interval (CrI): $(0.00020$ to $0.030)$; 95% Highest Posterior Density Interval (HPDI): $(8.057 \times 10^{-7}$ to $0.025)$, which represented a near-complete discounting of the historical data. Conversely, when the normalising constant is approximated and properly accounted for using the NPP approach, $\hat{a}_0 = 0.310$, 95% CrI: $(0.055$ to $0.89)$ 95% HPDI: $(0.005$ to $0.81)$. The latter result seems more reasonable; the historical data were, unsurprisingly, still discounted fairly heavily, but does nonetheless allow a reasonable amount of information to be incorporated into the overall analysis.

Figure 3.2 illustrates the posterior density of a_0 under the ICPP and the NPP approaches. The visualisation further emphasises how concentrated a_0 was towards 0 under the ICPP approach. It can also be seen that estimation of a_0 under the NPP approach had a great deal of uncertainty around it, although this uncertainty is fully accounted for in the final inference on the treatment effect. As a skewed distribution, the importance of presenting not only CrIs, but also HPDIs, is evident.

Treatment Effect

Recall that the pilot trial signalled that the HeLP intervention could plausibly achieve the clinically relevant average reduction in BMI SDS of 0.25, and was worthy of further investigation through a fully-powered, definitive trial. Recall also that the final analysis of the definitive trial data found no evidence of a clinically or statistically significant

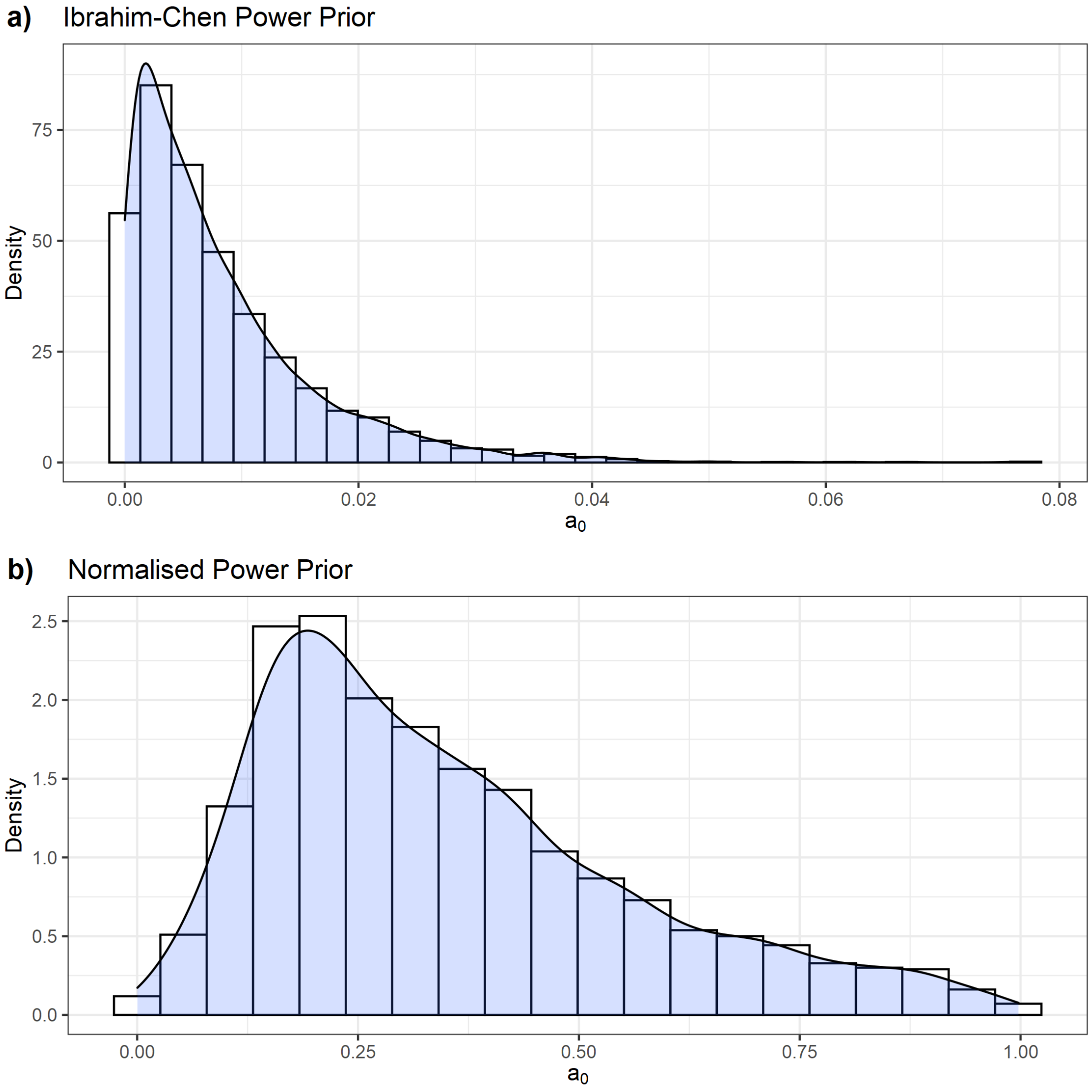


Figure 3.2: Density of the posterior distribution of a_0 when utilising the ICPP (a) and the NPP (b).

treatment effect. It is therefore intuitive to note that the average treatment effect displayed in Table 3.1 increases in magnitude as the value of a_0 increases and a greater degree of information borrowing from the pilot data is facilitated. Note also that $a_0 = 0$ and $a_0 = 1$ pertain to special cases in which the pilot data is entirely excluded from, and completely pooled with, the definitive trial data, respectively. It was for this reason that the results of the analysis of the definitive data alone are the same as that in which $a_0 = 0$, and the results of the pooled analysis are the same as that in which $a_0 = 1$. Furthermore, as the ICPP approach resulted in an estimated a_0 very close to 0, the result of this analysis is also similar to that in which $a_0 = 0$. However, the NPP approach was able to achieve a sensible estimate of a_0 and borrowed information from the historical data accordingly, resulting in an increased average treatment effect.

Table 3.1: Estimation of the Treatment Effect (θ) from the HeLP data

	Model	Treatment effect	95% CrI	$P(\theta > 0)$
1.	Definitive	-0.024	(-0.103, 0.057)	0.264
2.	Pooled	-0.052	(-0.133, 0.029)	0.100
3.	**Fixed $\mathbf{a_0}$**			
(i)	$a_0 = 0$	-0.023	(-0.101, 0.057)	0.278
(ii)	$a_0 = 0.2$	-0.035	(-0.114, 0.045)	0.193
(iii)	$a_0 = 0.4$	-0.042	(-0.129,0.041)	0.152
(iv)	$a_0 = 0.6$	-0.047	(-0.127, 0.037)	0.126
(v)	$a_0 = 0.8$	-0.050	(-0.132, 0.034)	0.118
(vi)	$a_0 = 1$	-0.052	(-0.139, 0.029)	0.108
4.	ICPP $(\hat{a}_0 = 0.006)$	-0.023	(-0.107,0.059)	0.278
5.	NPP $(\hat{a}_0 = 0.31)$	-0.039	(-0.121, 0.040)	0.164

Intracluster Correlation Coefficient

Estimates of the ICC, alongside 95% HPDIs, are provided in Table 3.2. HPDIs are used as they provide a more appropriate summary of skewed distributions such as posterior distribtions of ICCs. Incorporation of the pilot data resulted in an increase in the estimate of the ICC, with the magnitude of this increase growing as the amount of borrowing, according to the discounting parameter, increases.

3.5.4 Sensitivity Analyses

In order to assess the robustness of the proposed NPP methodology (and specifically the method used to approximate $C(a_0)$), the following sensitivity analyses were undertaken:

SA.1 Change Δ, the number of a_0's for which to approximate $C(a_0)$, to:

Table 3.2: Estimation of the ICC (ρ) from the HeLP data

	Model	ICC	95% HPD Interval
1.	Definitive	0.030	(0, 0.059)
2.	Pooled	0.043	(0.014, 0.079)
3.	**Fixed $\mathbf{a_0}$**		
(i)	$a_0 = 0$	0.030	(0.001, 0.062)
(ii)	$a_0 = 0.2$	0.030	(0, 0.060)
(iii)	$a_0 = 0.4$	0.035	(0.006, 0.069)
(iv)	$a_0 = 0.6$	0.038	(0.009, 0.070)
(v)	$a_0 = 0.8$	0.040	(0.012, 0.073)
(vi)	$a_0 = 1$	0.042	(0.012, 0.076)
4.	ICPP $(\hat{a}_0 = 0.006)$	0.030	(0, 0.059)
5.	NPP $(\hat{a}_0 = 0.31)$	0.033	(0.003, 0.066)

SA.1.1 $\Delta = 10$

SA.1.2 $\Delta = 40$

SA.2 Use linear regression, $\mathscr{M}_1$, to predict the grid of values of $C(a_0)$

SA.3 Change the length of arrays of a_0, to:

SA.3.1 1000

SA.3.2 $100,000$

SA.4 Use the "Warp-III" proposal distribution for the bridge sampler, instead of the normal proposal distribution.

Table 3.3: Results of the main NPP analysis, alongisde six sensitivity analyses of the HeLP data using the NPP

Model	$\hat{a}_0$	95% CrI	95% HPDI	Parameter[a]	Estimate	95% Interval[b]	$P(\theta > 0)$
NPP	0.310	(0.055, 0.886)	(0.005, 0.810)	θ	-0.039	(-0.121, 0.040)	0.164
				ρ	0.033	(0.003, 0.066)	
SA.1.1	0.286	(0.018,0.872)	(0, 0.786)	θ	-0.038	(-0.122,0.043)	0.176
				ρ	0.032	(0.004, 0.067)	
SA.1.2	0.312	(0.085,0.902)	(0.059, 0.836)	θ	-0.042	(-0.125,0.039)	0.155
				ρ	0.033	(0.004, 0.066)	
SA.2	0.107	(0.004,0.892)	(0, 0.805)	θ	-0.033	(-0.115,0.049)	0.212
				ρ	0.032	(0.002,0.064)	
SA.3.1	0.299	(0.066,0.882)	(0.026, 0.783)	θ	-0.038	(-0.12,0.043)	0.168
				ρ	0.033	(0.004, 0.067)	
SA.3.2	0.304	(0.061,0.873)	(0.028, 0.811)	θ	-0.039	(-0.119,0.044)	0.168
				ρ	0.033	(0.005, 0.067)	
SA.4	0.308	(0.057,0.873)	(0.021, 0.815)	θ	-0.039	(-0.123,0.041)	0.169
				ρ	0.033	(0.005, 0.065)	

[a] θ is the treatment effect; ρ is the ICC

[b] For θ, interval derived from the 2.5% and 97.5% quantiles; for ρ, Highest Posterior Density interval is shown

Table 3.3 shows the estimates of the discounting factor, the treatment effect, and the ICC, for each of the sensitivity analyses, alongside appropriate 95% intervals.

SA.1.1 involved reducing the number of $C(a_0)$ values calculated using Bridge Sampling to inform the later predictive model, $\mathscr{M}$. It is not surprising, therefore, that reducing the "sample size" for fitting this model changed the results, although only slightly and mainly with regard to estimation of a_0 rather than the treatment effect. It is encouraging that increasing from $\Delta = 20$ to $\Delta = 40$ as in analysis SA.1.2 did not change the results, suggesting that $\Delta = 20$ is adequate, but that reducing below this, as in SA.1.1, may have impacted the final inferences.

In SA.2, $\mathscr{M}_1$ (linear regression), instead of $\mathscr{M}_2$ (a generalised additive model), was used within the procedure for approximation of $C(a_0)$. As can be seen in Table 3.3, there was a notable decrease in the estimated discounting factor, a_0, compared to the primary analysis, resulting in a greater discounting of information from the historical data. As shown in §3.5.2, $\mathscr{M}_2$ is the superior model, and is likely to outperform linear regression in the majority of cases as it is able to handle non-linear relationships, and so this contrast between results is expected.

In SA.3.1 and SA.3.2, the length of the array used to fit the predictive model, $\mathscr{M}$, was altered. It may be expected that the use of more values (i.e. a finer array) would result in an improvement in model fit, and therefore impact the final inference. However, the results of the sensitivity analysis do not appear to indicate that there are any differences between either inference and the results of the primary analysis.

SA.4 involves using the "Warp-III" proposal distribution instead of the normal proposal distribution within the bridge sampling procedure, the details of which are provided in [Gronau et al., 2019, Meng and Schilling, 2002]. It can be seen in Table 3.3 that the results of SA.4 remain robust to the results of the primary analysis.

The results of the sensitivity analyses shown in Table 3.3 demonstrate the robustness of the proposed NPP method, with the exceptions of SA.1.1 and SA.2 as discussed above, in terms of estimation of the treatment effect, the ICC and the discounting factor, a_0. As a result, it is evident that the choices (i.e. $\Delta = 20$, a generalised additive model for $\mathscr{M}$, an array of length $10,000$, and a normal proposal distribution for the bridge sampling procedure) specified for the primary analyses are reasonable and appropriate.

***

3.6 Simulation Study

3.6.1 Design

A simulation study was designed and conducted in order to: (i) confirm that the proposed NPP method for CRCT data estimates the discounting factor, a_0, appropriately (i.e. discounts data more, on average, when the differences between the data generating mechanisms are larger); (ii) explore whether the NPP method can improve estimation of the treatment effect and (iii) explore whether the NPP method can improve estimation of the ICC, a key metric to inform the design of future CRCTs.

The simulation study involved generating two datasets at each iteration: the pilot data (with which to construct the power prior), and the definitive trial data. For simplicity, it was assumed that analyses did not include adjustment for additional covariates whilst acknowledging that, in practice, key pre-specified variables are often included, but rarely accounted for at the study design stage, as adjustments tend only to improve the precision of treatment effect estimates. In addition, a within-group standard deviation of 1 was assumed. The between-cluster standard deviation, σ_c, was varied in order to reflect the strength of clustering within the data. This simulation study only considered continuous outcome data.

The data generating mechanism for the definitive trial data was

$$\begin{aligned} Y_{i,j} &\sim \mathrm{N}(\beta + \theta x_{i,j} + b_i, 1^2) \\ b_i &\sim \mathrm{N}(0, \sigma_c^2) \end{aligned} \tag{3.10}$$

and similarly for the pilot trial data was

$$\begin{aligned} Y_{0\tilde{i},\tilde{j}} &\sim \mathrm{N}(\beta + \theta_0 x_{0\tilde{i},\tilde{j}} + b_{0\tilde{i}}, 1^2) \\ b_{0\tilde{i}} &\sim \mathrm{N}(0, \widetilde{\sigma_c^2}) \end{aligned} \tag{3.11}$$

where β was the intercept term, θ and θ_0 were the treatment effects for the definitive and pilot trials, respectively, and the b_j and $b_{0\tilde{j}}$ were the cluster-level random effects for the definitive and pilot trials, respectively. The ICC is a measure of the degree of clustering in the data, and is the proportion of the overall variance which can be attributed to the cluster level variance σ_c^2. Formally,

$$\mathrm{ICC} = \frac{\text{Between-cluster variance}}{\text{Between-cluster Variance} + \text{Within-cluster variance}} \tag{3.12}$$

The ICC for the definitive data was denoted as ρ, and for the pilot data as ρ_0 (and note the unit variance for the within-cluster standard deviation), so that

$$\rho = \frac{\sigma_c^2}{\sigma_c^2 + 1}$$

and

$$\rho_0 = \frac{\tilde{\sigma}_c^2}{\tilde{\sigma}_c^2 + 1}$$

The Intracluster Correlation Coefficient

Adams et al. [Adams et al., 2004] conducted a reanalysis of 31 CRCTs in order to estimate ICCs, finding that the median of the 1039 unadjusted ICCs calculated was 0.01 with 5 and 95 percentiles of 0 and 0.095, respectively. Therefore in order to reflect a reasonable range of ICCs which may be encountered in practice, ICCs of 0.01, 0.05 and 0.1 for both the pilot and definitive trial data were considered, which according to Equation (3.12) when $\sigma = 1$, correspond to values of σ_c of 0.101, 0.229 and 0.333, respectively.

The Treatment Effect

Small and medium treatment effects of 0.2 and 0.4, respectively, were used to simulate the data for both the pilot and definitive trial datasets. Larger treatment effects were not considered, as the number of clusters required to achieve the desired level of power becomes too few to make a cluster randomised design appropriate, as discussed further below.

Sample Size and Number of Clusters

Let N and N_0 denote the total sample sizes for the definitive and pilot data, respectively, and let k and k_0 denote the number of clusters per arm in the definitive and pilot studies, respectively.

Sample size calculations for definitive CRCTs are possible using closed formulae [Rutterford et al., 2015], based on assumptions for the variability of the outcome (σ^2), the ICC, and the cluster size, m, in order to detect a pre-specified minimum effect size with a pre-specified level of power, usually 80% or 90%. In practice, cluster sizes are rarely fixed, and recently the importance of allowing for cluster size variability in sample size calculation has been highlighted [Eldridge et al., 2006]. However, for simplicity, a fixed cluster size (i.e. no variability in cluster size) of $m = 15$ was assumed for all simulations of both pilot and definitive trial data.

For the definitive trial data, k was chosen in each case to achieve 85% power to detect the pre-specified treatment effect. 85% was targeted in order to broadly reflect standard practice in the design of clinical trials (typically powered at 80% or 90%) whilst allowing identification of potential improvements in power as a result of using the NPP to incorporate the pilot data. Sample size requirements to achieve 85% power, according to standard, closed-formula frequentist methods, for the different combinations of the ICC and treatment effect, are shown in Table 3.4, and were used to calculate the number of clusters per arm simulated for the definitive trial datasets within this simulation study (k).

Table 3.4: Sample size requirements to achieve 85% power, assuming $\sigma = 1$ and $m = 15$

ICC	Treatment Effect	Total Sample Size	Number of Clusters per arm
0.01	0.2	1080	36
0.05	0.2	1560	52
0.1	0.2	2190	73
0.01	0.4	300	10
0.05	0.4	420	14
0.1	0.4	570	19

It is not appropriate to specify the size of the simulated pilot data (k_0) on the basis of statistical power. The purpose of a pilot or feasibility study (both cluster randomised and individually randomised) is not to determine whether a treatment effect exists, and as such the design, analysis and reporting should be reflective of this [Eldridge et al., 2016]. As a result, justification of the pre-specified sample size for pilot and feasibility studies is less straightforward, and can vary depending on the purpose of the study. For example, to address logistical uncertainties surrounding feasibility, to estimate recruitment or retention rates, or to estimate parameters to inform a sample size calculation for a subsequent definitive trial, although Eldridge et al. [Eldridge et al., 2015] advised against relying on the results of pilot or feasibility CRCTs alone to inform sample size calculations, and also acknowledged that in most cases, it would be impossible to obtain precise estimates of rates of interest (e.g. recruitment, retention). Eldridge et al. [Eldridge et al., 2015], in order to inform discussion around how large a pilot or feasibility CRCT should be, conducted a small review of pilot and feasibility CRCTs, and found that the number of clusters analysed ranged from three to 29. In order to broadly reflect this finding, both “small” ($k_0 = 4$) and “large” ($k_0 = 8$) pilot study data were simulated.

Scenarios

A total of 72 scenarios were simulated, where the ICCs, ρ and ρ_0, the study sizes, k and k_0, and the effect sizes, θ and θ_0 were all varied sequentially. One subset of 36 scenarios included small pilot studies ($k_0 = 4$), and the other subset of 36 included large pilot studies ($k_0 = 8$). Within each of these subsets, the remaining design parameters were varied, with further details of these parameters within each scenario outlined in Table 3.5.

Within each scenario, five analyses were undertaken: analysis of the definitive trial data alone using a Bayesian and a frequentist model; analysis of the pooled definitive and pilot trial data using a Bayesian and a frequentist model; and analysis using the NPP approach.

Table 3.5: Simulation study scenarios

Scenario	k_0	θ_0	θ	ρ_0	ρ	k	$\tilde{\sigma}_c$	σ_c
1.1.1	4	0.2	0.2	0.01	0.01	36	0.101	0.101
1.1.2	4	0.2	0.2	0.01	0.05	52	0.101	0.229
1.1.3	4	0.2	0.2	0.01	0.1	73	0.101	0.333
1.1.4	4	0.2	0.2	0.05	0.01	36	0.229	0.101
1.1.5	4	0.2	0.2	0.05	0.05	52	0.229	0.229
1.1.6	4	0.2	0.2	0.05	0.1	73	0.229	0.333
1.1.7	4	0.2	0.2	0.1	0.01	36	0.333	0.101
1.1.8	4	0.2	0.2	0.1	0.05	52	0.333	0.229
1.1.9	4	0.2	0.2	0.1	0.1	73	0.333	0.333
1.2.1	4	0.2	0.4	0.01	0.01	10	0.101	0.101
1.2.2	4	0.2	0.4	0.01	0.05	14	0.101	0.229
1.2.3	4	0.2	0.4	0.01	0.1	19	0.101	0.333
1.2.4	4	0.2	0.4	0.05	0.01	10	0.229	0.101
1.2.5	4	0.2	0.4	0.05	0.05	14	0.229	0.229
1.2.6	4	0.2	0.4	0.05	0.1	19	0.229	0.333
1.2.7	4	0.2	0.4	0.1	0.01	10	0.333	0.101
1.2.8	4	0.2	0.4	0.1	0.05	14	0.333	0.229
1.2.9	4	0.2	0.4	0.1	0.1	19	0.333	0.333
1.3.1	4	0.4	0.2	0.01	0.01	36	0.101	0.101
1.3.2	4	0.4	0.2	0.01	0.05	52	0.101	0.229
1.3.3	4	0.4	0.2	0.01	0.1	73	0.101	0.333
1.3.4	4	0.4	0.2	0.05	0.01	36	0.229	0.101
1.3.5	4	0.4	0.2	0.05	0.05	52	0.229	0.229
1.3.6	4	0.4	0.2	0.05	0.1	73	0.229	0.333
1.3.7	4	0.4	0.2	0.1	0.01	36	0.333	0.101
1.3.8	4	0.4	0.2	0.1	0.05	52	0.333	0.229
1.3.9	4	0.4	0.2	0.1	0.1	73	0.333	0.333
1.4.1	4	0.4	0.4	0.01	0.01	10	0.101	0.101
1.4.2	4	0.4	0.4	0.01	0.05	14	0.101	0.229
1.4.3	4	0.4	0.4	0.01	0.1	19	0.101	0.333
1.4.4	4	0.4	0.4	0.05	0.01	10	0.229	0.101
1.4.5	4	0.4	0.4	0.05	0.05	14	0.229	0.229
1.4.6	4	0.4	0.4	0.05	0.1	19	0.229	0.333
1.4.7	4	0.4	0.4	0.1	0.01	10	0.333	0.101

Table 3.5: Simulation study scenarios (continued)

Scenario	k_0	θ_0	θ	ρ_0	ρ	k	$\tilde{\sigma}_c$	σ_c
1.4.8	4	0.4	0.4	0.1	0.05	14	0.333	0.229
1.4.9	4	0.4	0.4	0.1	0.1	19	0.333	0.333
2.1.1	8	0.2	0.2	0.01	0.01	36	0.101	0.101
2.1.2	8	0.2	0.2	0.01	0.05	52	0.101	0.229
2.1.3	8	0.2	0.2	0.01	0.1	73	0.101	0.333
2.1.4	8	0.2	0.2	0.05	0.01	36	0.229	0.101
2.1.5	8	0.2	0.2	0.05	0.05	52	0.229	0.229
2.1.6	8	0.2	0.2	0.05	0.1	73	0.229	0.333
2.1.7	8	0.2	0.2	0.1	0.01	36	0.333	0.101
2.1.8	8	0.2	0.2	0.1	0.05	52	0.333	0.229
2.1.9	8	0.2	0.2	0.1	0.1	73	0.333	0.333
2.2.1	8	0.2	0.4	0.01	0.01	10	0.101	0.101
2.2.2	8	0.2	0.4	0.01	0.05	14	0.101	0.229
2.2.3	8	0.2	0.4	0.01	0.1	19	0.101	0.333
2.2.4	8	0.2	0.4	0.05	0.01	10	0.229	0.101
2.2.5	8	0.2	0.4	0.05	0.05	14	0.229	0.229
2.2.6	8	0.2	0.4	0.05	0.1	19	0.229	0.333
2.2.7	8	0.2	0.4	0.1	0.01	10	0.333	0.101
2.2.8	8	0.2	0.4	0.1	0.05	14	0.333	0.229
2.2.9	8	0.2	0.4	0.1	0.1	19	0.333	0.333
2.3.1	8	0.4	0.2	0.01	0.01	36	0.101	0.101
2.3.2	8	0.4	0.2	0.01	0.05	52	0.101	0.229
2.3.3	8	0.4	0.2	0.01	0.1	73	0.101	0.333
2.3.4	8	0.4	0.2	0.05	0.01	36	0.229	0.101
2.3.5	8	0.4	0.2	0.05	0.05	52	0.229	0.229
2.3.6	8	0.4	0.2	0.05	0.1	73	0.229	0.333
2.3.7	8	0.4	0.2	0.1	0.01	36	0.333	0.101
2.3.8	8	0.4	0.2	0.1	0.05	52	0.333	0.229
2.3.9	8	0.4	0.2	0.1	0.1	73	0.333	0.333
2.4.1	8	0.4	0.4	0.01	0.01	10	0.101	0.101
2.4.2	8	0.4	0.4	0.01	0.05	14	0.101	0.229
2.4.3	8	0.4	0.4	0.01	0.1	19	0.101	0.333
2.4.4	8	0.4	0.4	0.05	0.01	10	0.229	0.101
2.4.5	8	0.4	0.4	0.05	0.05	14	0.229	0.229

Table 3.5: Simulation study scenarios (continued)

Scenario	k_0	θ_0	θ	ρ_0	ρ	k	$\tilde{\sigma}_c$	σ_c
2.4.6	8	0.4	0.4	0.05	0.1	19	0.229	0.333
2.4.7	8	0.4	0.4	0.1	0.05	10	0.333	0.101
2.4.8	8	0.4	0.4	0.1	0.05	14	0.333	0.229
2.4.9	8	0.4	0.4	0.1	0.1	19	0.333	0.333

3.6.2 Model Formulation and Posterior Sampling

Frequentist linear hierarchical models for the pooled and hierarchical data were fitted using the `lme4` [Bates et al., 2015] package within R.

For the simple linear Bayesian hierarchical model used to analyse the pooled and definitive data, as well as the NPP model, the prior distributions for the parameters were specified as: $\beta, \theta \sim$ N$(0,5)$; $\sigma \sim$ Exp(1) and $\sigma_c \sim$ Half-Normal$(0,1.5)$, as recommended by Gelman [Gelman, 2006]. For the NPP, the prior for the discounting factor was non-informative. Specifically, $a_0 \sim$ Beta$(1,1)$. The simple hierarchical models were run for a total of 2000 iterations across four chains, including a warmup period of 1000 iterations per chain. In acknowledgement of the additional complexity and to ensure sufficient posterior samples for reliable inference, the NPP models were run for 3000 iterations across four chains, including a warmup period of 1000 iterations. All Bayesian models were fitted using the probabilistic programming language Stan [Carpenter et al., 2017], which samples from the posterior distribution using HMC methods. For the approximation of $C(a_0)$, the method for which is outlined in §3.3, $\Delta = 20$ values of a_0 were used, and each model was run for 3000 iterations, across two chains including a warmup of 1500 iterations per chain. The number of chains was reduced in order to ease computational burden. Parallel computing techniques were utilised within R using the `parallel` package [R Core Team, 2019] and the University of Plymouth High Peformance Computing cluster. Specifically parallelisation occured at: (i) the scenario level, where each scenario is run simultaneously; (ii) during the calculation of $C(a_0)$, where $\Delta = 20$ values of a_0 were used and (iii) to run each chain of each MCMC procedure concurrently.

Typically, a key element of a Bayesian workflow is examination of diagnostic plots to ensure convergence of the Monte Carlo simulation approaches used, but this is not feasible in a simulation study, as it would require manual inspection at each iteration. However, some diagnostic information was captured at each iteration. Specifically, $\hat{R}$, and the effective sample size, N_{eff}. $\hat{R}$ is a measure that uses multiple chains to determine convergence, where values < 1.1 indicate that a model has converged [Vehtari et al., 2021]. Given N dependent samples, the effective sample size, N_{eff}, is the number of independent samples with the same estimation power as the N autocorrelated samples. An effective sample size of at least 400 (100 per chain) for each model [Vehtari et al., 2021] is required. Any results obtained from an MCMC procedure in which $\hat{R} \geq 1.1$ or $N_{eff} < 400$ were discarded. A further indication of a well fitted, properly converged model in Stan is an absence of *divergent transitions*, which occur when the MCMC sampler encounters numerical instability [Carpenter et al., 2017]. Encountering divergent transitions is inevitable when simulating a large number of datasets. For each Bayesian model, the number of post-warmup divergent transitions was recorded and, in a similar manner to Fuglstad et al. [Fuglstad et al., 2020], any iterations in which at least 0.1% (i.e. a total of 8 or more) of the post-warmup iterations were divergent transitions were discarded. Similarly, any iterations in which the frequentist models failed to converge were also discarded.

3.6.3 Outcome Measures

Discounting Factor

At each iteration, the mean and median of the posterior distribution of a_0 was captured. In addition, the upper and lower 95% credible intervals (2.5 and 97.5 percentiles) were captured, as well as the 95% HPDIs in order to account for the possibility of a skewed posterior density.

Treatment Effect

The primary value of interest within most RCTs is the average treatment effect. The simulation study explored whether the NPP has the potential to improve estimation of the average treatment effect, through a reduction in bias or mean squared error (MSE) with regards to the mean of the posterior treatment effect. The 95% credible intervals for the treatment effect within the Bayesian models, and the 95% confidence intervals for the frequentist models, were captured, and the power, coverage and interval width derived, as shown in Table 3.6.

Intracluster Correlation Coefficient

Whilst not of primary interest in the context of CRCT analysis, the ICC can provide useful insight into the degree of clustering within a population, and provides valuable evidence to inform study design for future research. As such, the effect of the NPP approach in estimating the ICC was examined, in comparison to estimating the ICC based on the analysis of the definitive data alone, or the pooled data. In order to account for the likely skewed shape of the posterior distribution of the ICC, medians and 95% HPDIs were obtained. The performance measures shown in Table 3.6 were again derived in order to assess the performance of the proposed NPP approach (with the exception of power, which is not an appropriate metric to consider in the context of ICC estimation). For the frequentist models, 95% confidence intervals for the ICCs were calculated using Swiger's method [Ukoumunne, 2002], with truncation of the lower limit at 0 if required.

Table 3.6: A table of performance measures with formulae

Performance Measure	Formula
Properties of the estimator	
Bias	$\frac{1}{n_{sims}}\sum_{i=1}^{n_{sims}}(\hat{\theta}_i - \theta)$
Mean Squared Error	$\frac{1}{n_{sims}}\sum_{i=1}^{n_{sims}}(\hat{\theta}_i - \theta)^2$
Empirical Standard Error	$\sqrt{\text{Var}(\hat{\theta})}$
Properties of the 95% interval	
Power	$\frac{1}{n_{sims}}\sum_{i=1}^{n_{sims}}\mathbb{1}_{\hat{\theta}_{i,0.025}>0 \text{ or } \hat{\theta}_{i,0.975}<0}$
Coverage	$\frac{1}{n_{sims}}\sum_{i=1}^{n_{sims}}\mathbb{1}_{\hat{\theta}_{i,0.025}\leq\theta\leq\hat{\theta}_{i,0.975}}$
Width of CI	$\frac{1}{n_{sims}}\sum_{i=1}^{n_{sims}}(\hat{\theta}_{i,0.975} - \hat{\theta}_{i,0.025})$

3.6.4 Results

A full table of results from the simulation study, including each of the performance measures, and the summary statistics for a_0, can be found in Appendix E.

Estimating a_0

In order to assess whether the NPP approach can effectively distinguish the degree of commensurability between the pilot and definitive data, summary statistics pertaining to the posterior distribution of a_0 are illustrated in Figures 3.3 - 3.6.

Figure 3.3 shows, for each scenario, the mean of the point estimates (median) of each posterior distribution of a_0 calculated at each iteration of the simulation study, alongside the corresponding 95% Credible Intervals (CrIs) and overlaid with the density of the median values of a_0 estimated at each iteration. Only the distribution of the *point estimates* for a_0 across each iteration of the simulation study are illustrated in Figure 3.3, rather than the *entirety* of all posterior distributions. Point estimates ranged from approximately 0.7, where the data generating mechanisms underpinning the pilot and definitive data were similar (e.g. scenario 2.1.9), to approximately 0.4, where the data generating mechanisms became increasingly contradictory (e.g. scenario 2.3.7). In particular, scenarios in which the degree of clustering in the pilot data was greater than the degree of clustering in the definitive trial data (scenarios 1.1.7, 2.1.7, 1.2.7, 2.2.7, 1.3.7, 2.3.7, 1.4.7, 2.4.7) appear to be subject to particularly heavy discounting, as can be seen by the low point estimates of a_0 for these scenarios. There also appears to be a greater deal of uncertainty around the discounting factors associated with these scenarios, as indicated by the wider 95% CrIs and heavier-tailed posterior densities.

Figure 3.4 illustrates key summary statistics for a_0 with dichotomisation of each scenario into one in which the treatment effects for the underlying data generating mechanisms are the same (i.e. $\theta = \theta_0$), shown in grey, versus different (i.e. $\theta \neq \theta_0$), shown in blue. Each point in the boxplot represents the average (mean) of the summary statistic for a_0 under consideration, across the iterations within the simulation study, for each scenario. It can be seen that both the mean and the median values of a_0 were larger amongst the scenarios in which the treatment effects were the same, indicating a greater degree of information borrowing in these cases. Equally, the lower 95% CrI and HPDI limits appear to be equally sensitive to the differences in treatment effect. It can be seen that, across both categories, the upper CrI and HPDI limits were close one, the maximum allowable value for a_0. Encouragingly, however, there was still some distinction between the two sets of scenarios, with the upper limits appearing slightly larger for the scenarios with the same treatment effect. There was no obvious difference between the summary statistics of a_0 between the scenarios in which $k_0 = 4$ and $k_0 = 8$.

Figure 3.5 shows summary statistics for a_0 with scenarios grouped by whether the ICCs for the data generating mechanisms for the definitive and pilot data are the same (i.e. $\rho = \rho_0$), shown in grey, or different (i.e. $\rho \neq \rho_0$), shown in blue. Across the mean, median and lower limits of a_0, it can again be seen that the value of a_0 estimated through the NPP approach appears proportionate to the similarity between the pilot and definitive trial datasets with different degrees of clustering. Once again the upper CrI and HPDI limits were close to one, although it appears that there was less variation amongst the scenarios in which the ICCs were the same for the two datasets.

Figure 3.6 shows the summary statistics for a_0 grouped by differences in both treatment effects and ICCs as follows: (i) different treatment effects and ICCs (i.e. $\theta \neq \theta_0$ and $\rho \neq \rho_0$), shown in blue; (ii) same treatment effect and different ICCs (i.e. $\theta = \theta_0$ and $\rho \neq \rho_0$), shown in grey; (iii) same ICC and different treatment effects (i.e. $\theta \neq \theta_0$ and $\rho = \rho_0$), shown in green, and (iv) same ICC and treatment effect (i.e. $\theta = \theta_0$ and $\rho = \rho_0$), shown in orange. It can be seen that, as the similarity between the data generating mechanisms increased, so did the value of a_0, across all summary statistics, indicating that the NPP was working well in terms of appropriately accounting for differences between data sets. Furthermore, there appears to be no clear distinction in the values of a_0 between the category where the treatment effects are the same and the ICCs different (grey), and the category where the ICCs are the same but the treatment effects are different (green). As a result, it is not possible to conclude which of the two parameters (the treatment effect or the ICC) have a larger impact on the value of a_0.

A second, supplementary simulation study, exploring inference upon a_0 for more extreme differences between pilot and definitive trial data than could usually be expected within the context of an RCT is outlined and reported in Appendix D. Within this simulation study, sensitivity of the posterior distributions of a_0 are further demonstrated, showing substantial discounting of the historical data both when the treatment effect and the ICC differs, and in some cases resulting in near-complete discounting.

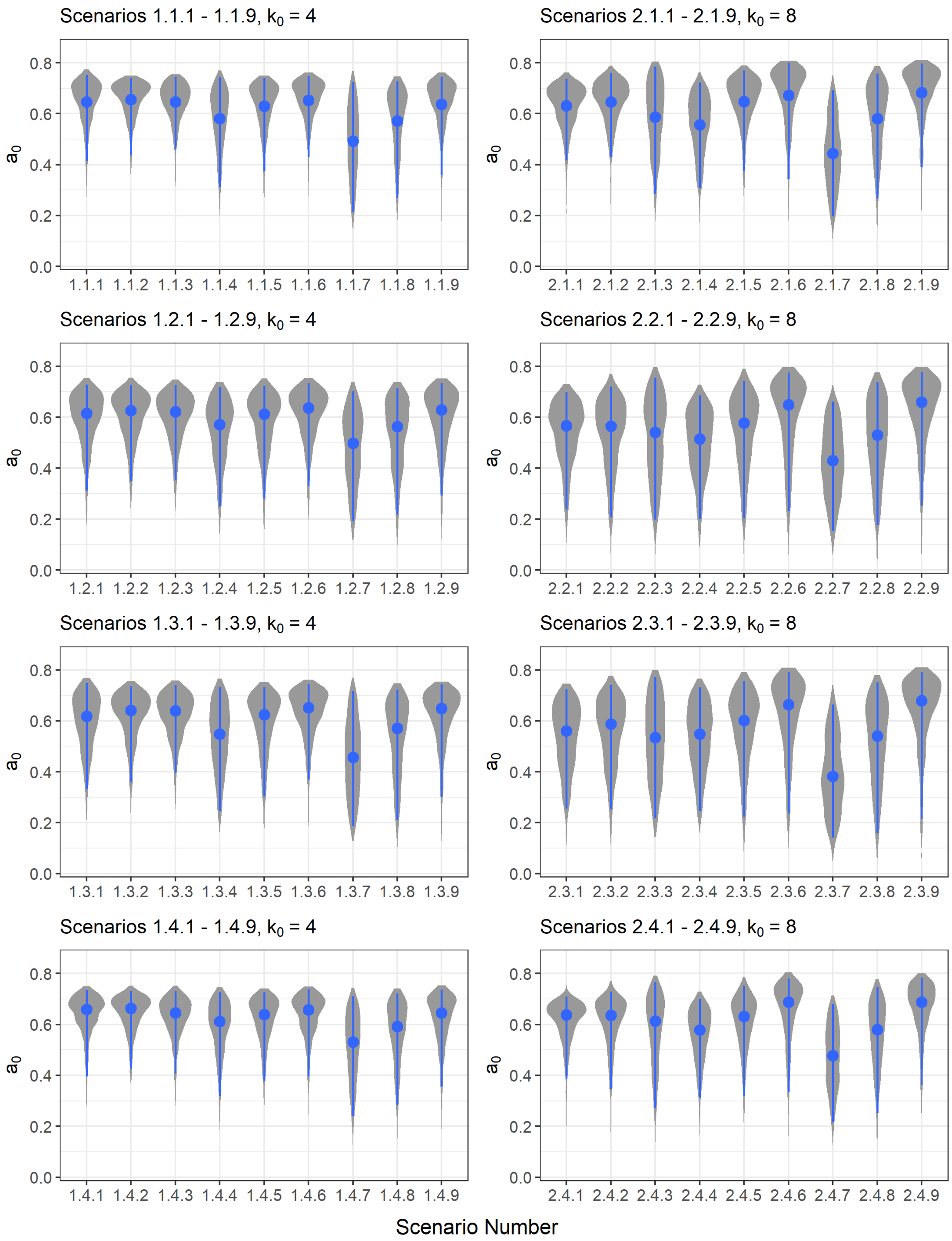


Figure 3.3: Eye plots showing the mean of the point estimates (median) of a_0 for each iteration within each scenario, with 95% CrIs and overlaid with the density of the median of a_0 for each scenario.

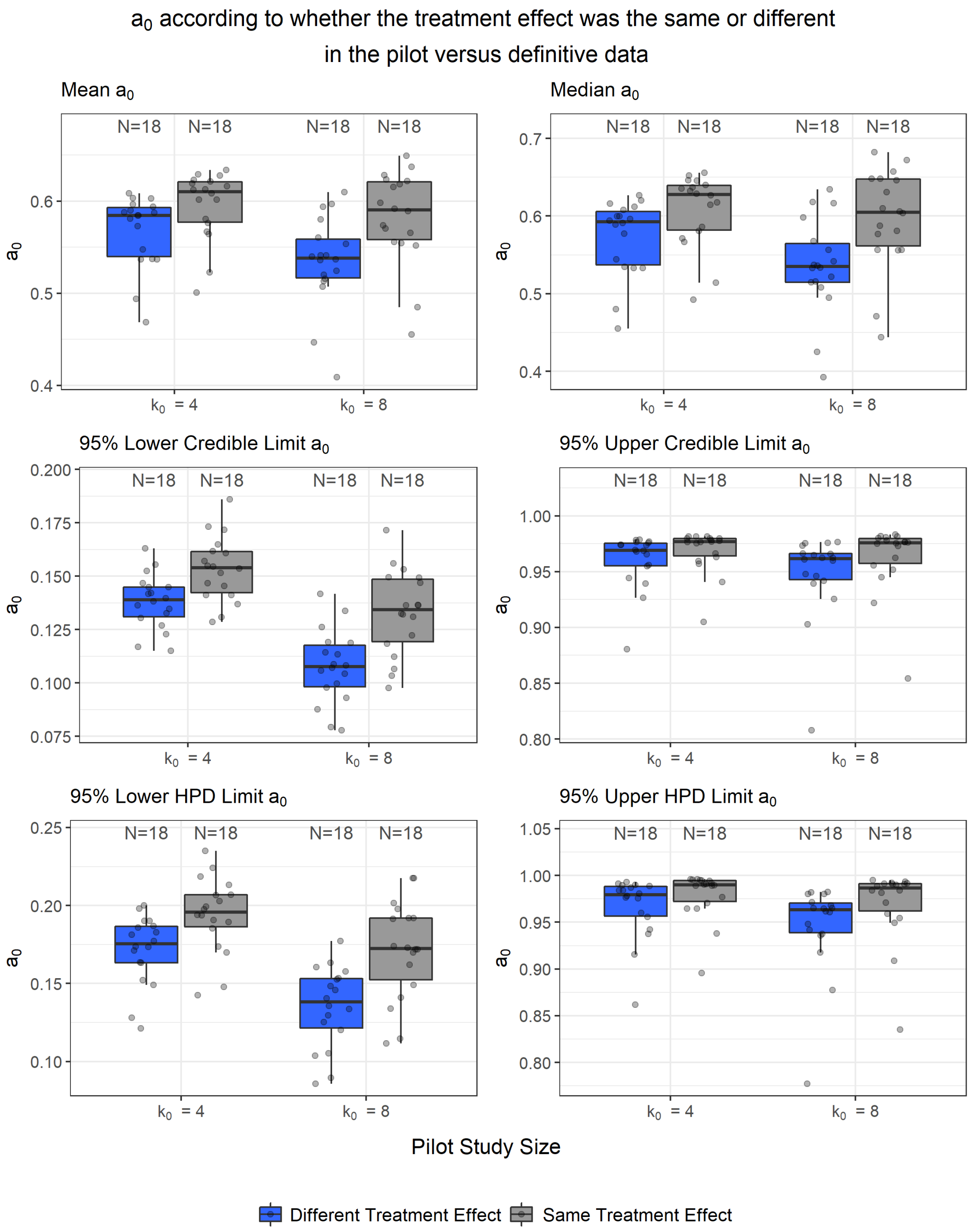


Figure 3.4: Boxplots of summary statistics for a_0 grouped according to whether the treatment effect from the underlying data generating mechanism was different (blue) or the same (grey) for the definitive and pilot data.

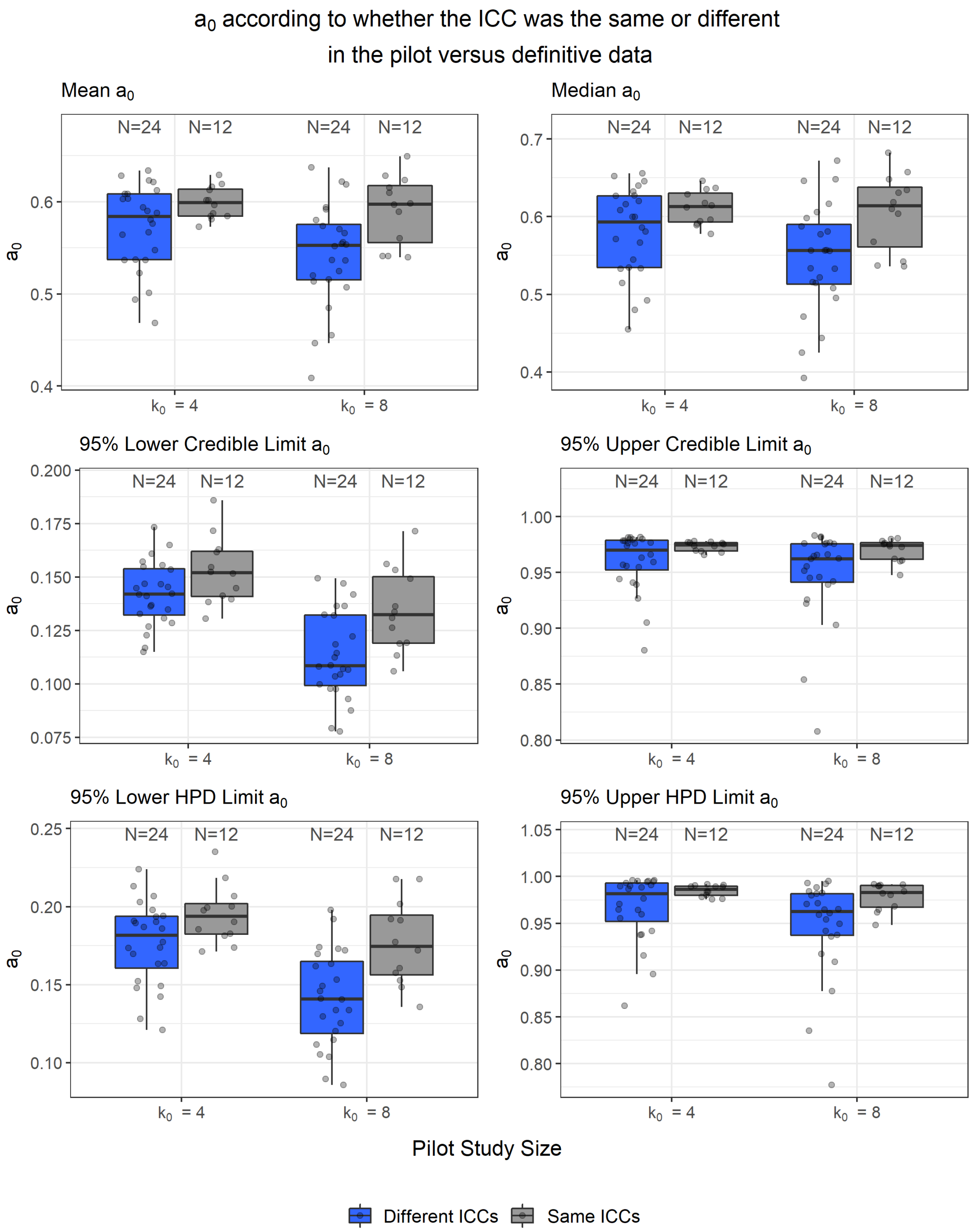


Figure 3.5: Boxplots of summary statistics for a_0 grouped according to whether the ICC from the underlying data generating mechanism were different (blue) or the same (grey) for the definitive and pilot data.

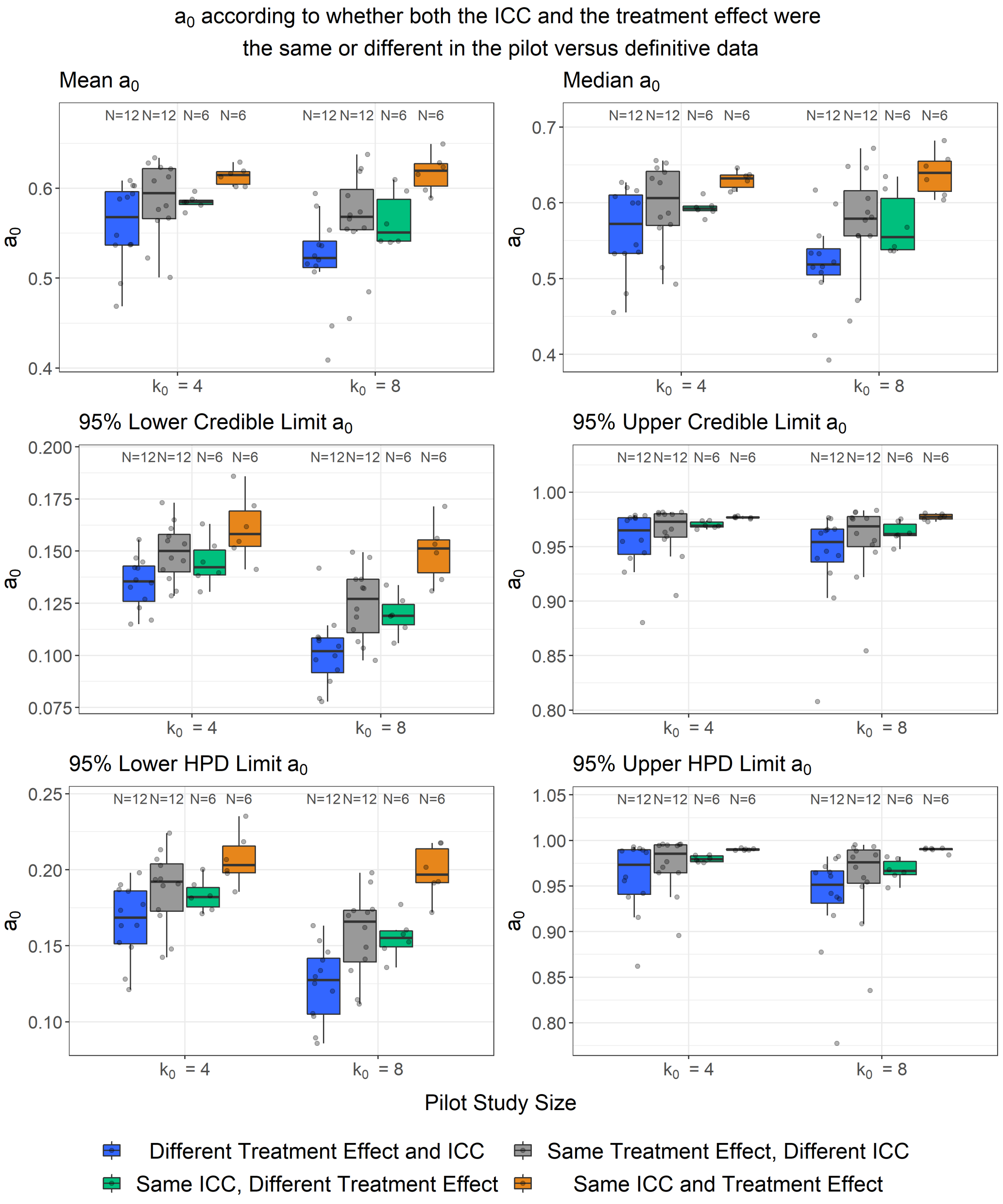


Figure 3.6: Boxplots of summary statistics for a_0 grouped according to whether, for the definitive and pilot data: (i) the treatment effects and the ICCs were different (blue); (ii) the treatment effects were the same, but the ICCs were different (grey); (iii) the treatment effects were different, but the ICCs were the same, and (iv) both the treatments effects and the ICCs were the same (orange).

Treatment Effect

Figures 3.7 - 3.12 illustrate the performance of the NPP compared to the other analysis strategies, as measured by a range of performance metrics pertaining both to the estimated treatment effect and to the associated 95% intervals.

Figure 3.7 shows the mean bias according to the difference between the estimated treatment effect from each of the five models and the treatment effect from the underlying data generating mechanism for the definitive trial data. Analysis of the definitive data alone was, by definition, unbiased, which is reflected in the visualisation under both the Bayesian and frequentist framework across all scenarios. Furthermore, the scenarios represented in rows (a) and (d) pertain to data generating mechanisms in which the treatment effect was the same across the definitive and pilot data, and so analysis remained unbiased when combining the two datasets, either through simple pooling or through the NPP approach. Within rows (b) and (c), it can be seen that incorporation of pilot data with differing treatment effects did, unsurprisingly, introduce bias in to the analyses. Both Bayesian and frequentist analysis of the pooled data performed similarly, but the NPP approach appears to consistently realise less biased results than either of these approaches.

Figure 3.8 shows the MSE of the treatment effect estimates for each simulation within each scenario. As illustrated, it can be seen that, on the whole, incorporation of pilot data, whether through simple pooling, or through the NPP approach, outperformed analysis of the definitive data alone in terms of MSE. Moreover, in almost every scenario, the NPP approach was either consistent with, or outperformed the simple pooling approaches.

The empirical standard error of each modelling approach for each scenario is shown in Figure 3.9. The Empirical Standard Error is a measure of precision, or efficiency, of a modelling technique over each iteration of the simulation [Morris et al., 2019], where smaller values represent a higher degree of efficiency. The incorporation of pilot data into the analysis resulted in a smaller Empirical Standard Error compared with analysis of the definitive data alone, regardless of whether this was simply pooled, or incorporated using the NPP approach. For the majority of the scenarios presented, the NPP approach either performed similarly to the simple pooling approaches, or slightly underperformed, with the notable exception of scenarios in which the ICC for the pilot data was 0.1, and the ICC for the definitive trial data was 0.01, where the NPP outperformed all other approaches (i.e. scenarios 1.1.7, 2.1.7, 1.2.7, 2.2.7, 1.3.7, 2.3.7, 1.4.7, 2.4.7), likely due to the more heavy discounting of the pilot data (shown in Figure 3.3). Furthermore, the simple pooling and NPP approaches consistently outperformed the analyses of the definitive trial data alone.

Figure 3.10 shows the power of each of the modelling approaches to detect the underlying treatment effect in the definitive data, defined as the proportion of iterations in which the 95% interval excludes 0. Recall that the sample sizes for the definitive data within each scenario were calculated in order to ensure at least 85% power according to frequentist methodology, which is illustrated by the dashed horizontal line in each of the plots. It can be seen that the incorporation of pilot data often had a material impact on power, either through simple pooling or the NPP approach. For the scenarios shown in rows (c) and (d), there were notable increases in power as a result of incorporation of the pilot data. Within row (c), the scenarios shown all pertain to data generating mechanisms in which the treatment effect for the pilot data was greater than the treatment effect for the definitive trial data, and therefore the increase in power was predominantly a result of the overall increase in estimated treatment effect when including this pilot data. The scenarios shown in row (d) are associated with data generating mechanisms in which the treatment effect was the same for the pilot and definitive data, and for which the size of the pilot data was large (in terms of both the number of participants and the number of clusters) relative to the size of the definitive trial data. As a result, the increases seen in power can be attributed simply to an increase in sample size, with more modest gains realised through the NPP approach due to the partial discounting of the information obtained from this additional sample size. A similar situation can be observed within the scenarios illustrated in row (a), where again the treatment effects for the pilot and definitive data were the same, although here the increase was more modest as the sizes of the pilot data were smaller relative to the definitive data. Finally, row (b) illustrates scenarios in which the treatment effect for the pilot data was smaller than that of the definitive data, a situation in which intuitive expectation of the impact on the power of incorporating this pilot data is less clear. The increased sample size may lead to an increase in precision and a narrowing of the 95% intervals, Or alternatively, incorporating this additional data may shift the overall treatment effect estimate towards zero. For the simple pooling approach, the simulation results suggest that both can occur. In particular, in cases where the ICC in the pilot data was small (scenarios 1.2.1, 1.2.2, 1.2.3, 2.2.1, 2.2.2, 2.2.3), there appears to be modest increases in power. However, as this ICC increases relative to the ICC for the definitive trial data, simply pooling the data can have detrimental effects on the power, in particular in scenarios 1.2.7 and 2.2.7. In contrast, the NPP approach was able to achieve improvements in power for scenarios in which the pilot ICC was small (scenarios 1.2.1, 1.2.2, 1.2.3, 2.2.1, 2.2.2, 2.2.3), whilst also mitigating against the loss of power induced by the simple pooling approach for scenarios in which the pilot ICC was large (in particular, scenarios 1.2.7 and 2.2.7).

Figure 3.11 shows the coverage of the 95% intervals for the treatment effect for each scenario, which by definition should be 95%, as shown by the dashed horizontal line. For most scenarios, there was no obvious discernible difference in performance between analysis methods, with the exception of the scenarios shown in row (b), in which the treatment effect for the pilot data was smaller than that of the definitive trial data. It can be seen that in some scenarios, simply pooling the two datasets had a detrimental effect on the coverage, particularly in scenarios 2.1.1 - 2.1.9. However, the NPP approach mitigated this, maintaining coverage at approximately 95%.

Finally, Figure 3.12 shows the width of the 95% intervals for the treatment effect. As expected, the introduction of additional data in to the analysis, either through simple pooling, or through the NPP approach, reduced the width of the interval for the treatment effect, thus improving the precision of the estimate. This reduction was more modest through the NPP approach, compared to simply pooling datasets, reflecting the discounting of some of this additional information.

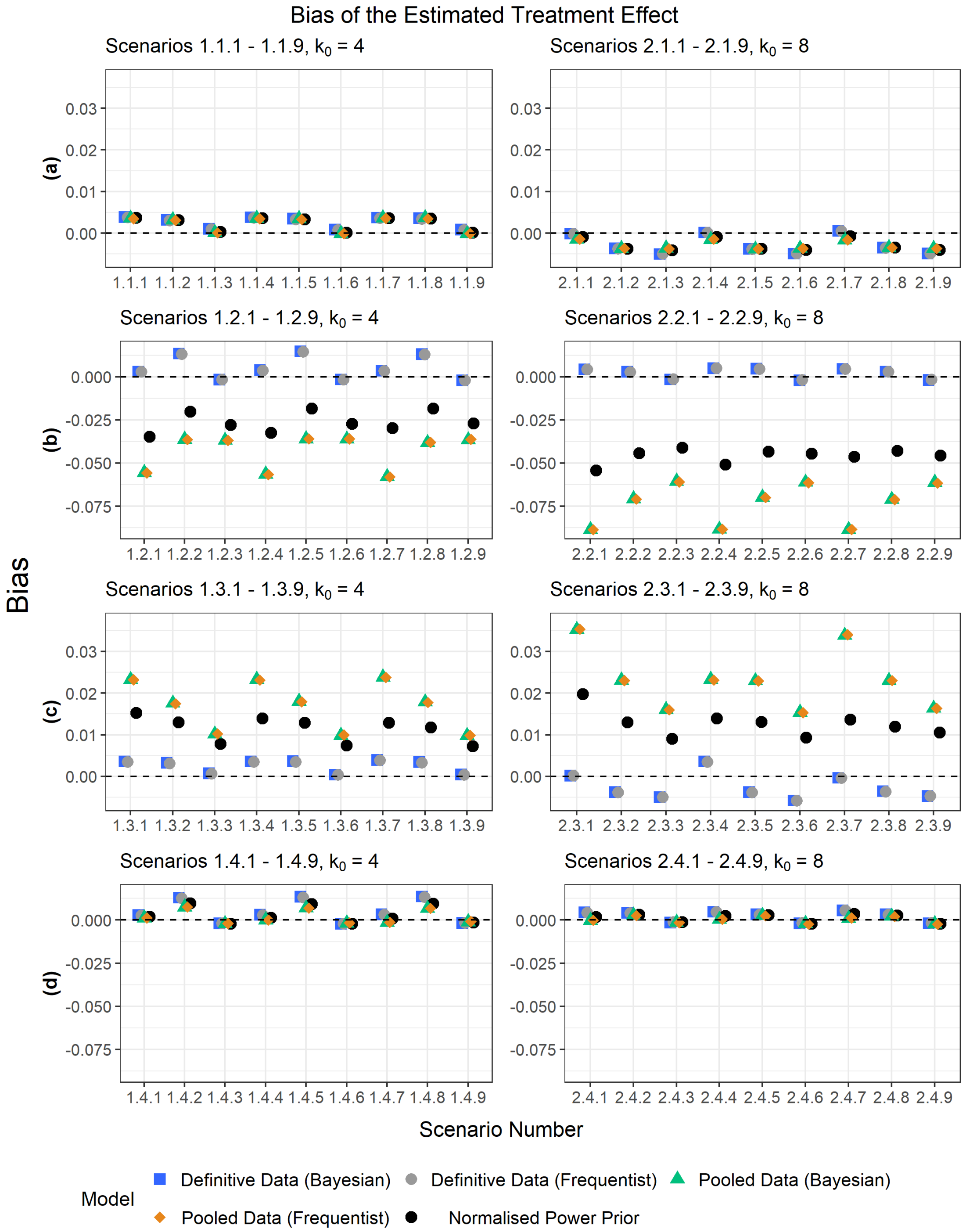


Figure 3.7: Average bias of the estimated treatment effect by scenario and modelling approach.

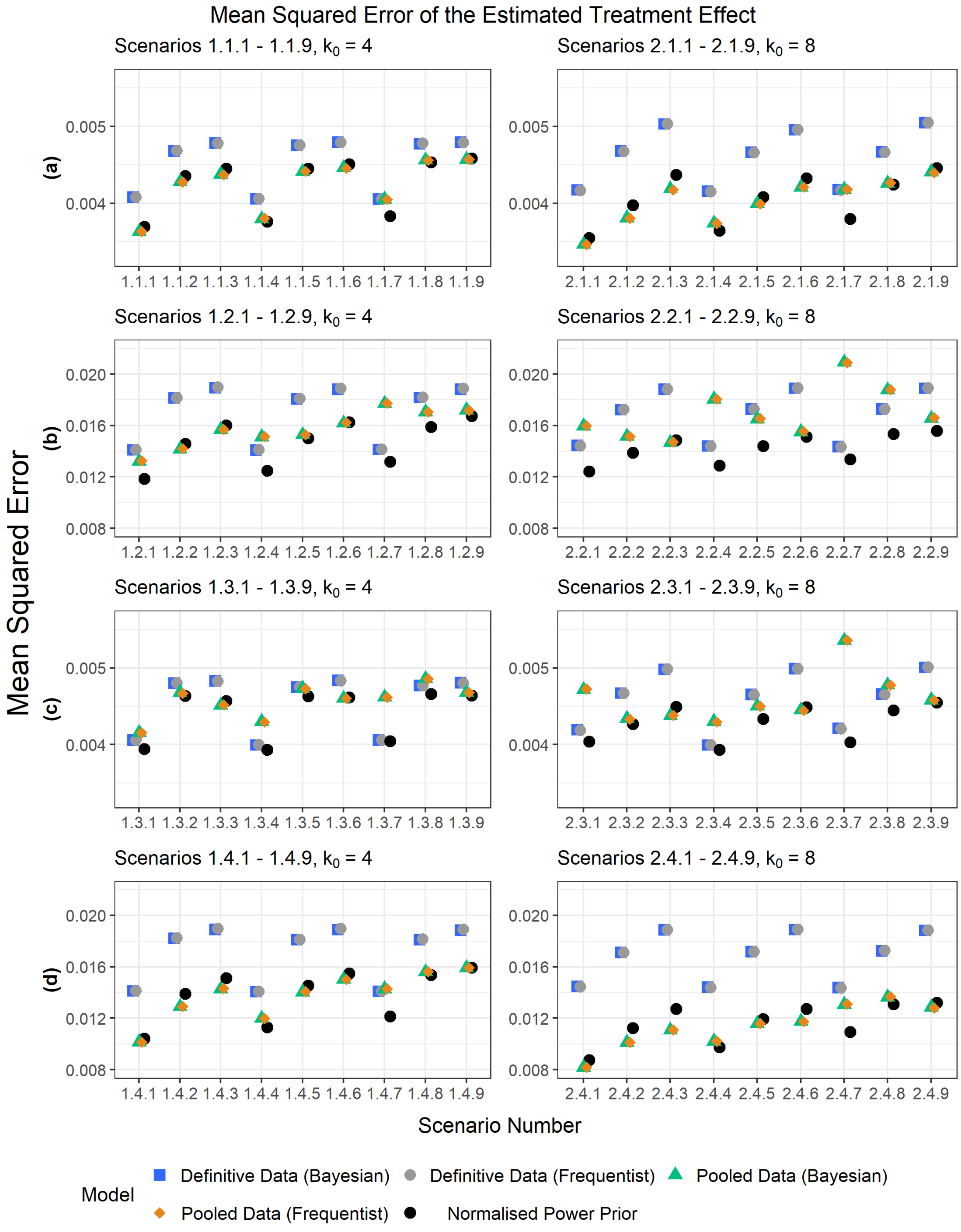


Figure 3.8: Mean squared error of the estimated treatment effect by scenario and modelling approach.

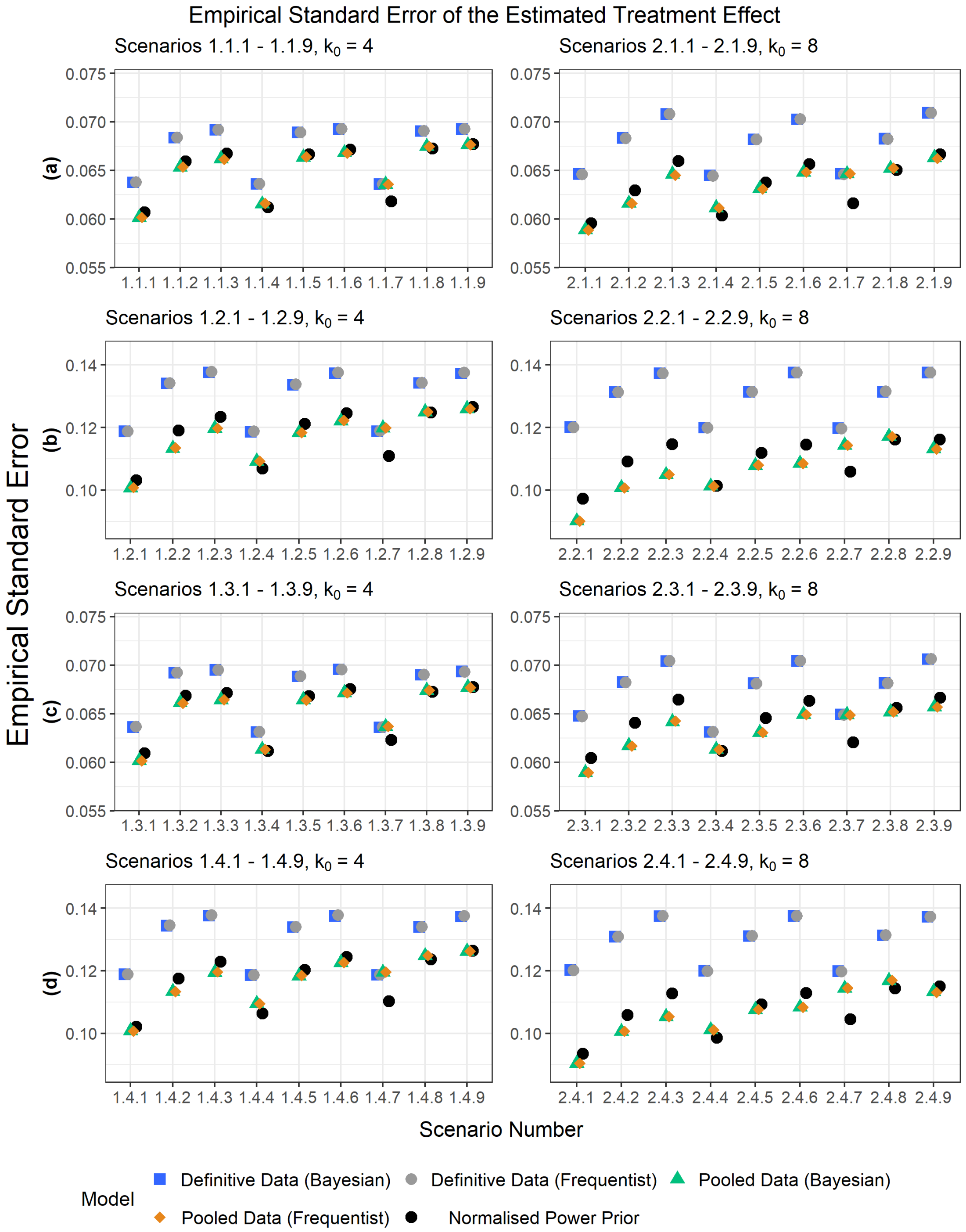


Figure 3.9: Empirical standard error of the treatment effect by scenario and modelling approach.

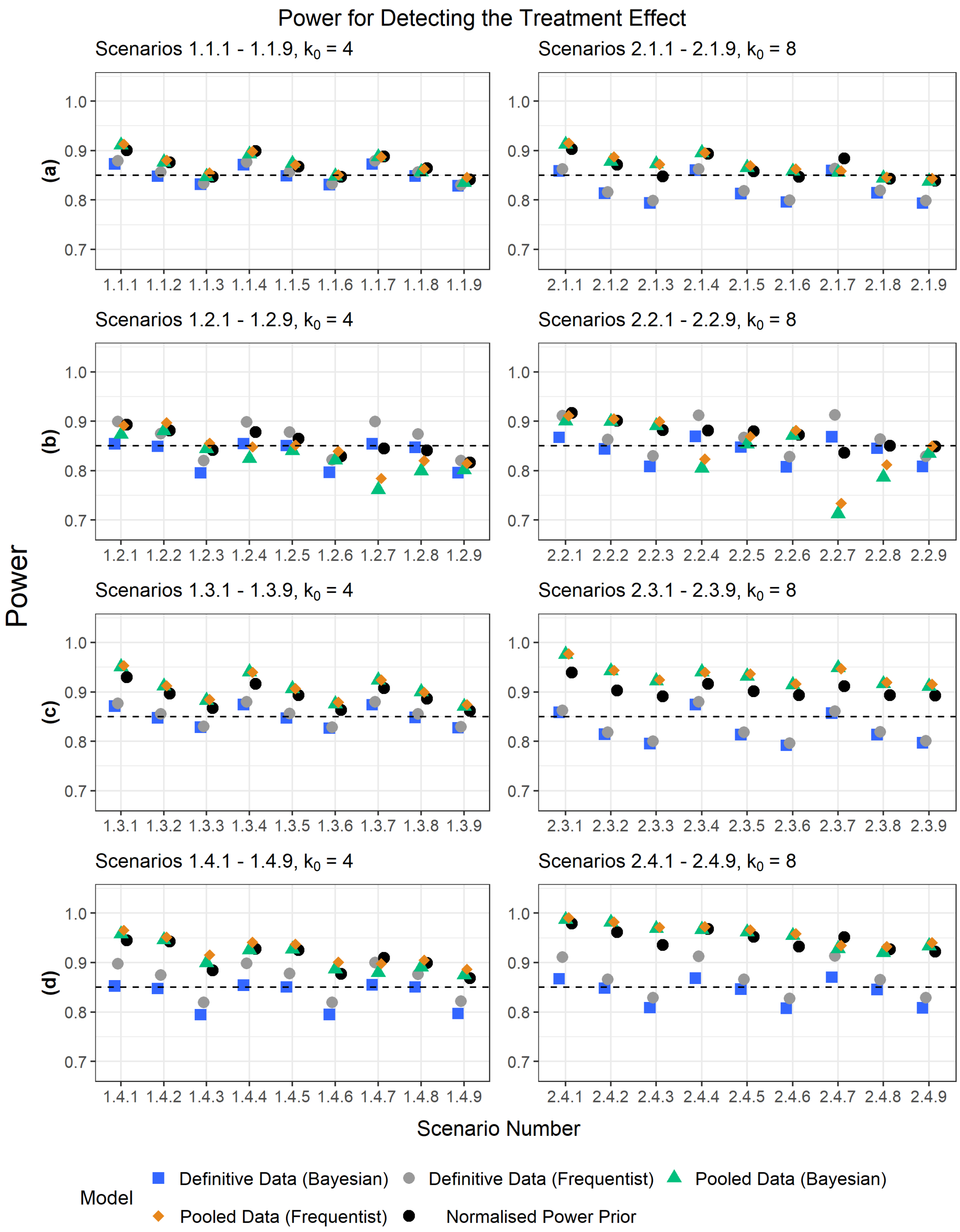


Figure 3.10: Overall power for detecting the treatment effect used to simulate the definitive trial data by scenario and modelling approach, with dashed horizontal line indicating 85% power.

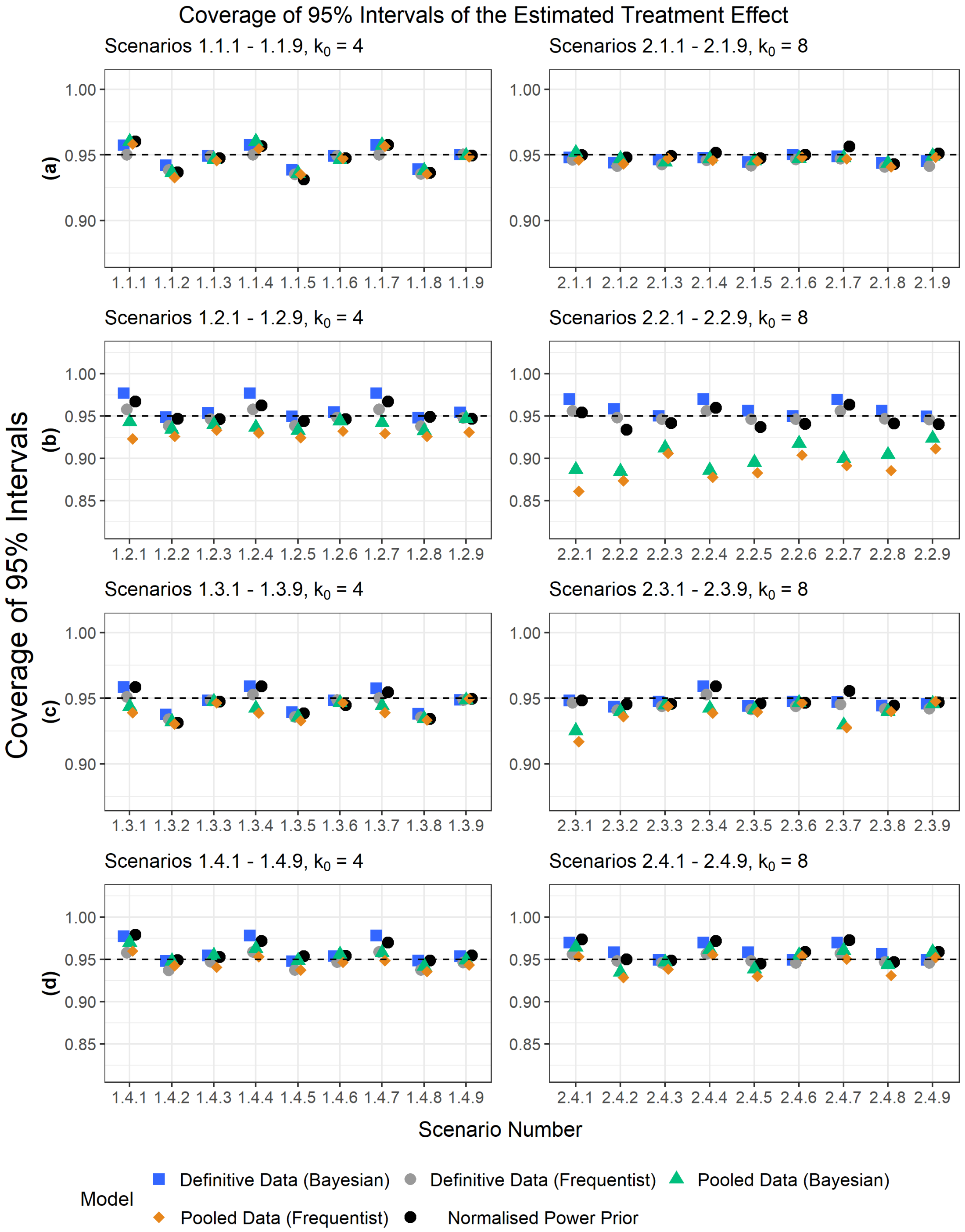


Figure 3.11: Coverage of the 95% Intervals for the estimated treatment effect by scenario and modelling approach.

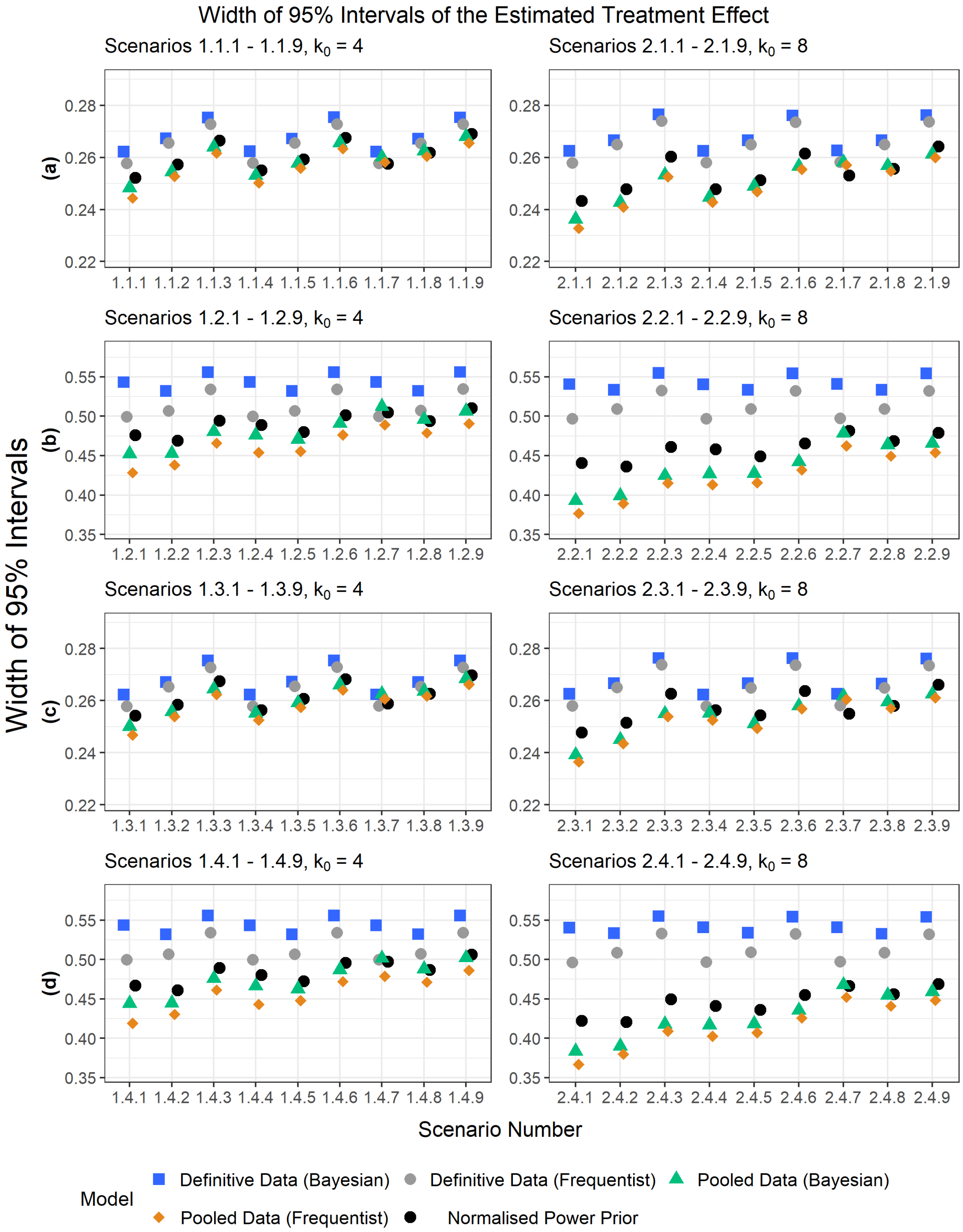


Figure 3.12: Average width of the 95% Intervals for the estimated treatment effect by scenario and modelling approach.

Intracluster Correlation Coefficient

The ICC is a key metric often estimated and reported from CRCT data, not least because it provides valuable data to inform power calculations for subsequent trials. As a result, the performance of the NPP was also considered in the context of estimation of the ICC. Figures 3.13 - 3.17 illustrate this performance against the other modelling approaches across a range of metrics. As ICCs are often skewed, these performance metrics were calculated according to the median of the posterior distribution of the ICC for the bias, MSE and empirical standard error, and according to 95% HPDIs for the coverage and interval widths for the Bayesian models. For the frequentist models, the ICC was calculated from the model estimates of the variance parameters, and the confidence intervals according to Swiger's method [Ukoumunne, 2002], with truncation of the lower limit at 0 if required.

From Figure 3.13 it can be seen that, as expected, incorporation of additional data in to the analysis, either through the NPP approach or simple pooling, introduced bias when the underlying ICC between the two datasets are different (e.g. scenarios 2.2.2, 2.2.3). However, it can also be seen that discounting of the pilot data through the NPP approach reduced this bias compared to simple pooling in some scenarios. In fact in some cases, particularly when the pilot ICC was large relative to the definitive trial ICC (e.g. scenarios 2.2.7, 2.2.8), the NPP approach was able to maintain an unbiased estimate of the ICC whilst the simple pooling approach was biased.

As shown by Figure 3.14, for many scenarios, the performance in terms of the mean squared error was consistent across all five modelling approaches. However, there were scenarios, particularly where the pilot data ICC is significantly larger than the definitive trial data (scenarios 1.1.7, 2.1.7, 1.2.7, 2.2.7, 1.3.7, 2.3.7, 1.4.7, 2.4.7) where the mean squared error increased substantially when pooling the two data sets. However, the simulation results suggest that the NPP approach effectively mitigated against this increase in mean squared error, placing its performance broadly in line with the analyses of the definitive trial data alone, across all scenarios. A similar pattern emerged when considering the efficiency of the modelling techniques as measured by the empirical standard error and illustrated in Figure 3.15.

Figures 3.16 and 3.17 show the coverage and the width of the intervals for the ICC, respectively. For scenarios where the pilot ICC is smaller than the definitive ICC (e.g. scenarios 1.2.3, 2.2.3), there were modest reductions in coverage when combining the data, either through simple pooling or through the NPP approach. However, simple pooling of the data again results in large performance reductions for scenarios in which the pilot ICC was substantially larger than the definitive trial data (scenarios 1.1.7, 1.2.7, 2.1.7, 2.2.7, 1.3.7, 2.3.7, 1.4.7, 2.4.7), although the NPP again performed in line with the analyses of the definitive data alone. Across the majority of scenarios, incorporation of pilot data resulted in a reduction in the width of the 95% interval for the ICC. The exception once again was the scenarios in which the pilot ICC was larger than the definitive data ICC (scenarios 1.1.7, 1.2.7, 2.1.7, 2.2.7, 1.3.7, 2.3.7, 1.4.7, 2.4.7). In these cases, simple pooling of the data resulted in an increase in the width of the intervals, and therefore a decrease in the precision of the estimation of the ICC. This decrease, however, was mitigated through the use of the NPP approach.

Divergent Transitions

Within a Bayesian model fitted using HMC and Stan, even a small number of divergent transitions can be indicative of a problem, suggesting thorough exploration of the posterior distribution may not have been possible, and thus the possibility of unreliable results. Within this simulation study, any model in which at least 0.1% (8) of the post-warmup iterations were divergent transitions were removed from the final analysis (note that the maximum number of iterations with ≥ 8 divergent transitions within a single scenario was 55). However, the number of iterations with divergent transitions were further explored.

Figure 3.18 shows a clear, positive correlation between the similarity of the simulated datasets (as indicated by larger values of a_0, as well as data generating mechanisms with the same treatment effect and/or ICC, as indicated by the grey and green points on the graph). In practice, this suggests that as definitive and historical datasets become more similar, and samples from the posterior distribution of a_0 approach 1, divergent transitions become more prevalent. This is further supported by Figure D.2 (Appendix D) pertaining to the supplementary simulation study, in which the number of iterations with divergent transitions became as low as zero when the data generating mechanisms used to simulate each of the two datasets were substantially different.

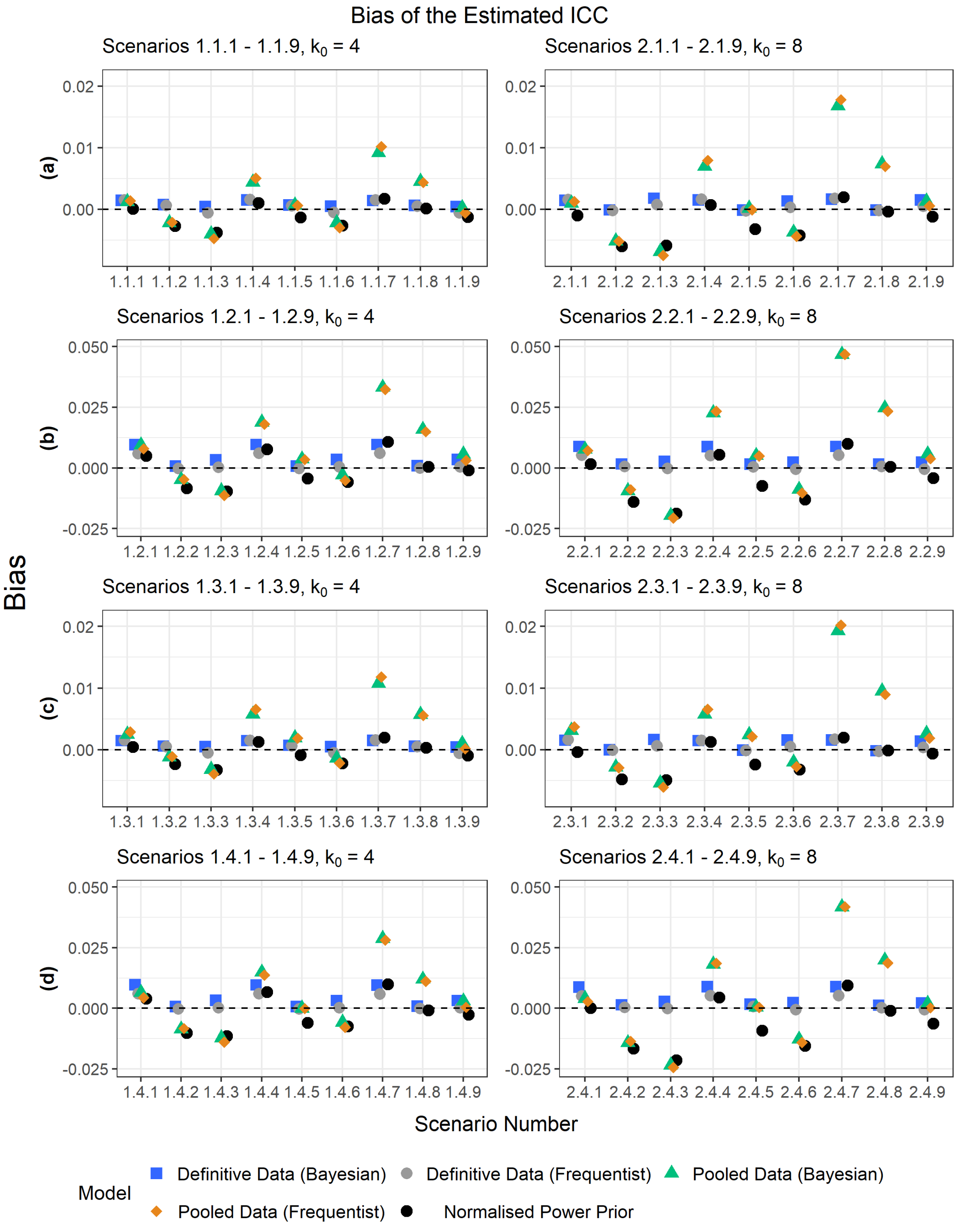


Figure 3.13: Average bias of the estimated ICC by scenario and modelling approach.

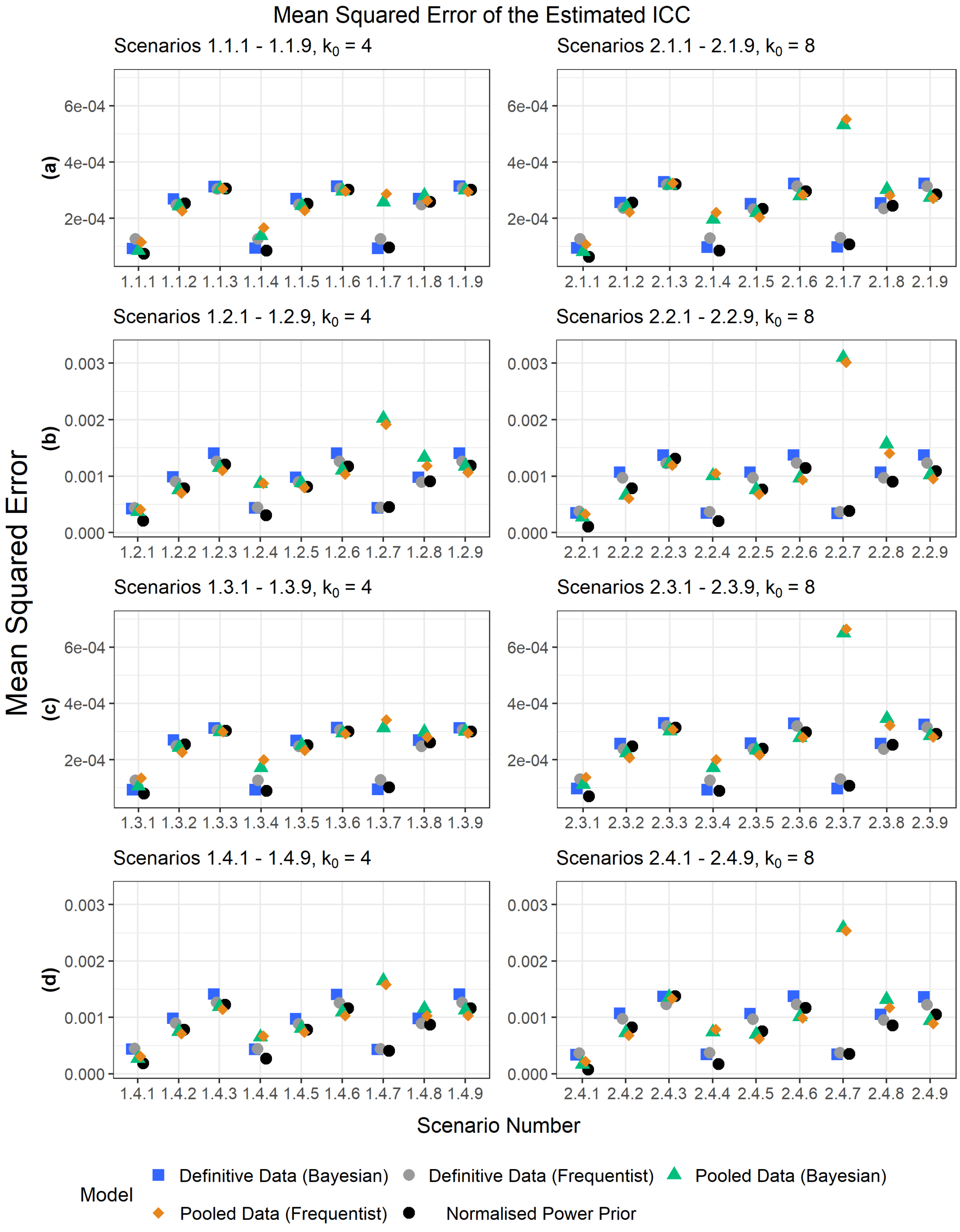


Figure 3.14: Mean squared error of the estimated ICC by scenario and modelling approach.

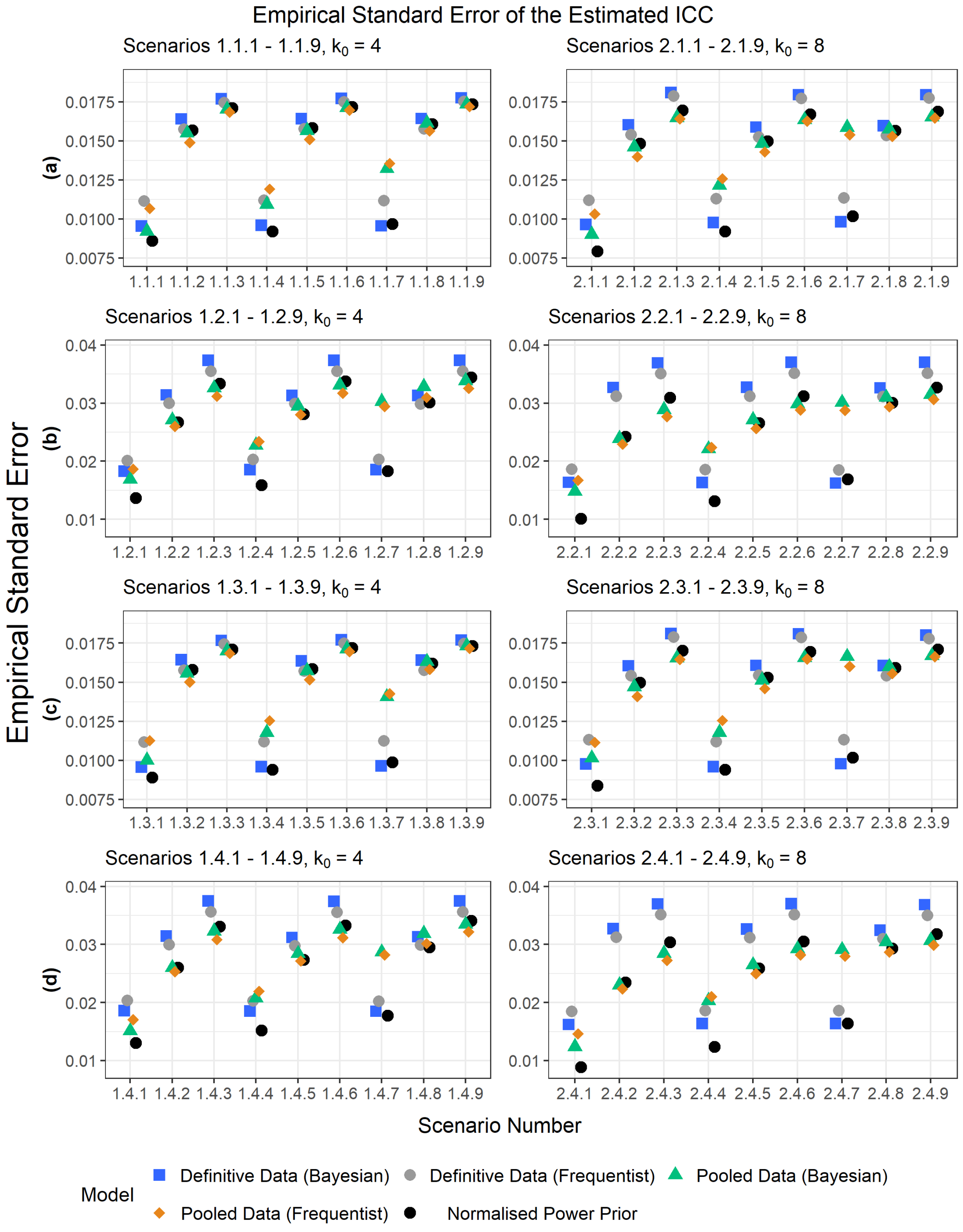


Figure 3.15: Empirical standard error of the estimated ICC by scenario and modelling approach.

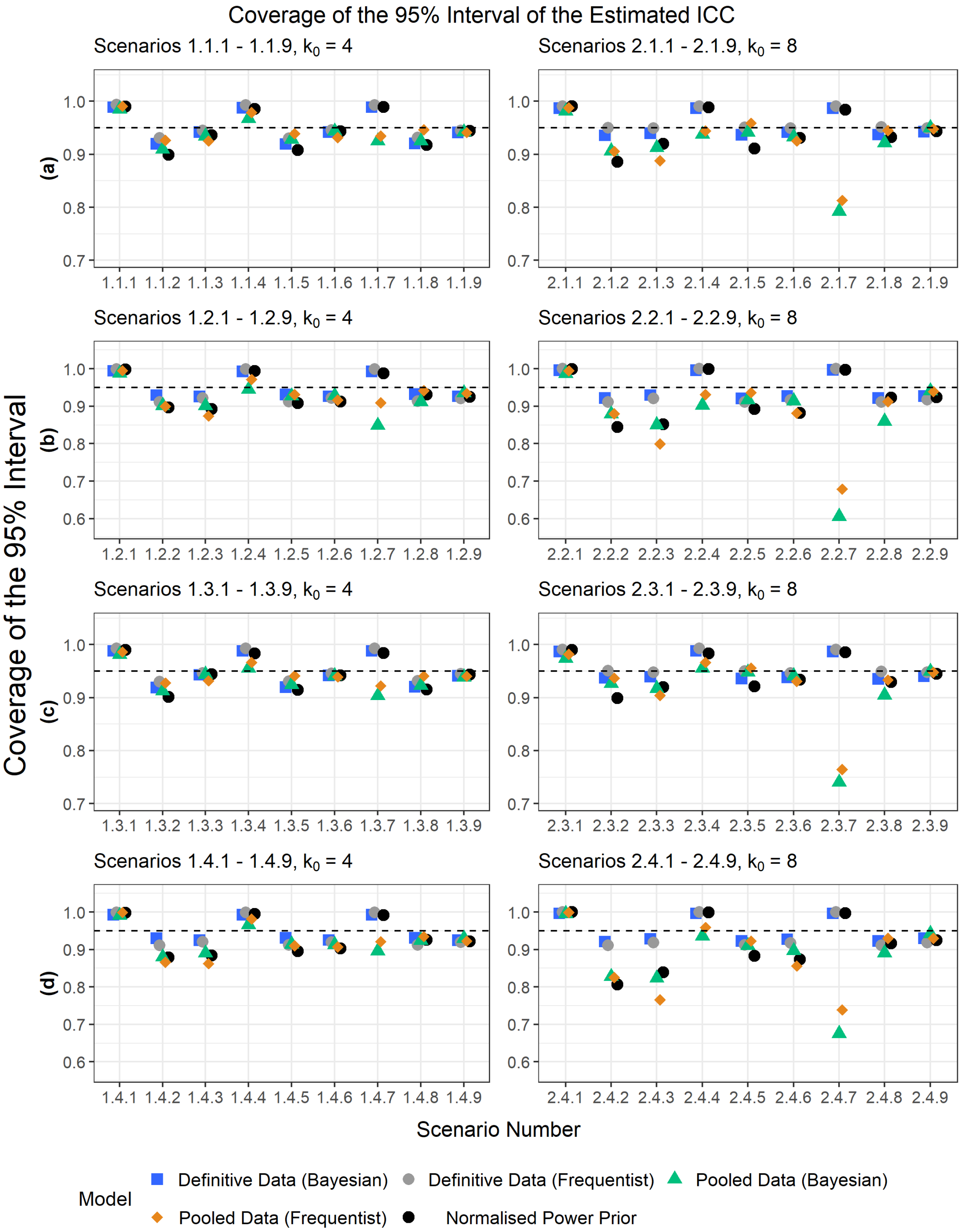


Figure 3.16: Coverage of the 95% interval for the estimated ICC by scenario and modelling approach.

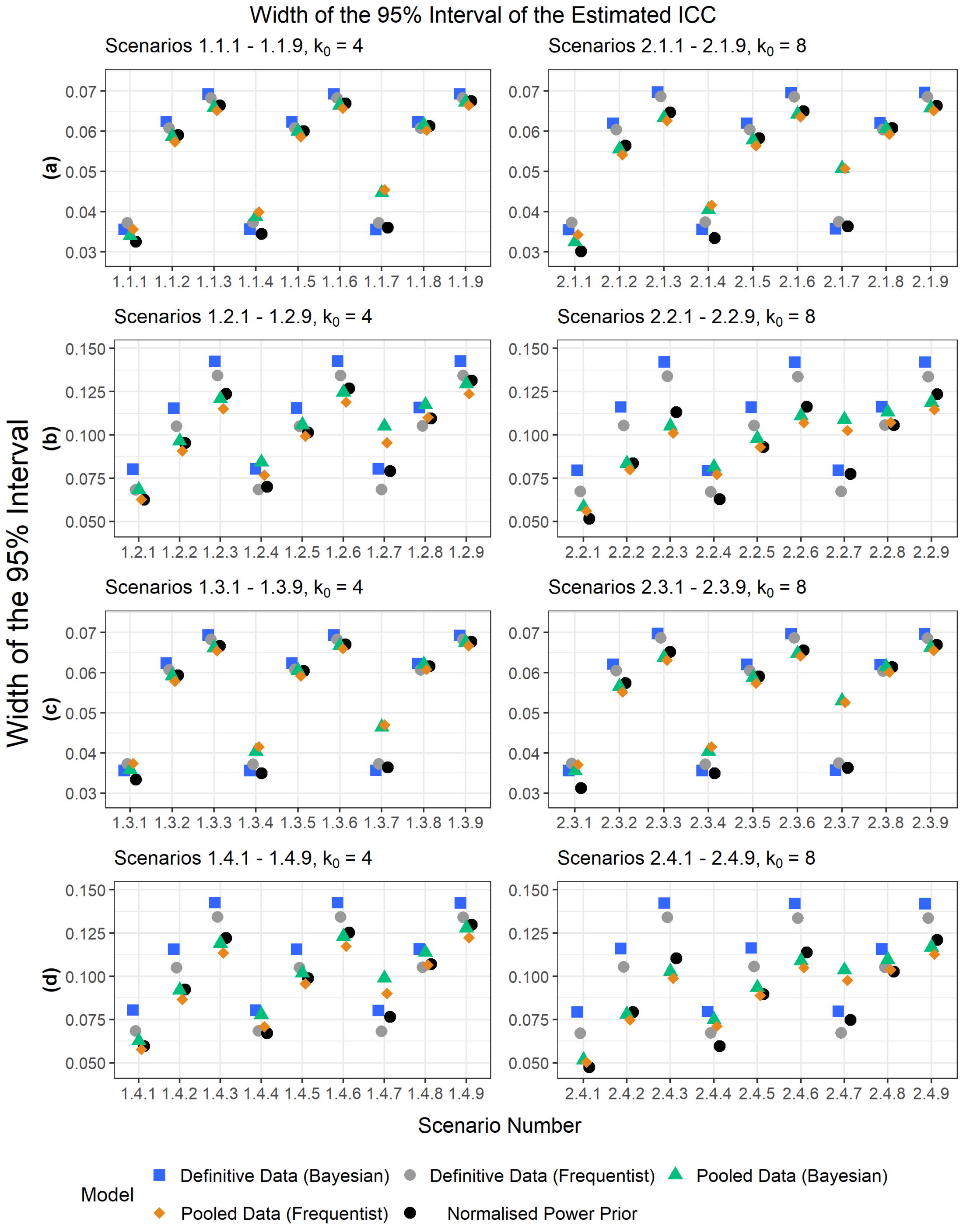


Figure 3.17: Average width of the 95% interval for the estimated treatment effect by scenario and modelling approach.

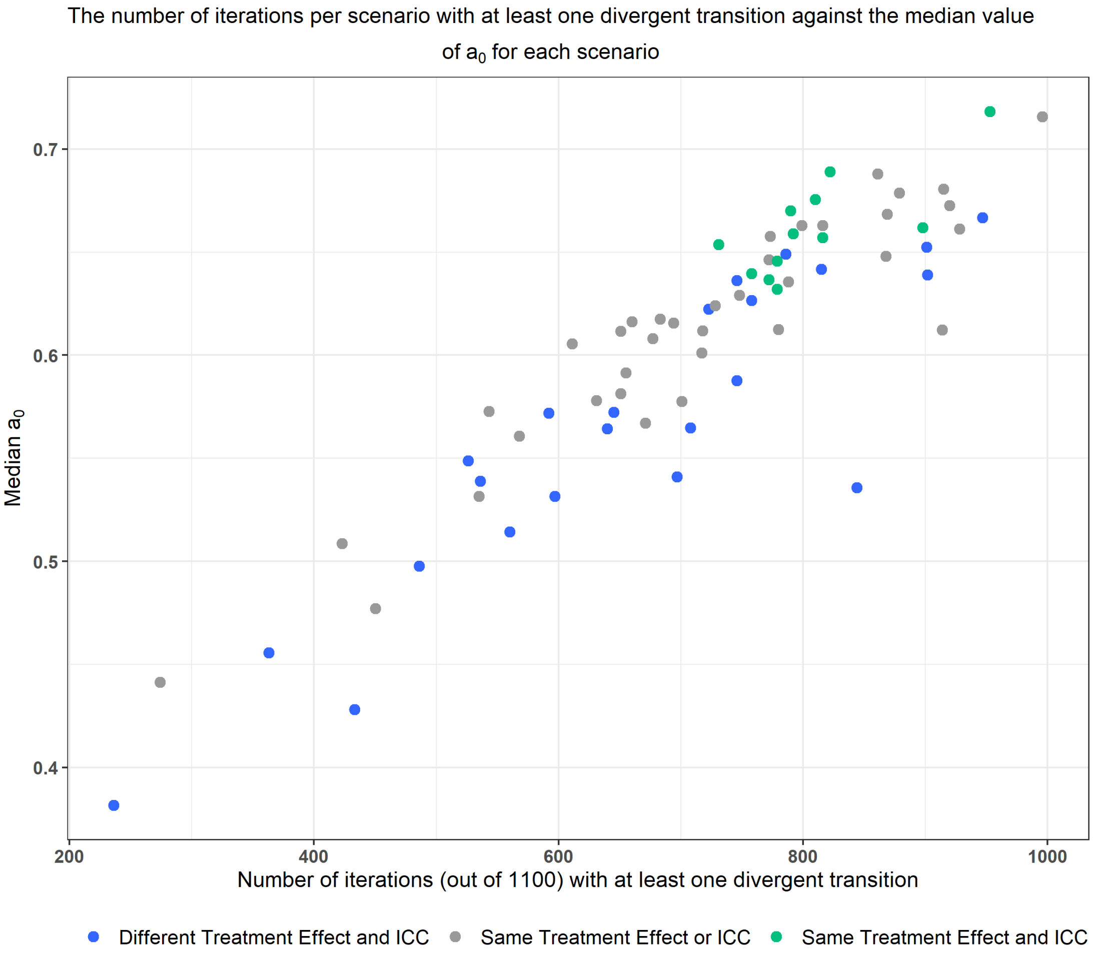


Figure 3.18: A scatterplot of the number of iterations per scenario with at least one divergent transition against the median value of a_0 for each scenario.

***

3.7 Discussion

In this chapter, a novel NPP was proposed which facilitates incorporation of continuous historical data into the analysis of data from a definitive CRCT, where both datasets are clustered and therefore analysed using hierarchical models. The proposed methodology assumes a fully Bayesian approach to model formulation, where the discounting parameter, a_0, is treated as random and estimated using MCMC methods according to the commensurability between the two datasets of interest.

The proposed method has been demonstrated with real data from a CRCT and the preceding pilot study. The conclusions drawn from analysing the two datasets independently (whilst acknowledging that it is not good practice to perform inferential analysis on pilot or feasibility data) were substantially different. Using the newly developed methodology, this contrast was reflected in the fairly heavy discounting of the information obtained from the pilot data, as shown in the estimations of a_0. It was also shown that incorporation of this historical data was enough to change the estimated treatment effect but, in this case, not enough to change the overall conclusions.

An extensive simulation study was undertaken across a range of scenarios which may plausibly be encountered in the context of CRCTs. The sensitivity of the discounting parameter, a_0, to varying similarity between datasets has been demonstrated. In estimating the ICC, it was shown that the NPP approach may result in some additional bias, but has the potential to facilitate more precise estimates of the ICC through narrower intervals, thus contributing to a more robust evidence base for informing future study design. Furthermore, it was shown that, when estimating the treatment effect, whilst some bias may be introduced if the two datasets are derived from different underlying data generating mechanisms, a reduction in mean squared error can be expected, as well as more precise estimation intervals, and an increase in power. As a result, the NPP approach has the potential to facilitate more efficient CRCT study design. However, because a closed formula approach to sample size calculation is not possible for the NPP method, an algorithmic, simulation-based approach must be developed.

The simulation study has highlighted the relationship between the similarity of the pilot and definitive trial datasets, and the chance of encountering divergent transitions within the HMC procedure during inference. This observation is further explored in a supplementary simulation study (Appendix D). The presence of divergent transitions was likely due to the fact that the justification for specifying $a_0 \leq 1$ is practical rather than mathematical; in practice, it is difficult to imagine a scenario in which one would give greater weighting to pilot data than to data obtained from a definitive trial, but mathematically, values of a_0 exceeding one do not pose a problem. The divergent transitions, therefore, are likely to be as a result of the numerical procedure attempting to explore the posterior distribution of a_0 for values very close to one, but being prevented from going above one by the bounded Beta$(1,1)$ prior. Divergent transitions occur far less often when the two datasets are dissimilar, because under such scenarios a_0 is further from one. However, whilst encountering at least one divergent transition was common, the maximum number encountered in a single iteration was as small as 16 (0.2%). These findings highlight the need for robust sensitivity analyses using a range of fixed a_0, which should always be recommended when implementing these methods in the analysis of CRCTs in practice.

The new methodology proposed in this chapter pertains only to continuous data, and only facilitates the incorporation of one historical data set. However, there is potential to extend the methodology both to allow for other types of data, such as binary or count data, and to allow for multiple historical datasets, as has been considered previously within the wider power prior literature [Ibrahim et al., 2015]. Furthermore, it is now well established that variability in cluster size should be accounted for in the design of CRCTs. The simulation study presented in this chapter assumes fixed cluster sizes, and further work could consider the performance of the NPP under varying degrees of cluster size variability.

Chapter 4

An Exploration of the Impact of Using Information Borrowing Methods During Study Design and Sample Size Calculation

This chapter explores the effect of implementing a normalised power prior analysis, introduced in Chapter 3, on the design of a cluster randomised controlled trial. A simulation-based approach to sample size calculation is outlined. The impact of the incorporation of evidence from historical data is quantified in the context of its effect on type I error and statistical power, both in the context of the Healthy Lifestyles Programme, and through a simulation study. The impact of placing sampling priors on the design parameters is also explored.

***

4.1 Introduction

THE determination of the required sample size is a crucial exercise undertaken in the design of randomised trials. In general, the premise of sample size determination involves selecting a sufficient number of participants to ensure that the research question can be definitively answered, whilst simultaneously ensuring that an excess of participants is not recruited and therefore unnecessarily subjected to an experimental intervention, often at additional financial expense. A frequentist sample size calculation typically seeks to ensure a minimum of 80% probability (with 90% increasingly specified in definitive randomised controlled trials) of correctly rejecting a null hypothesis given the alternate hypothesis is true (statistical power), whilst ensuring that the probability of erroneously rejecting the null hypothesis is no greater than 5% (type I error).

For an individually randomised trial, sample size determination for a continuous outcome relies on an estimate of the variability of the (primary) outcome, as well as the specification of a minimum clinically important difference (MCID) or the between-group target difference. The problem of sample size determination is complicated in the context of CRCTs. Due to the homogeneity amongst participants within the same cluster induced by cluster-level randomisation, additional considerations are required to ensure that adequate statistical power is maintained. Specifically, the required sample size calculated assuming an individually randomised trial must be inflated by a value known as the “design effect”, which is a function of the ICC, as well as the estimated average cluster size and an estimate of the variability in cluster size [Eldridge et al., 2006]. Further details of sample size determination in CRCTs is provided in Chapter 1.

CRCTs are typically designed and analysed within the frequentist framework, and within this framework the problem of sample size determination is well understood. Closed formulae are available for calculating the required sample size for for CRCTs for a range of outcome types, including continuous, binary, count, ordinal, time-to-event and rates [Rutterford et al., 2015]. However, one of the key challenges associated with the determination of sample size for CRCTs is the justification of the assumed value of the ICC. Whilst pilot or feasibility studies are often undertaken in order to estimate key design parameters, pilot CRCTs usually recruit too few clusters to estimate the ICC with any useful degree of precision. The CONSORT extension for CRCTs [Campbell et al., 2012] recommends reporting estimates of the ICC alongside trial results for each primary outcome in order to inform the design of future similar studies. Sensitivity analyses are often undertaken during sample size calculations to understand the impact on statistical power of differing assumed values of the ICC.

An alternative method of sample size determination to the usual closed formula approach is via a simulation-based approach. Such an approach can be advantageous over the more commonly applied formula-based approach in its ability to more flexibly accommodate complex study designs or uncertainty in key design parameters, such as the ICC or the variability in a continuous outcome. Simulation-based approaches for sample size determination involve the generation of a large number of datasets according to a set of design assumptions, and calculating the proportion of datasets for which the null hypothesis is rejected using the planned analysis strategy to obtain an estimate of statistical power. Simulation-based approaches for sample size determination in the context of CRCTs are discussed by Arnold et al. [Arnold et al., 2011]. Hybrid approaches that utilise both closed formula and simulation-based methods also offer potential advantages over formulae-based approaches alone. For example, Turner et al. [Turner et al., 2004] proposed a Bayesian meta-analytic approach to synthesising ICC estimates from multiple previous studies to obtain a posterior distribution for statistical power by in turn “plugging in” each sample from the posterior distribution of the ICC to the closed power formula.

The simulation-based approach to sample size determination was extended by Wang and Gelfand [Wang and Gelfand, 2002] who placed the problem entirely into a Bayesian framework, arguing that doing so allows for more thorough expression of the uncertainty surrounding the assumptions for the design parameters underpinning the sample size calculation. Specifically, rather than simply specifying a single, fixed MCID/target difference as in a frequentist setting, they proposed to elicit an informative prior distribution for the treatment effect at the design stage using historical evidence or expertise, known as a *sampling prior*. Following this, they proposed a simulation-based algorithm, with data at each iteration being generated using draws from the *sampling prior* for the target difference/MCID, but then analysed using a (non-informative) *fitting prior*. Performance criteria can then be obtained from the results of the analysis of each simulated dataset. For example, statistical power can be estimated by calculating the proportion of simulated datasets for which the null hypothesis is rejected. It is important to note that the traditional frequentist interpretation of statistical power and type I error do not hold under the fully Bayesian approach to sample size determination, as the target difference is treated as random, rather than fixed, within a Bayesian framework. However, by specifying a point mass sampling prior distribution (i.e. distributions in which the entire mass is placed at a single value) for the treatment effect, frequentist interpretations of statistical power and type I error can be retained.

When adopting more complex analysis strategies, such as the NPP, it is necessary to adopt such simulation-based approaches to sample size determination. The use of power prior methods has been considered in the context of sample size determination in recent previous work; for example, Psioda and Ibrahim [Psioda and Ibrahim, 2019] proposed using fixed values of the discounting parameter, chosen to control type I error at a nominal level. The use of power priors has also been examined in the context of the design of non-inferiority trials [Chen et al., 2011], sequential meta-analysis designs [Chen et al., 2014a, Ibrahim et al., 2012] and trials with recurrent event data [Chen et al., 2014b]. They have also been explored in the context of CRCTs where the discounting parameter is fixed according to the Kullback-Liebler divergence measure quantifying the distance between the historical and current datasets [Xiao, 2017].

In this chapter, the primary aim was to understand whether the incorporation of historical data, such as pilot data, through the use of the NPP analysis method can be used during study design to justify smaller sample sizes, and thus the delivery of a more efficient CRCT. The secondary aim was to understand the impact of different choices of sampling priors for the design parameters, used in the simulation-based sample size calculations, on the required sample size. The focus throughout remains on CRCTs with a continuous outcome.

***

4.2 Bayesian Sample Size Calculation - A Simulation-Based Approach

Within this chapter, a Bayesian simulation-based approach to sample size determination was adopted to investigate the impact of the incorporation of historical (e.g. pilot) data in CRCT design on estimated statistical power and required sample size. An approach similar to that outlined by Wang and Gelfand [Wang and Gelfand, 2002] was proposed, extended in two ways: firstly through the specification of *sampling priors* not only for the treatment effect parameter, but also on the other design parameters for the study; and secondly, through the use of an informative *fitting prior*, namely the NPP introduced in Chapter 3.

Denote a sampling prior as $\pi^{(s)}(\cdot)$, and a fitting prior as $\pi^{(f)}(\cdot)$. Furthermore, recall that the design parameters to be specified prior to a sample size calculation for a CRCT with a continuous outcome are: (i) the SD of the outcome, denoted σ; (ii) the ICC, denoted ρ; and (iii) the MCID/target difference, denoted θ. In the simulation-based scenarios outlined within this chapter, the intercept term (i.e. the mean of the outcome in the control arm), β, must also be specified.

Suppose further that a reduction in the outcome of interest (i.e. a negative value of θ) corresponds to a successful outcome. Then a one-sided hypothesis test can be specified as

$$\begin{gathered} H_0 : \theta > 0 \\ \text{vs} \\ H_1 : \theta \leq 0 \end{gathered} \tag{4.1}$$

and a similar two-sided hypothesis test can be denoted as

$$\begin{gathered} H_0 : \theta = 0 \\ \text{vs} \\ H_1 : \theta \neq 0 \end{gathered} \tag{4.2}$$

Recall that specification of a point mass sampling prior for the treatment effect retains the frequentist interpretation of statistical power (alternative specifications for the sampling prior for the treatment effect are explored in Chapter 5). After pre-specification of a non-zero point mass *sampling prior*, $\pi^{(s)}(\theta)$ for the treatment effect, a simulation-based approach to determining the smallest required sample size to achieve a desired level of statistical power (denoted $\Pi\%$) can be devised. Specifically,

$$\pi^{(s)}(\theta) = \begin{cases} 1, & \text{if } \theta = \theta^{\star} \\ 0 & \text{otherwise} \end{cases}$$

where $\theta^{*} \neq 0$, and θ^{*} is typically specified as the MCID/target difference.

Furthermore, let $\hat{\Pi}_k$ denote the estimated statistical power simulated using k clusters per arm (assuming a two-arm trial). To initiate the algorithm, choose some value of k (denoting number of clusters per arm) such that $\hat{\Pi}_{k-1} < \Pi$. That is, that the study is underpowered with $k-1$ clusters. Let μ_m and σ_m denote the expected mean and standard deviation of the cluster sizes. Furthermore, specify a total of N iterations chosen to be large enough to ensure that power can be estimated with a suitable degree of precision, and let $i = 1,\ldots,N$. Then proceed according to the following steps:

1. Simulate N datasets, $\mathbf{D}_i, i = 1,\ldots,N$, comprising k clusters per arm of size $m_j, j = 1,\ldots,k$ where m_j is drawn from a $N(\mu_m, \sigma_m^2)$ distribution, using a sample from each of the *sampling priors* $\pi^{(s)}(\sigma)$, $\pi^{(s)}(\rho)$, $\pi^{(s)}(\beta)$ and $\pi^{(s)}(\theta)$

2. Fit the analysis model to $\mathbf{D}_i$ to obtain samples from the posterior distribution of the treatment effect:

$$\pi(\beta,\theta,\rho,\sigma^2|D) \propto L(\beta,\theta,\rho,\sigma|D)\pi^{(f)}(\beta)\pi^{(f)}(\theta)\pi^{(f)}(\sigma)\pi^{(f)}(\rho)$$

3. Store the upper and lower $(100-\alpha)\%$ credible intervals (CrIs) for the treatment effect from each $\mathbf{D}_i$, denoted $(\hat{\theta}_{i,\frac{\alpha}{2}}, \hat{\theta}_{i,1-\frac{\alpha}{2}})$

4. Calculate $\hat{\Pi}_k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\hat{\theta}_{i,\frac{\alpha}{2}} > 0 \text{ OR } \hat{\theta}_{i,1-\frac{\alpha}{2}} < 0}$

5. If $\hat{\Pi}_k < \Pi$, let $k = k+1$ and return to step 1. Else, if $\hat{\Pi}_k \geq \Pi$, terminate the algorithm and declare the minimum study size required to achieve $\Pi\%$ power to be k clusters per arm.

Once the number of clusters per arm, k, required to achieve the pre-specified level of statistical power has been determined, a similar procedure can be undertaken in order to calculate the type I error. However, in order to do so, a point mass sampling prior for θ at zero must be specified such that

$$\pi^{(s)}(\theta) = \begin{cases} 1, & \text{if } \theta = 0 \\ 0 & \text{otherwise} \end{cases}$$

One-sided type I error can be calculated as the proportion of credible intervals which fall entirely below zero, $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{\hat{\theta}_{i,1-\frac{\alpha}{2}}<0}$. Similarly, the two sided type I error can be calculated as the proportion of credible intervals which do not include zero, $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{\hat{\theta}_{i,\frac{\alpha}{2}}>0 \text{ OR } \hat{\theta}_{i,1-\frac{\alpha}{2}}<0}$.

***

4.3 An Example: The Healthy Lifestyles Programme

The HeLP study was a school-based CRCT exploring whether the implementation of a school-based intervention was effective in obesity prevention, measured using the primary outcome of BMI SDS, compared with standard education provision. During the design of the HeLP study, an SD of 1.3 and an ICC of 0.02 were assumed in order to calculate the required sample size, as well as an average cluster size of 35 and a coefficient of variation in cluster size of 0.5 (implying an SD of cluster size of 17.5). More details on the study are provided in Chapter 1, the protocol [Wyatt et al., 2013] and the results paper [Lloyd et al., 2018]. Prior to this definitive CRCT, a pilot CRCT was first undertaken [Wyatt et al., 2011].

In this chapter, the effect of incorporating the HeLP pilot data via the NPP, introduced in Chapter 3, on statistical power and required sample size/number of clusters is examined. A number of sampling-prior strategies were explored, as outlined in further detail in §4.3.1. In all cases, point mass sampling priors were specified for the treatment effect parameter in order to maintain a frequentist interpretation of power and type I error.

4.3.1 Sampling Strategies for the Design Parameters

As outlined in §4.2, uncertainty in design parameters used to determine the required sample size to achieve a pre-specified level of statistical power can be incorporated into a simulation-based power calculation through the use of a *sampling prior*. Within the context of the design of the HeLP study, a range of sampling-prior strategies for the design parameters were considered to assess the impact on power and type I error. Details of each sampling-prior strategy are outlined below.

Point Mass Sampling Priors (Sampling-Prior Strategy I)

The scenario in which point mass sampling priors are specified for each design parameter was considered, where the entire mass of the sampling priors are placed at the values that were pre-specified in the original sample size calculation for the HeLP study. Specifically, $\pi^{(s)}(\sigma) = 1$ if $\sigma = 1.3$ and $\pi^{(s)}(\sigma) = 0$ otherwise and $\pi^{(s)}(\rho) = 1$ if $\rho = 0.02$ and $\pi^{(s)}(\rho) = 0$ otherwise. The intercept term is also assumed to be drawn from a point mass sampling prior, where $\pi^{(s)}(\beta) = 1$ if $\beta = 0.5$ (the mean of the outcome in the control arm from the HeLP pilot study) and $\pi^{(s)}(\beta) = 0$ otherwise. This sampling-prior strategy is in line with the frequentist approach, which assumes at the study design stage that each parameter is fixed.

Partial Sampling Priors (Sampling-Prior Strategy II)

Under this sampling-prior strategy, a point mass prior for the ICC, ρ, was specified, with all mass at $\rho = 0.02$, and sampling priors for the SD, σ, and the intercept term, β, were derived from analysis of the pilot HeLP data. Specifically, by fitting a hierarchical linear regression model to the pilot data, using the notation and framework outlined in Equation (3.11), it was possible to obtain posterior distributions for these parameters from the model. Specifically, $\pi^{(s)}(\sigma) = \pi(\sigma|D_0, \theta, \beta, \sigma_c, \mathbf{b}_0)$ and $\pi^{(s)}(\beta) = \pi(\beta|D_0, \theta, \sigma, \sigma_c, \mathbf{b}_0)$.

Specification of a point mass sampling prior for the ICC was in recognition of the fact that there is often a great deal of uncertainty around estimation of ICCs from pilot CRCTs, particularly from those with a small number of clusters such as in the HeLP pilot study, which included only 185 participants across four schools (clusters). This uncertainty has the potential to reduce statistical power within a simulation-based sample size calculation as the posterior distribution of the ICC may include very large values. Furthermore, Eldirdge et al. [Eldridge et al., 2015] cautioned against using pilot or feasibility data alone to inform estimates of the ICC for subsequent sample size calculation.

Full Sampling Priors (Sampling-Prior Strategy III)

The full sampling-prior strategy involved the use of sampling priors informed by analysis of the pilot data for all design parameters. In addition to the sampling priors for σ and β informed by the pilot data as specified in sampling-prior strategy II, the posterior distribution of the ICC obtained from analysis of the pilot data, was also used as a sampling prior. Specifically, $\pi^{(s)}(\rho) = \pi(\rho|D_0, \theta, \beta, \mathbf{b}_0)$.

Full Sampling Priors with Meta-Analysed ICCs (Sampling-Prior Strategy IV)

Turner et al. [Turner et al., 2004] proposed a meta-analytic approach for synthesising ICC estimates from multiple studies, in order to obtain a posterior distribution to inform power or sample size calculation. In this final sampling-prior strategy, in which full sampling priors were specified for all design parameters, the sampling prior for the ICC was determined using this meta-analytic method instead of simply relying on the analysis of the pilot data.

The approach proposed by Turner et al. [Turner et al., 2004] proceeds as follows. Suppose that r relevant ICC estimates are obtained from the literature, denoted $\hat{\rho}_{1l}, l = 1, \ldots, r$. Furthermore, denote the total sample size and total number of clusters (across both arms) associated with $\hat{\rho}_{1l}$ by N_{1l} and k_{1l} respectively. Assume that each $\hat{\rho}_{1l}$ is normally distributed around an underlying ICC, ρ_l (which is assumed to be heterogeneous), and a distribution is assumed for the set of ICCs $\rho_l, l = 1, \ldots, r$. In this case, a Normal distribution with mean μ_ρ and variance σ^2_ρ, truncated to the interval $[0,1]$, is assumed. Using the distributional assumption for the variance of an ICC proposed by Swiger et al. [Swiger et al., 1964], the meta-analytic method proposed by Turner et al. [Turner et al., 2004] can be expressed as follows:

$$\hat{\rho}_{1l} \sim \mathrm{N}(\rho_l, \mathrm{Var}(\widehat{\rho_{1l}})), \mathrm{Var}(\hat{\rho}_{1l}) = \frac{2(N_{1l}-1)(1-\rho_l)^2\left[1+\left(\frac{N_{1l}}{k_{1l}}-1\right)\rho_l\right]^2}{\left(\frac{N_{1l}}{k_{1l}}\right)^2(N_{1l}-k_{1l})(k_{1l}-1)}, l = 1, \ldots, r$$

$$\rho_l \sim \text{truncated } \mathrm{N}(\mu_\rho, \sigma^2_\rho)$$

$$\mu_\rho \sim \mathrm{Uniform}(0,1)$$

$$\sigma^2_\rho \sim \mathrm{Uniform}(0,10)$$

Samples from the posterior distribution of each ρ_l are obtained, and can be subsequently used to inform future study design. It is worth noting that alternative distributional assumptions for the ICC (as also outlined by Turner et al. [Turner et al., 2004]), as well as for the mean and variance parameters, may also be appropriate, but are not explored further here.

A literature review was undertaken in order to identify relevant CRCTs of weight loss or obesity prevention interventions delivered in a school setting. Six studies were identified, although they did differ from the HeLP target population in key demographic characteristics such as country and age range: (i) the Health in Adolescents (HEIA) study [Grydeland et al., 2014]; (ii) the "fun 'n healthy in Moreland" (FnHiM) trial [Waters et al., 2017]; (iii) the WAVES trial [Adab et al., 2018]; (iv) the Childhood Obesity in China (COiC) trial [Liu et al., 2019]; (v) the CHIRPY DRAGON trial [Li et al., 2019] and (vi) the Daily Mile trial [Breheny et al., 2020]. For each of these trials, the point estimate of the ICC ($\hat{\rho}_{1l}$), the total sample size (N_{1l}) and the total number of clusters (k_{1l}) were extracted. The relevant statistics are shown in Table 4.1.

Synthesis of these ICC estimates using the methodology proposed by Turner et al. [Turner et al., 2004] provided an evidence-based method of specifying a sampling prior

Table 4.1: Table of ICC point estimates, total sample sizes and total number of clusters for each of the six relevant studies.

Study	$\hat{\rho}_{1l}$	N_{1l}	k_{1l}
HEIA [Grydeland et al., 2014]	0.02	1324	37
FnHiM [Waters et al., 2017]	0.008	2806	22
WAVES [Adab et al., 2018]	0.01	1392	54
COiC [Liu et al., 2019]	0.05	1839	12
CHIRPY DRAGON [Li et al., 2019]	0.118	1562	40
The Daily Mile [Breheny et al., 2020]	0.001	1670	37

for the ICC using a range of sources, and is more in line with the recommendations of Eldridge et al. [Eldridge et al., 2015] who recommended that ICC assumptions used in study design are informed not by a single pilot study, but rather by a variety of evidence. Formally, the sampling prior for the ICC under this sampling strategy can be written as $\pi^{(s)}(\rho) = \pi(\rho|D_{1l}), l = 1, \ldots, 6$, where D_{1l} denotes data from the l^{th} study identified from the literature. The sampling priors for the other design parameters (σ, β) remain as in sampling-prior strategies II and III, above.

4.3.2 Calculating Power and Type I Error

Simulation-based estimates of power and type I error (one-sided and two-sided) were calculated according to the methodology outlined in §4.2, with the dataset $\mathbf{D}$ simulated according to the following data generating mechanism:

$$\begin{aligned} Y_{i,j} &\sim \mathrm{N}(\beta + \theta x_{i,j} + b_i, \sigma^2) \\ b_i &\sim \mathrm{N}(0, \sigma_c^2) \end{aligned}$$

where $Y_{i,j}$ is the outcome for participant j in cluster i, $i = 1, \ldots, 2k$ and $j = 1, \ldots, m_i$, where k is the number of clusters per arm and m_i is the number of participants in cluster i. Furthermore, $x_{i,j}$ is the indicator term for the treatment arm, β is the intercept term, θ is the treatment effect, σ^2 is the within-cluster variance, b_i is the cluster-level random effect term and σ_c^2 is the between-cluster variance. An equal number of clusters per arm was assumed (i.e. a $1:1$ randomisation, in line with the original HeLP design), and the variability in cluster size was accounted for in the simulation-based power calculation by drawing $2k$ samples from $m_i \sim \mathrm{Normal}(35, 17.5)$ distribution (derived by assuming a mean cluster size of 35 and a coefficient of variation of cluster size of 0.5) at each iteration in order to determine the size of each cluster.

For each of the four sampling strategies outlined in §4.3.1, operating characteristics were explored for three analysis strategies: (i) frequentist hierarchical model; (ii) Bayesian hierarchical model with non-informative prior distributions and (iii) the normalised power prior approach. Power and Type I error, calculated according to the closed formula [Rutterford et al., 2015], are presented, using the assumptions prespecified in the HeLP sample size calculation [Wyatt et al., 2013]. Specifically, an SD of 1.3 units, an MCID of 0.25 units, an average cluster size of 35 and a coefficient of variation in cluster size of of 0.5 were assumed.

A total of 1500 simulations were run for each combination of sampling and analysis strategies, chosen to ensure a manageable computational burden, but also allowing estimation of power at 80% alongside a 95% confidence interval (CI) with precision of ± 2%, or 90% with precision ± 1.5%. Any iterations in which the MCMC diagnostic criteria indicated potential convergence issues were removed, using the same criteria outlined in §3.6.2. Specifically, an iteration was removed if the effective sample size was below 400, if the value of $\hat{R}$ was less than 1.1, or if more than 0.1% of the post-warm up iterations were divergent transitions in the Hamiltonian Monte Carlo procedure.

4.3.3 Results

Sampling Priors

The first stage in the simulation-based power calculation was to obtain samples from the sampling priors outlined in §4.3.1. The sampling prior for the standard deviation of the outcome, σ, is illustrated in Figure 4.1. The median value of the sampling prior was 1.2 units, which is lower than the point mass prior of 1.3 units assumed in the original sample size calculation, although this value is contained within the 95% CrI (1.1 to 1.4).

The sampling priors used for the ICC with sampling strategies III and IV are shown in Figure 4.2, with the sampling prior derived from the pilot data shown in Figure 4.2a) (sampling-strategy III), and from the multiple historical trials shown in Figure 4.2b) (sampling-strategy IV). The median value of both sampling priors is slightly larger than the point mass prior concentrated at $\rho = 0.02$. Furthermore, it can be seen that the tail of the sampling prior derived from the pilot data alone is far heavier than the one derived from the multiple historical trials, with an upper 95% CrI limit of 0.399 for the former, and 0.131 for the latter. This indicated a higher degree of uncertainty in the posterior distribution when considering the pilot data alone compared with considering the wider literature. This is consistent with the findings of Eldridge et al. [Eldridge et al., 2015], who stressed the importance of using multiple sources to justify the ICC used in sample size calculation, rather than simply using an estimate from a pilot study.

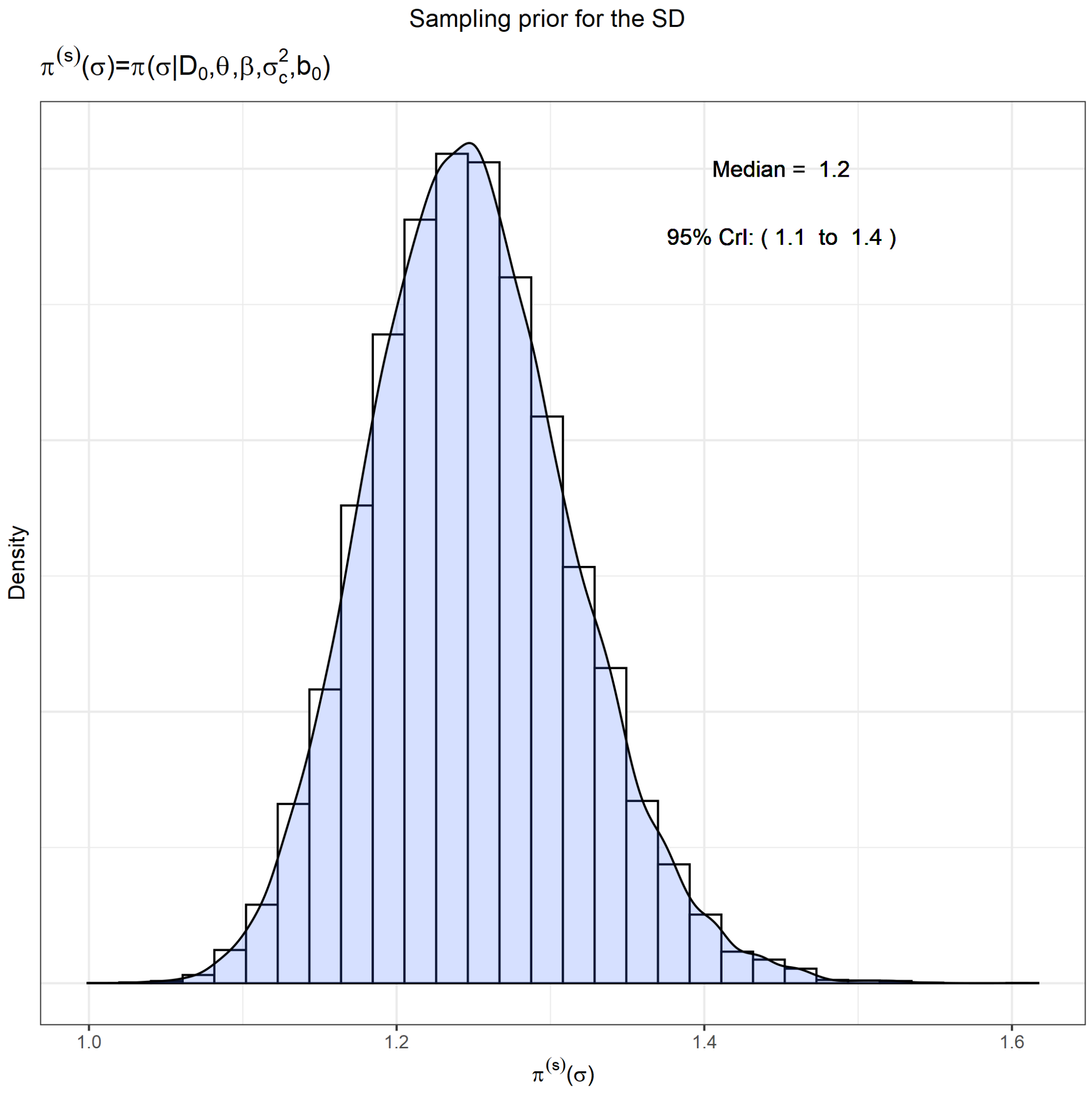


Figure 4.1: Posterior density of the SD derived through analysis of the pilot HeLP data.

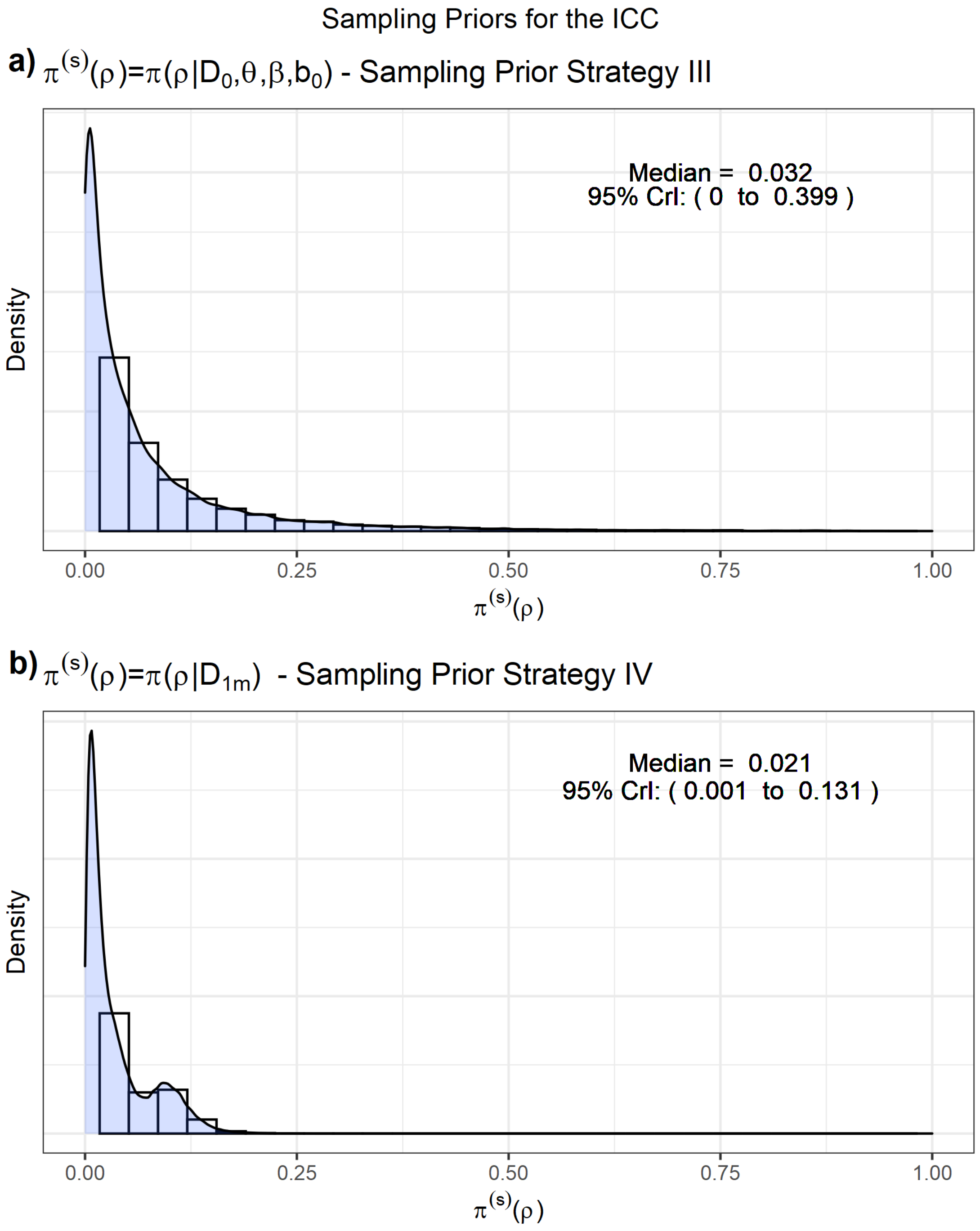


Figure 4.2: Posterior density of the ICC derived through analysis of the pilot HeLP data a) and through synthesis of ICC estimates obtained from six trials reported in the literature b).

Power and Required Sample Size

Incorporation of the data collected from the HeLP pilot study through the NPP analysis strategy resulted in an increase in statistical power when compared to both frequentist and Bayesian hierarchical models across all four sampling strategies.

The results of the power calculations are shown in Figure 4.3. The formula-based power calculation indicates that a total of 31 clusters per arm was required to achieve 90% power to detect a difference in BMI SDS of 0.25 units, assuming a 5% two-sided significance level (equivalently, a one-sided 2.5% significance level). Under sampling-prior strategy I (point mass sampling priors), the simulation-based power calculation indicated that a total of 26 clusters per arm was required to achieve 90% statistical power to detect a difference in BMI SDs of 0.25 units when using an NPP analysis; using either frequentist or Bayesian analyses required a total of 31 clusters per arm. The results under sampling-prior strategy II (partial sampling priors) were similar; again, the simulation-based power calculation indicated that 26 clusters per arm were needed to achieve 90% power when employing the NPP analysis strategy, and 29 per arm were needed when employing a hierarchical Bayesian or frequentist model. Under sampling-prior strategy III (full sampling priors), significant reductions in power were observed across all three analysis strategies compared with sampling-prior strategies I and II, with more than 41 clusters per arm (the upper limit of the number of clusters simulated) required to achieve 90% power. Despite this, statistical power was consistently higher when using the NPP analysis compared to either the Bayesian or frequentist random effects models. Sampling-prior strategy IV (with meta-analysed ICCs) also saw a substantial reduction in statistical power compared to sampling-prior strategies I and II, although to a lesser extent than sampling-prior strategy III. The simulation-based power calculations indicated that 38 clusters per arm were required to achieve 90% power when using the NPP analysis strategy, 39 per arm were required to achieve 90% power when using a frequentist random effects model, and 42 per arm were required when using a Bayesian random effects model.

Type I Error

Given two-sided 95% CrIs or CIs were used to determine acceptance or rejection of the null hypothesis at each iteration within the simulation study, the expected two-sided type I error is 5%, which is equivalent to a one-sided type I error of 2.5%. As can be seen in Figure 4.4, under all sampling and analysis strategies, the two-sided type I error was reasonably well-controlled at around 5% in most cases. However, Figure 4.5 illustrates substantial inflation of the one-sided type I error rate above the nominal rate of 2.5% when using the NPP strategy, although the one-sided type I error remained well-controlled without any information borrowing.

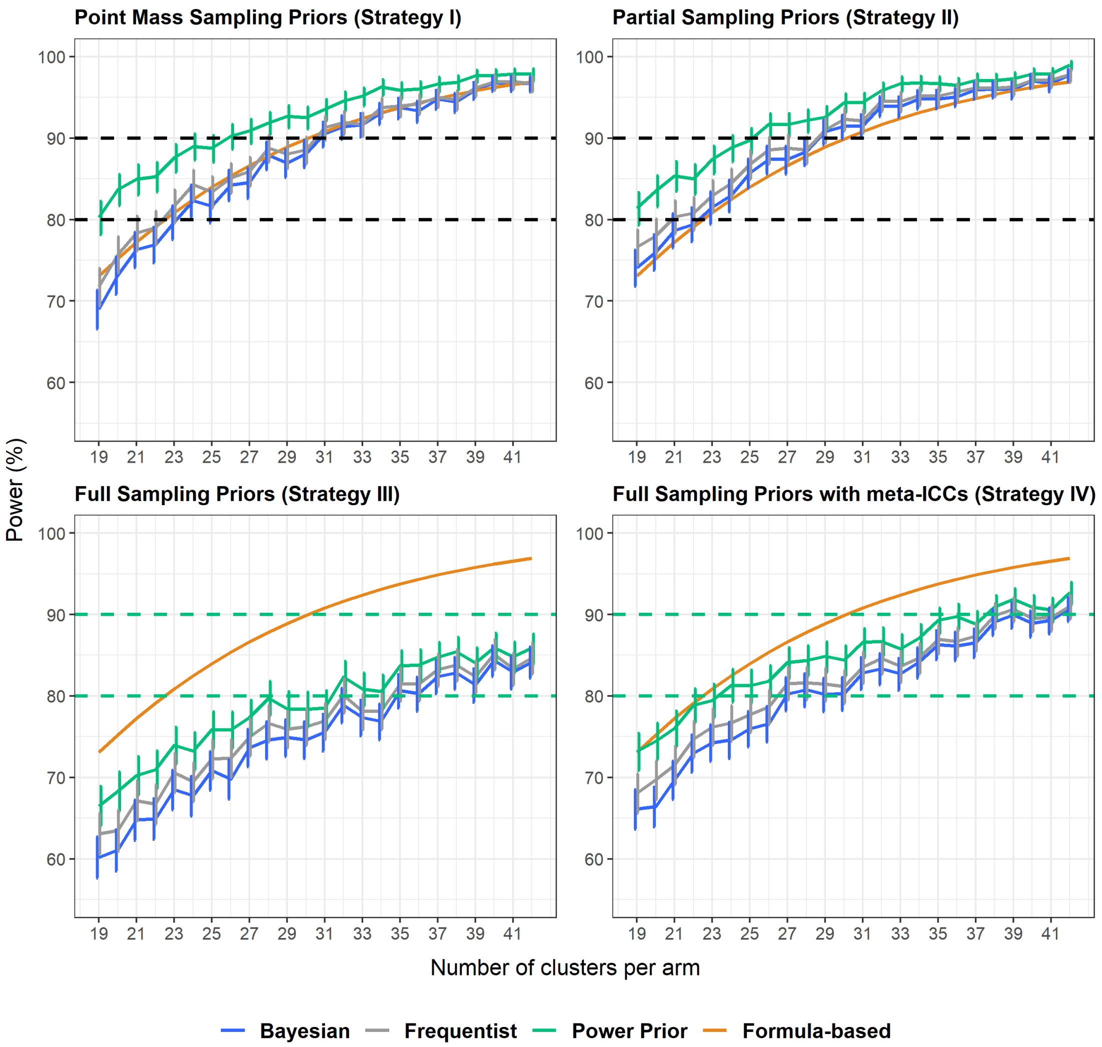


Figure 4.3: Power curves for each of the four sampling strategies, when analysing the simulated data using Bayesian or frequentist random effects models, using the normalised power prior (NPP) or when using a formula-based approach for calculating power. Error bars represent 95% confidence intervals.

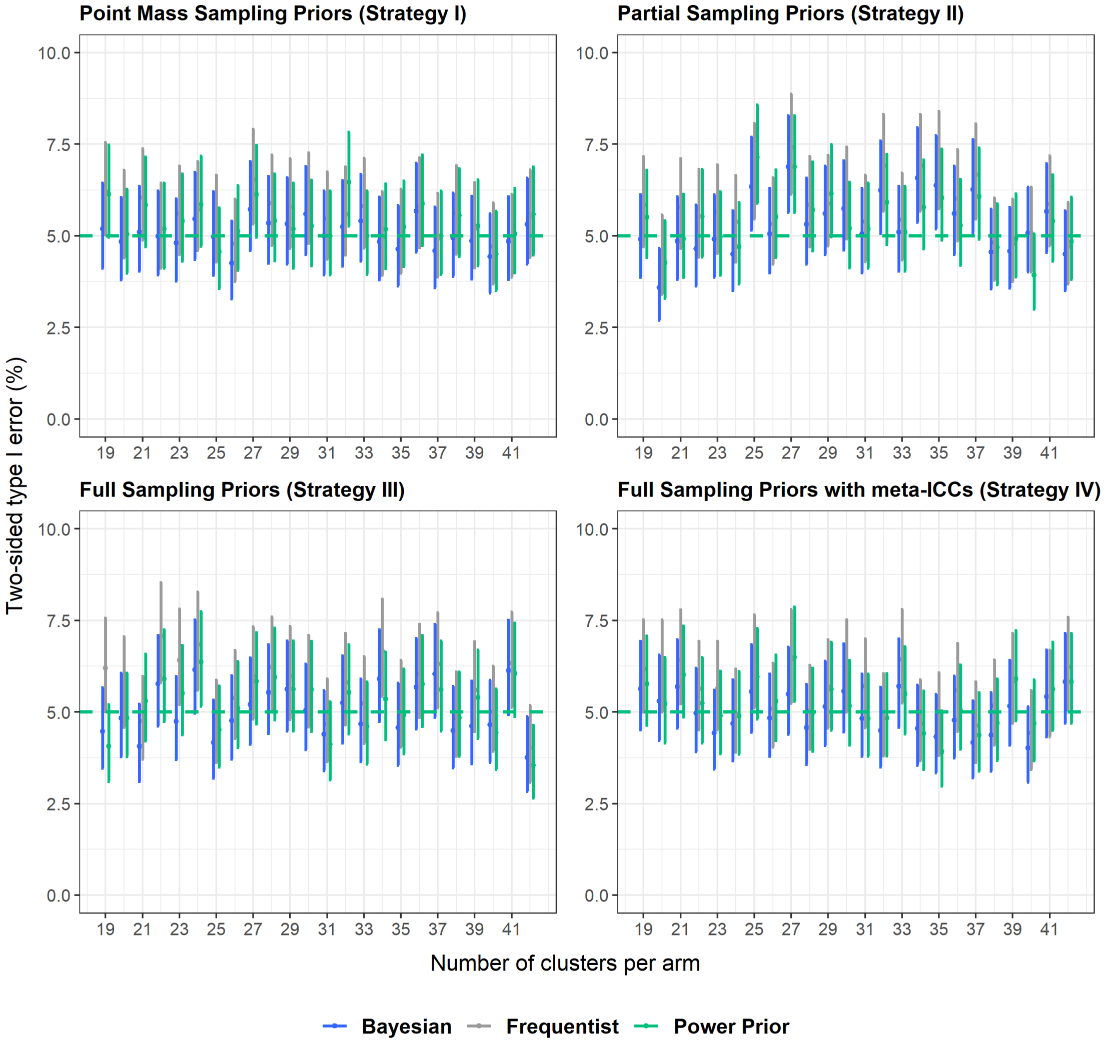


Figure 4.4: Estimated two-sided type I error for each of the four sampling strategies, when analysing the simulated data using Bayesian or frequentist random effects models, or using the normalised power prior (NPP). Error bars represent 95% confidence intervals.

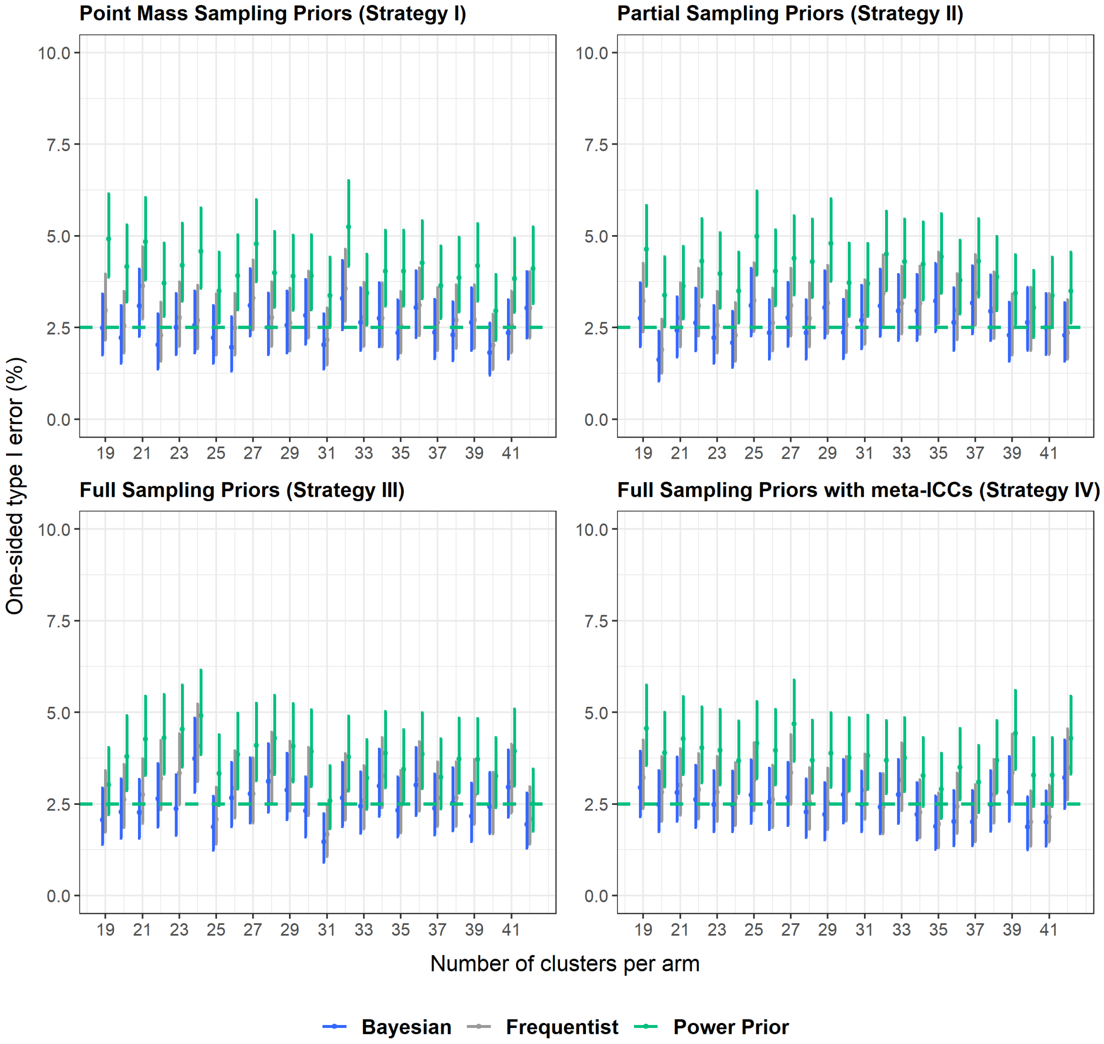


Figure 4.5: Estimated one-sided type I error for each of the four sampling strategies, when analysing the simulated data using Bayesian or frequentist random effects models, or using the normalised power prior (NPP). Error bars represent 95% confidence intervals.

The fact that the two-sided type I error was well-controlled, whereas its one-sided counterpart was not, at first appears surprising. However, this can be explained by the fact that the historical information from the HeLP study incorporated within the analysis through the NPP provides evidence for a negative treatment effect (i.e. a reduction in BMI SDS). Recall that the two-sided type I error rate was estimated as the proportion of iterations for which the null hypothesis was rejected in favour of an alternative hypothesis *in either direction* (i.e. a treatment effect above or below zero). Introducing the HeLP pilot data through the NPP resulted in a reduction in the number of iterations rejecting the null hypothesis *above* zero, but an increase in the number of iterations *below* zero. As a result, in this particular scenario, the two-sided type I error rate balanced out, giving the (incorrect) impression of reasonable control. By considering the one-sided type I error, only the increase in incorrect rejections of the null hypothesis *below* zero was captured. As a result, when incorporating external evidence through the power prior approach (or indeed other informative prior distributions), a one-sided type I error is a more appropriate measure to evaluate and control for.

A summary of the results of these simulation-based power calculations, including the number of clusters per arm required to achieve 80% and 90% power under the NPP analysis, alongside the estimated one-sided and two-sided type I error rates is shown in Table 4.2.

Table 4.2: Required number of clusters per arm to achieve 80% and 90% power under the NPP approach, alongside estimated one-sided and two-sided type I error, for each of the four sampling strategies.

	80% Power			90% Power		
Sampling Strategy	k	One-sided error	Two-sided error	k	One-sided error	Two-sided error
I	19	4.9%	6.1%	26	3.9%	5.1%
II	19	4.6%	5.5%	26	4.0%	5.5%
III	32	3.8%	5.5%	$>$42	-	-
IV	24	3.7%	4.8%	38	3.7%	4.7%

Conclusion

It is clear that the incorporation of the HeLP pilot data into the planned definitive trial analysis through the NPP method could have justified a reduction in the required sample size compared to the traditional formula-based approach to sample size determination. However, this would have resulted in an inflated one-sided type I error rate. Furthermore, sampling priors could have been considered to formally account for the uncertainty in the underpinning assumptions for the SD without notable detriment to statistical power. However, placing a sampling prior other than a point-mass sampling prior on the ICC would have significantly increased the required number of clusters, beyond even the reduction that could have been realised through the NPP analysis.

***

4.4 A Simulation Study

4.4.1 Design

A simulation study was undertaken to explore and quantify the potential impact of the incorporation of historical data, such as data obtained from pilot or feasibility studies, on statistical power or required sample size for a CRCT with a continuous primary outcome. Specifically, interest lies in determining if and when it may be possible to justify a reduction in required sample size for a definitive study as a result of constructing an informative power prior for a CRCT analysis based on historical data (such as data obtained from a pilot CRCT), and understanding any compromise of doing so in the context of type I error inflation. In this simulation study, the NPP was compared to a frequentist analysis of the simulated definitive trial data alone using a hierarchical linear model.

The simulation study involved generation of both pilot data (from which the power prior was constructed) and definitive CRCT data. In order to replicate a typical scenario in which a pilot study has been completed, and a subsequent fully-powered trial is being designed, each pilot dataset was simulated only once, whereas the definitive trial data was simulated at each iteration. Similarly to the simulation study presented in §3.6.1, it was assumed that no adjustment for additional covariates was made in the sample size calculation. As previously, the data generating mechanism for the definitive trial data was

$$\begin{aligned} Y_{i,j} &\sim \mathrm{N}(\beta + \theta x_{i,j} + b_i, \sigma^2) \\ b_i &\sim \mathrm{N}(0, \sigma_c^2) \end{aligned} \tag{4.3}$$

and similarly for the pilot trial data was

$$\begin{aligned} Y_{0\tilde{i},\tilde{j}} &\sim \mathrm{N}(\beta + \theta_0 x_{0\tilde{i},\tilde{j}} + b_{0\tilde{i}}, \sigma^2) \\ b_{0\tilde{i}} &\sim \mathrm{N}(0, \tilde{\sigma}_c^2) \end{aligned} \tag{4.4}$$

where $Y_{i,j}$ represents the outcome for participant j in cluster i, $i = 1,\dots,2k$, $j = 1,\dots,m_i$ in the definitive trial and $Y_{0\tilde{i},\tilde{j}}$ represents the outcome for participant $\tilde{j}$ in cluster $\tilde{i}$,

$\tilde{i} = 1, \ldots, 2k_0$, $\tilde{j} = 1, \ldots, m_{\tilde{i}}$ in the pilot trial, where k and k_0 represent the number of clusters per arm in the definitive and pilot trials, respectively, and m_i represents the number of participants in cluster i. In addition, β is the intercept term, θ and θ_0 are the treatment effects for the definitive and pilot studies, respectively, and b_i and $b_{0\tilde{i}}$ are the cluster level random effects for the definitive and pilot trials, respectively. Furthermore, σ denotes the within-cluster standard deviation (so assumes homogeneous variation across clusters), and σ_c denotes the between-cluster variation.

Simulation of the Pilot Data

A total of twelve historical datasets were simulated. In each, the cluster size (m) was fixed at fifteen (thus simulating a typical situation for a cluster trial, for example in schools), the ICC, ρ_0, at 0.05 (a typical guideline often used in CRCT sample size calculations) and the standard deviation, σ, at 1 (to facilitate consideration of standardised effect sizes). The number of clusters per arm (k_0) and the treatment effect (θ_0) were varied to reflect a range of possible scenarios for a pilot CRCT. Specifically, all combinations of $k_0 = 2,\ 4,\ 6,\ 8$ and $\theta_0 = 0,\ -0.2,\ -0.4$ were generated. The parameters underpinning the data generating mechanism and key summary statistics are presented for each simulated pilot dataset in Table 4.3.

Table 4.3: Summary statistics for the simulated pilot datasets.

Pilot Dataset #	k_0	θ_0	$\hat{\theta}_0$ (95% CrI)	$\hat{\rho}_0$ (95% CrI)
1	2	-0.4	-0.37 (-1.2 to 0.44)	0.03 (0 to 0.41)
2	4	-0.4	-0.40 (-0.99 to 0.19)	0.05 (0 to 0.28)
3	6	-0.4	-0.37 (-0.73 to 0.01)	0.02 (0 to 0.15)
4	8	-0.4	-0.42 (-0.72 to -0.13)	0.02 (0 to 0.12)
5	2	-0.2	-0.17 (-1.06 to 0.63)	0.03 (0 to 0.47)
6	4	-0.2	-0.24 (-0.66 to 0.17)	0.03 (0 to 0.22)
7	6	-0.2	-0.21 (-0.58 to 0.16)	0.02 (0 to 0.13)
8	8	-0.2	-0.21 (-0.54 to 0.13)	0.05 (0 to 0.18)
9	2	0	0.04 (-0.77 to 0.84)	0.03 (0 to 0.39)
10	4	0	-0.06 (-0.61 to 0.50)	0.06 (0 to 0.32)
11	6	0	0.00 (-0.38 to 0.36)	0.02 (0 to 0.13)
12	8	0	0.09 (-0.28 to 0.48)	0.05 (0 to 0.18)

Simulation of the Definitive Trial Data

A total of 36 scenarios were simulated across 1500 iterations. Whilst each pilot dataset was simulated only once, the definitive trial data was repeatedly generated over each iteration in order to replicate the situation under which this simulation-based power calculation may be used in practice, namely in the period following completion of a pilot/feasibility study when a definitive study is being designed. A range of target effect sizes was considered, and the range of number of clusters used (k_{min} to k_{max}) was informed by the number of clusters required to achieve 80% power using the frequentist formula-based approach. Target power was set at 80% in order to ease the computational burden of the simulation study, as to identify the study size required for 90% power required a much larger range of study sizes to be considered. In each simulated dataset, the cluster size was fixed at fifteen.

The parameters underpinning the data generating mechanism for each of the scenarios are shown in Table 4.4, as well as the number of clusters per arm required to achieve 80% power using the frequentist formula-based approach to sample size calculation. Three strategies were used for defining sampling priors: (i) Point mass sampling priors for all parameters, with mass for the ICC at $\rho = 0.05$, for the SD at $\sigma = 1$ and for the intercept at $\beta = 1$; (ii) Partial sampling priors, with sampling priors for the SD and the intercept defined by posterior samples obtained through fitting a Bayesian random effects model to the pilot data, and a point mass sampling prior for the ICC at $\rho = 0.05$ and (iii) full sampling priors, with sampling priors for the SD, the intercept and the ICC obtained through fitting a Bayesian random effects model to the pilot data.

Model Formulation and Posterior Sampling

The prior distributions for both the Bayesian hierarchical model and the NPP model were the same as specified in §3.6.2. Specifically, the regression coefficients had N$(0,5)$ priors, the within-group SD had a prior of Exp(1) and the between-group SD had a Half-Normal$(0,1.5)$ prior. The discounting factor, a_0, had a non-informative prior, namely Beta$(1,1)$. Analyses of the simulated pilot data alone was run for $4,000$ iterations across four parallel chains, with the first $2,000$ of each discarded. Each of the NPP models was run for $3,000$ iterations across four chains with the first $1,500$ discarded. The approximation of $C(a_0)$, as outlined in §3.3, used $\Delta = 20$ values of a_0, with the model at each value of δ run across four chains for a total of $3,500$ iterations per chain with the first $1,750$ discarded. The Bayesian hierarchical models were each run for $2,500$ iterations across four chains, with the first $1,250$ discarded.

The numbers of iterations used for each model were chosen to balance the need for more iterations to reach convergence in more complex models, against the computational cost of running computationally intensive MCMC procedures repeatedly within each iteration of a simulation study. A large number of iterations was used for the analysis of each of the pilot datasets alone in order to ensure convergence of the posterior distributions, particularly given the relatively small sample sizes (in terms of numbers of clusters) in some of the scenarios. Computational cost was not a significant issue in this instance, as these datasets were not re-analysed during each iteration of the sim-

Table 4.4: Simulation study scenarios

	Pilot Data			Definitive Trial Design Parameters			
Scenario #	Pilot Dataset #	k_0	θ_0	Target Difference	k^{a}	k_{min}	k_{max}
1.1	1	2	-0.4	-0.4	13	6	17
1.2	2	4	-0.4	-0.4	13	6	17
1.3	3	6	-0.4	-0.4	13	6	17
1.4	4	8	-0.4	-0.4	13	6	17
2.1	5	2	-0.2	-0.4	13	6	17
2.2	6	4	-0.2	-0.4	13	6	17
2.3	7	6	-0.2	-0.4	13	6	17
2.4	8	8	-0.2	-0.4	13	6	17
3.1	9	2	0	-0.4	13	6	17
3.2	10	4	0	-0.4	13	6	17
3.3	11	6	0	-0.4	13	6	17
3.4	12	8	0	-0.4	13	6	17
4.1	1	2	-0.4	-0.3	21	17	29
4.2	2	4	-0.4	-0.3	21	17	29
4.3	3	6	-0.4	-0.3	21	17	29
4.4	4	8	-0.4	-0.3	21	17	29
5.1	5	2	-0.2	-0.3	21	17	29
5.2	6	4	-0.2	-0.3	21	17	29
5.3	7	6	-0.2	-0.3	21	17	29
5.4	8	8	-0.2	-0.3	21	17	29
6.1	9	2	0	-0.3	21	17	29
6.2	10	4	0	-0.3	21	17	29
6.3	11	6	0	-0.3	21	17	29
6.4	12	8	0	-0.3	21	17	29
7.1	1	2	-0.4	-0.2	46	35	68
7.2	2	4	-0.4	-0.2	46	35	68
7.3	3	6	-0.4	-0.2	46	35	68
7.4	4	8	-0.4	-0.2	46	35	68
8.1	5	2	-0.2	-0.2	46	35	68
8.2	6	4	-0.2	-0.2	46	35	68
8.3	7	6	-0.2	-0.2	46	35	68
8.4	8	8	-0.2	-0.2	46	35	68
9.1	9	2	0	-0.2	46	35	68
9.2	10	4	0	-0.2	46	35	68
9.3	11	6	0	-0.2	46	35	68
9.4	12	8	0	-0.2	46	35	68

a: Number of clusters required to achieve 80% power using the formula-based approach ulation study. The NPP models were more complex, and so it was especially important to maximise the number of iterations but also crucial to consider the computational cost given these models were run at each iteration during the simulation study. It was again important to allow sufficient iterations for convergence of the models used to calculate $C(a_0)$, as datasets with few clusters were modelled, and again these analyses were not fitted at each iteration within the simulation study. Finally, fewer iterations were used for the "standard" Bayesian random effects models, as these models were less complex and therefore achieving convergence was more straightforward. Given these models were fitted at each iteration of the simulation study, this presented an opportunity to reduce the overall computational cost.

Convergence diagnostic statistics were collected at each iteration in the simulation study, and the iteration was discarded if these statistics suggested that valid and reliable inference based on the posterior samples was not possible. Specifically, any iteration with an $\hat{R} \geq 1.1$, an effective sample size of less than 400, or where more than 0.1% of the post warmup iterations were divergent transitions, was discarded.

4.4.2 Results

Table 4.5 shows the results of the simulation study, indicating the minimum number of clusters per arm required to achieve 80% power for each of the 36 scenarios under both the frequentist and NPP analysis strategies, using fixed, partial or full sampling priors for the simulation-based power calculations. The one-sided type I error for the NPP analysis strategy is also presented. The green arrows indicate a reduction in the required number of clusters when adopting the NPP analysis method in comparison to the frequentist method, the red arrows indicate an increase and the black arrows indicate no change.

Table 4.6 shows the mean point estimate of the treatment effect and the associated mean precision (half the width of the 95% CrI for the Bayesian models, or of the 95% CI for the frequentist model) for each scenario under each sampling-prior strategy, when simulating the smallest number of clusters per arm required to achieve 80% power when employing the NPP analysis method (k).

Table 4.5: The minimum number of clusters required to achieve 80% power when using frequentist and NPP analysis methods, alongside the one-sided type I error under the NPP analysis.

	Fixed Sampling Priors				Partial Sampling Priors				Full Sampling Priors			
Scenario	Frequentist	NPP[a]		Type I Error	Frequentist	NPP[a]		Type I Error	Frequentist	NPP[a]		Type I Error
1.1	12	10	↓	6.3%	14	11	↓	4.3%	14	13	↓	3.7%
1.2	13	8	↓	6.4%	15	11	↓	5.0%	16	12	↓	5.8%
1.3	12	7	↓	10.2%	13	7	↓	10.1%	11	6	↓	9.6%
1.4	12	6	↓	16.6%	11	6	↓	15.2%	9	6	↓	13.9%
2.1	12	12	←→	3.7%	14	13	↓	3.4%	15	15	←→	2.6%
2.2	13	10	↓	5.7%	8	7	↓	5.1%	8	7	↓	5.3%
2.3	12	11	↓	5.3%	13	11	↓	5.4%	11	9	↓	4.7%
2.4	12	10	↓	7.8%	12	9	↓	5.7%	12	9	↓	5.1%
3.1	12	13	↑	2.4%	13	15	↑	2.2%	14	17	↑	1.9%
3.2	13	14	↑	2.2%	12	13	↑	2.1%	16	16	←→	1.9%
3.3	12	14	↑	2.2%	13	14	↑	1.5%	11	12	↑	1.5%
3.4	12	>17	↑	-	13	>17	↑	-	15	>17	↑	-
4.1	22	19	↓	5.1%	23	20	↓	4.1%	24	21	↓	3.0%
4.2	21	17	↓	7.0%	25	20	↓	5.2%	>29	24	↓	3.6%
4.3	21	17	↓	7.5%	22	17	↓	8.2%	19	17	↓	6.0%
4.4	22	17	↓	9.0%	18	17	↓	10.8%	17	17	←→	8.1%

Table 4.5: The minimum number of clusters required to achieve 80% power when using frequentist and NPP analysis methods, alongside the one-sided type I error under the NPP analysis (continued).

	Fixed Sampling Priors			Partial Sampling Priors			Full Sampling Priors		
Scenario	Frequentist	NPP[a]	Type I Error	Frequentist	NPP[a]	Type I Error	Frequentist	NPP[a]	Type I Error
5.1	21	20 ↓	3.2%	23	22 ↓	2.9%	25	25 ←→	3.4%
5.2	21	18 ↓	5.0%	17	17 ←→	5.0%	17	17 ←→	4.7%
5.3	21	18 ↓	4.7%	22	18 ↓	4.5%	19	17 ↓	3.8%
5.4	22	17 ↓	6.1%	20	17 ↓	6.4%	21	17 ↓	5.5%
6.1	22	22 ←→	3.1%	23	23 ←→	1.7%	24	26 ↑	2.6%
6.2	21	22 ↑	2.4%	20	21 ↑	2.8%	27	27 ←→	2.0%
6.3	21	23 ↑	2.0%	24	25 ↑	2.2%	19	21 ↑	2.3%
6.4	22	27 ↑	1.3%	25	29 ↑	1.4%	26	>27 ↑	-
7.1	47	43 ↓	3.7%	52	47 ↓	4.2%	54	49 ↓	3.6%
7.2	47	37 ↓	4.3%	55	47 ↓	4.2%	65	54 ↓	4.4%
7.3	48	35 ↓	6.1%	48	35 ↓	5.9%	40	35 ↓	5.5%
7.4	47	35 ↓	7.1%	42	35 ↓	7.9%	35	35 ←→	6.6%
8.1	48	46 ↓	3.7%	52	50 ↓	3.5%	52	52 ←→	3.0%
8.2	47	42 ↓	3.8%	35	35 ←→	4.7%	35	35 ←→	3.4%
8.3	48	40 ↓	3.9%	52	44 ↓	4.3%	41	35 ↓	4.5%
8.4	47	37 ↓	5.8%	43	35 ↓	5.3%	45	36 ↓	5.7%

Table 4.5: The minimum number of clusters required to achieve 80% power when using frequentist and NPP analysis methods, alongside the one-sided type I error under the NPP analysis (continued).

	Fixed Sampling Priors			Partial Sampling Priors			Full Sampling Priors		
Scenario	Frequentist	NPP[a]	Type I Error	Frequentist	NPP[a]	Type I Error	Frequentist	NPP[a]	Type I Error
9.1	45	48 ↑	2.5%	51	52 ↑	2.7%	55	56 ↑	2.3%
9.2	47	48 ↑	2.5%	45	45 ←→	2.3%	58	58 ←→	3.0%
9.3	48	48 ←→	2.2%	52	52 ←→	2.8%	40	41 ↑	2.8%
9.4	47	56 ↑	2.0%	55	62 ↑	1.8%	60	63 ↑	1.4%

[a] ↓ illustrates a reduction in the required number of clusters when using the NPP analysis method compared to the frequentist method; ←→ indicates no difference in the required number of clusters; and ↑ indicates an increase in the number of clusters.

Table 4.6: Mean point estimates of the treatment effect and 95% interval widths for each scenario when using the minimum number of clusters per arm required to achieve 80% power when using the NPP analysis method (from Table 4.5).

	Fixed Sampling Priors					Partial Sampling Priors					Full Sampling Priors				
Scenario		Frequentist		NPP			Frequentist		NPP			Frequentist		NPP	
	k	PE[a]	IW[b]	PE[a]	IW[b]	k	PE[a]	IW[b]	PE[a]	IW[b]	k	PE[a]	IW[b]	PE[a]	IW[b]
1.1	10	-0.42	0.30	-0.41	0.28	11	-0.40	0.30	-0.40	0.28	13	-0.40	0.29	-0.40	0.28
1.2	8	-0.40	0.33	-0.40	0.29	11	-0.40	0.31	-0.40	0.28	12	-0.40	0.33	-0.40	0.28
1.3	7	-0.40	0.36	-0.39	0.27	7	-0.40	0.36	-0.39	0.28	6	-0.39	0.37	-0.39	0.29
1.4	6	-0.41	0.38	-0.41	0.27	6	-0.40	0.36	-0.41	0.26	6	-0.40	0.33	-0.41	0.25
2.1	12	-0.40	0.27	-0.37	0.26	13	-0.40	0.27	-0.38	0.26	15	-0.40	0.27	-0.38	0.26
2.2	10	-0.39	0.30	-0.37	0.26	7	-0.39	0.30	-0.35	0.26	7	-0.40	0.30	-0.36	0.26
2.3	11	-0.40	0.28	-0.36	0.25	11	-0.40	0.30	-0.35	0.25	9	-0.40	0.30	-0.35	0.26
2.4	10	-0.40	0.30	-0.35	0.25	9	-0.40	0.30	-0.34	0.24	9	-0.40	0.31	-0.34	0.25
3.1	13	-0.40	0.26	-0.36	0.26	15	-0.40	0.26	-0.37	0.25	17	-0.40	0.26	-0.37	0.25
3.2	14	-0.40	0.25	-0.35	0.25	13	-0.40	0.26	-0.35	0.24	16	-0.40	0.27	-0.36	0.25
3.3	14	-0.40	0.25	-0.35	0.25	14	-0.40	0.26	-0.33	0.24	12	-0.40	0.26	-0.33	0.24
3.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4.1	19	-0.30	0.22	-0.31	0.21	20	-0.30	0.22	-0.31	0.21	21	-0.30	0.23	-0.30	0.22
4.2	17	-0.30	0.23	-0.32	0.21	20	-0.30	0.23	-0.31	0.22	24	-0.30	0.23	-0.31	0.22
4.3	17	-0.30	0.23	-0.32	0.19	17	-0.30	0.23	-0.31	0.20	17	-0.30	0.22	-0.31	0.19
4.4	17	-0.30	0.23	-0.32	0.19	17	-0.30	0.22	-0.33	0.18	17	-0.30	0.20	-0.33	0.17

Table 4.6: Mean point estimates of the treatment effect and 95% interval widths for each scenario when using the minimum number of clusters per arm required to achieve 80% power when using the NPP analysis method (from Table 4.5) (continued).

	Fixed Sampling Priors					Partial Sampling Priors					Full Sampling Priors				
Scenario		Frequentist		NPP			Frequentist		NPP			Frequentist		NPP	
	k	PE[a]	IW[b]	PE[a]	IW[b]	k	PE[a]	IW[b]	PE[a]	IW[b]	k	PE[a]	IW[b]	PE[a]	IW[b]
5.1	20	-0.30	0.21	-0.29	0.20	22	-0.30	0.21	-0.29	0.20	25	-0.30	0.20	-0.29	0.20
5.2	18	-0.30	0.22	-0.30	0.20	17	-0.30	0.19	-0.29	0.18	17	-0.30	0.19	-0.29	0.18
5.3	18	-0.30	0.22	-0.29	0.20	18	-0.30	0.23	-0.28	0.20	17	-0.30	0.22	-0.28	0.19
5.4	17	-0.30	0.23	-0.28	0.20	17	-0.30	0.22	-0.28	0.19	17	-0.30	0.23	-0.28	0.19
6.1	22	-0.30	0.20	-0.28	0.20	23	-0.30	0.21	-0.28	0.20	26	-0.30	0.20	-0.29	0.20
6.2	22	-0.30	0.20	-0.27	0.20	21	-0.30	0.20	-0.28	0.19	27	-0.30	0.21	-0.28	0.20
6.3	23	-0.30	-020	-0.27	0.19	25	-0.30	0.20	-0.26	0.18	21	-0.30	0.19	-0.26	0.18
6.4	27	-0.30	0.18	-0.25	0.18	29	-0.30	0.19	-0.25	0.18	-	-	-	-	-
7.1	43	-0.20	0.15	-0.21	0.14	47	-0.20	0.15	-0.21	0.14	49	-0.20	0.15	-0.21	0.15
7.2	37	-0.20	0.16	-0.21	0.15	47	-0.20	0.15	-0.21	0.15	54	-0.20	0.15	-0.21	0.15
7.3	35	-0.20	0.16	-0.22	0.15	35	-0.20	0.16	-0.22	0.15	35	-0.20	0.15	-0.22	0.14
7.4	35	-0.20	0.16	-0.23	0.15	35	-0.20	0.15	-0.23	0.14	35	-0.20	0.14	-0.23	0.13
8.1	46	-0.20	0.14	-0.20	0.14	50	-0.20	0.14	-0.20	0.14	52	-0.20	0.14	-0.20	0.14
8.2	42	-0.20	0.15	-0.20	0.14	35	-0.20	0.14	-0.20	0.13	35	-0.20	0.13	-0.20	0.13
8.3	40	-0.20	0.15	-0.20	0.14	44	-0.20	0.15	-0.20	0.14	35	-0.20	0.15	-0.20	0.14
8.4	37	-0.20	0.16	-0.20	0.14	35	-0.20	0.15	-0.20	0.14	36	-0.20	0.15	-0.20	0.14

Table 4.6: Mean point estimates of the treatment effect and 95% interval widths for each scenario when using the minimum number of clusters per arm required to achieve 80% power when using the NPP analysis method (from Table 4.5) (continued).

	Fixed Sampling Priors					Partial Sampling Priors					Full Sampling Priors				
Scenario		Frequentist		NPP			Frequentist		NPP			Frequentist		NPP	
	k	PE[a]	IW[b]	PE[a]	IW[b]	k	PE[a]	IW[b]	PE[a]	IW[b]	k	PE[a]	IW[b]	PE[a]	IW[b]
9.1	48	-0.20	0.14	-0.19	0.14	52	-0.20	0.14	-0.19	0.14	56	-0.20	0.14	-0.20	0.14
9.2	48	-0.20	0.14	-0.19	0.13	45	-0.20	0.14	-0.19	0.14	58	-0.20	0.14	-0.19	0.14
9.3	48	-0.20	0.14	-0.19	0.13	52	-0.20	0.14	-0.19	0.13	41	-0.20	0.14	-0.19	0.13
9.4	56	-0.20	0.13	-0.18	0.12	62	-0.20	0.13	-0.18	0.12	63	-0.20	0.13	-0.18	0.13

[a] Point Estimate
[b] Interval Width

Required Number of Clusters and Precision

In the context of the number of clusters required to achieve 80% power, clear patterns emerged in the results shown in Table 4.5. It can be seen that consistent reductions in the required number of clusters were observed in scenarios in which the treatment effect parameter used to generate the pilot data (θ_0) was the same, or greater, in magnitude than the target difference specified in the design of the definitive trial - specifically scenarios 1.1 - 1.4, 4.1 - 4.4, 7.1 - 7.4, 8.1 - 8.4. Larger differences in the magnitude of treatment effect estimated from the pilot data relative to the target difference resulted in larger reductions in the required number of clusters. In scenarios 1.1 - 1.4 and 8.1 - 8.4, where $\theta = \theta_0$ (i.e. the target difference is the same as the treatment effect in the pilot data), this reduction in the required number of clusters was driven by an increased precision in the estimate of the treatment effect parameter, as illustrated in Table 4.6. In scenarios 4.1 - 4.4 and 7.1 - 7.4, this reduction was driven by both an increase in precision and an increase in the magnitude of the point estimate of the treatment effect. Furthermore, the results indicated that the incorporation of pilot data through the NPP method can even reduce the number of clusters required when the treatment effect observed from the pilot data was *smaller* in magnitude than the target effect size (scenarios 2.1 - 2.4, 5.1 - 5.4), although to a lesser extent than when the pilot effect size is the same or larger than the target effect size. This suggested that under such scenarios, the reduction in power as a result of shifting the point estimate towards zero was more than offset by the increased precision in the treatment effect estimate. For scenarios in which the pilot data were generated with a null treatment effect, incorporation of this data through the NPP consistently resulted in an increase in the required number of clusters compared to the frequentist approach, as can be seen in scenarios 3.1 - 3.4, 6.1 - 6.4 and 9.1 - 9.4. However, it is important to consider that, whilst pilot or feasibility studies are not designed to assess treatment effect, it is perhaps less likely that definitive studies are undertaken without any signal of potential effectiveness. That is to say, if the point estimate from the pilot study is at or close to the null, it may be somewhat more difficult to justify progressing to a definitive trial.

Increasing the size of the pilot dataset amplified the observed effects on required study size, point estimate and interval precision of incorporating said dataset through the NPP analysis approach, be those effects positive or negative. For example, scenario 1.1 included just two clusters per arm in the pilot data, and the NPP method when using fixed sampling priors resulted in a reduction in the number of clusters required in the definitive trial from 12 to 10 per arm, and a narrowing of the 95% CrI width for the treatment effect from 0.30 to 0.28. In scenario 1.4, where 12 clusters per arm were again required under the frequentist analysis, the required number of clusters was reduced further to six per arm, and the interval width reduced from 0.38 to 0.27

when using the NPP method. In scenario 9.1, where the use of the NPP method had a detrimental effect on statistical power, the size of the pilot data again amplified these effects. The incorporation of the pilot data with $k_0 = 2$ resulted in an increase of three required clusters per arm compared with the frequentist approach, and a small reduction in the point estimate from -0.20 to -0.19. By increasing the size of the pilot study to $k_0 = 8$ as in scenario 9.4, an additional nine clusters per arm were required to reach the desired level of statistical power, and the point estimate for the treatment effect was reduced further towards the null to -0.18.

The effect of incorporating the pilot data through the NPP was broadly consistent across the three strategies for defining sampling priors. That is to say, if a reduction in the required number of clusters was observed, it tended to be observed across all three sampling-prior strategies, and similarly if an increase was observed this tended to be consistent across all three sampling strategies. However, in comparing the three sampling-prior strategies, there was a loss of statistical efficiency when increasing the amount of uncertainty in the data generating mechanism for the definitive data when moving from fixed to partial to full sampling priors, both when applying the frequentist and the NPP analysis approaches. There were no noteworthy differences in the point estimates for the treatment effect or interval widths across the three sampling-prior strategies, which is not surprising given the summaries pertain to different study sizes depending on the number of clusters required to achieve 80% power.

Type I Error

The incorporation of the pilot data through the NPP analysis strategy resulted in an inflation of the one-sided type I error above the nominal rate of 2.5% in scenarios with a non-zero treatment effect used to generate the pilot data. This inflation in type I error was exacerbated by increasing the magnitude of the treatment effect in the pilot data, and became as large as 16.6% in scenario 1.4 when fixed sampling priors were used to simulate the trial data. This is not surprising, given larger treatment effects in the pilot data express more substantial support *against* the null hypothesis.

In a similar fashion to the impact on statistical power, larger pilot study datasets also amplified the effects of incorporating the pilot data through the NPP analysis on type I error. For example, the type I error rate of 6.3% when using fixed sampling priors in scenario 1.1, where $k_0 = 2$, was inflated to 16.6% in scenario 1.4, where $k_0 = 8$. Once again this is an intuitive result, where larger pilot studies express stronger support *against* the null hypothesis.

When applying the NPP analysis method, there were some differences in the type I error rates between the three strategies for defining sampling priors. By moving from fixed to partial to full sampling priors, in order to better capture the uncertainty of the study design parameters, reductions in type I error rates tended to be observed. However, this pattern was not consistent across all scenarios, and so it is not possible to conclude definitively whether the choice of sampling-prior strategy has an impact, positive or negative, on control of type I error. What is clear, however, is that the use of an NPP analysis results in significant inflation of type I error rates across all three sampling-prior strategies.

***

4.5 Discussion

In this chapter, evidence of the potential for the use of the NPP to facilitate more efficient CRCT design has been demonstrated in certain scenarios through an increase in statistical power and therefore a reduction in the required sample size by using information borrowing techniques. However, it was also shown that there was a trade off in the application of such an approach; specifically, whilst statistical power can be enhanced and therefore the required number of clusters reduced, an inflation of type I error above the nominal rate can also be expected under most scenarios.

In the first instance, data from the HeLP pilot study was used in order to undertake a hypothetical redesign of the definitive study. The simulation-based power calculations demonstrated that by incorporating the pilot data into the primary analysis of the primary outcome from the main HeLP trial, reductions in the number of clusters required to achieve 90% power were achieved in comparison to both frequentist hierarchical models, and Bayesian hierarchical models with non-informative prior distributions. This conclusion was observed across all four sampling-prior strategies for specification of sampling priors.

A “signal” of potential effectiveness was observed in the analysis of the pilot HeLP data, where the point estimate of the treatment effect, although estimated with a high degree of uncertainty, exceeded the minimum clinically important difference. As a result, it is an unsurprising and intuitive result that incorporation of this pilot data enhanced statistical power in comparison to traditional methods which would analyse the definitive trial data alone. By using the NPP analysis approach to construct an informative prior based on the pilot data, the point estimate of the treatment effect can be expected to shift away from the null at zero (as in Table 3.1), thus increasing the chance in any given iteration of the simulation-based power calculation of rejecting the null hypothesis. In addition, the extra information incorporated into the analysis as a result of utilising the NPP method can be expected to enhance the precision of the treatment effect estimate, therefore further compounding the efficiency gains.

However, by the very same logic, through which an increase in statistical power can be expected, an inflation of type I error rate can also be expected. By shifting the point estimate away from the null, there is a higher chance of rejecting the null hypothesis for each simulated dataset, conditional on a treatment effect of zero in the current data. This gets increasingly likely both as the treatment effect under the pilot data deviates further from the null, and as the strength of historical evidence incorporated through the prior distribution increases with its sample size. In the simulation study based on the HeLP study, there was no obvious inflation in two-sided type I error. This was likely the result of the inclusion of evidence from the pilot data shifting the estimated treatment effect below zero, with the inflated error rate below (negative values) the null being negated by the deflated error rate above (positive values) it. As a result, the two-sided type I error rate is an inappropriate measure to consider in the context of information borrowing, and a one-sided type I error rate should be used instead. Inflated one-sided type I error rates were, as expected, observed as a result of using the NPP method in the analysis of the HeLP trial data.

The exploration of the four sampling-prior strategies observed that specifying a sampling prior for the SD of the primary outcome in the HeLP study did not result in a significant decrease in power. As such, this may be an elegant means of formally incorporating the uncertainty in estimate of this value during study design, regardless of whether the planned analysis strategy is Bayesian or frequentist. Furthermore, the results demonstrated that specifying a sampling prior for the ICC can be detrimental to the efficiency of the design, particularly if based only on data from a single pilot study. This is as a result of the imprecision in attempting to estimate an ICC using a single study with a small sample size. As a result, this approach is not recommended, which concurs with the conclusions of Eldridge et al. [Eldridge et al., 2015] who recommend drawing from multiple sources when justifying ICC assumptions in CRCT design. Specifying a sampling prior for the ICC using the meta-analytic methods proposed by Turner et al. [Turner et al., 2004] may provide a compromise between the efficiency of the design and the desire to formally account for the uncertainty in the estimated ICC. However, the number of relevant studies and the size of each study will play a significant role in the degree of certainty with which the ICC can be estimated, and with few studies and/or small studies, the results could be similar to simply using the pilot data.

The subsequent simulation study further reiterated the findings from the redesign of the HeLP study. Incorporation of external data, such as pilot data, through the NPP analysis method has the potential to facilitate more efficient study design under certain scenarios. In situations where the point estimate from the pilot data exceeds the target difference in the study under design, it can be expected that statistical power will be enhanced, predominantly as a result of shifting the treatment effect estimate away from zero, although an increase in precision can also be expected, resulting in further efficiency gains. In scenarios where the treatment effect is the same across the two datasets, an increase in study efficiency can again be expected, driven by an increased precision around the treatment effect estimate. However, perhaps the most interesting result is that the NPP methodology has the potential to enhance study efficiency even in scenarios where the treatment effect for the pilot data is closer to the null than the target difference. That is, despite the fact that using the NPP method shifts the point estimate closer to the null, the overall statistical power is still enhanced as a result of the increased precision with which the treatment effect is estimated. It is likely that as the treatment effect for the pilot data approaches the null the efficiency gain will diminish and eventually reverse and result in the need for a larger study size to overwhelm the evidence from the pilot data, as indicated by those scenarios in which the pilot data was simulated using a null treatment effect.

The simulation study also reiterated the findings related to inflated type I error rates observed in the HeLP redesign. In scenarios where the null hypothesis was supported in the pilot dataset, type I error was well controlled. However, incorporation of data with a non-null treatment effect inflated the type I error proportionate to the extent of the strength of evidence against the null. That is, when larger treatment effects were observed in the pilot data, and when the pilot studes were larger, more substantial inflation of type I error was observed.

The exploration of the three sampling-prior strategies presented in the simulation study indicated that incorporating the uncertainty in the design parameters tended to reduce statistical power. However, this was not observed in all scenarios, and in some, especially where the size of the pilot dataset was large, gains in power were observed, likely because of the ability to more accurately estimate the design parameters from the larger pilot datasets. The potential benefits of incorporating uncertainty in design parameters in terms of conservatism of assumptions should be weighed up against the potential loss of statistical power and inflation of required sample size, particularly when the pilot study is small. The key thing to note, however, is that regardless of the sampling-prior strategy, the impact on statistical power of adopting an NPP analysis approach remained consistent.

To summarise, in many scenarios, implementation of the NPP analysis method can justify smaller CRCT study designs through enhanced statistical power, but can result in the need for larger studies if the pilot data contradicts the alternative hypothesis. In order to ensure that a decision to use NPP methods in study design is not driven by the extent to which the sample size will be reduced, it should be recommended that the intention to use these methods is pre-specified before analysis of the pilot or feasibility study data. The trade-off in the use of the methods outlined in this chapter is the inflated type I error rate. A possible solution to this is to adopt a Bayesian interpretation of statistical power and type I error rate by specifying a fully Bayesian sampling prior for the target difference (i.e. not a point mass prior). Such an approach has been considered in the context of individually randomised trials, where the discounting parameter for the NPP is fixed at a value that controls the Bayesian type I error rate at the nominal level [Psioda and Ibrahim, 2019]. This approach is explored in the context of CRCTs in Chapter 5.

Chapter 5

Fixed Discounting Power Priors in Cluster Randomised Controlled Trials

In previous chapters, the focus has been on the use of the Normalised Power Prior to facilitate information borrowing. Instead, in this chapter, two variations of the Fixed Discounting Power Prior are explored. Existing methodology on choosing fixed discounting factor based on trial operating characteristics is extended to Cluster Randomised Controlled Trials, and a hypothetical re-design of the Healthy Lifestyles Programme trial is undertaken to explore how these new methods may impact study design and sample size calculation.

***

5.1 Introduction

IN the previous chapter, a novel use of the NPP, to incorporate historical data such as pilot data, was explored through both an applied example using the HeLP study data and through a simulation study, in the context of CRCT design. As a reminder, the NPP treats the discounting factor, a_0, as a parameter rather than a fixed value, which is estimated as part of the MCMC procedure, and reflects the degree of similarity between the historical and the current datasets. The findings, both from the application to the HeLP study and the subsequent simulation study, demonstrated the potential of these methods to enhance statistical power and therefore facilitate more efficient study design through justification of reduced sample sizes compared with the typical frequentist, formula-based approach to sample size determination in CRCTs. However, these findings also identified a trade-off against this efficiency gain; specifically, type I error rates can be expected to inflate as a result of incorporating external evidence through the NPP, unless the treatment effect within this external evidence is equal to the null hypothesis value (usually zero). Larger deviations from the null hypothesis result in larger inflation of type I error. In the context of trial design, it is likely that evidence from pilot and feasibility studies have already shown a "signal" of potential effectiveness, and that this signal has formed part of the justification for progression to a definitive trial. As a result, it is likely that in most practical applications of information borrowing in trial design and analysis, there will be incorporation of evidence that is at least mildly in support of the alternative hypothesis.

Intuitively, it is unsurprising that the specification of an informative prior distribution results in inflated type I error rates, given that informative prior distributions often provide evidence *against* the null hypothesis. Furthermore, type I error is itself a frequentist concept. This chapter, therefore, explores whether a Bayesian interpretation of type I error can help to justify the use of informative prior distributions, namely the power prior, in CRCT study design. Specifically, in the first instance, the operating characteristics of the HeLP study design presented in Table 4.2 are extended to include Bayesian type I error rates. Secondly, a method recently proposed by Psioda and Ibrahim [Psioda and Ibrahim, 2019] is extended in the context of CRCT design with application to the HeLP study. In particular, Psioda and Ibrahim proposed a power prior approach with a fixed discounting factor, a_0, where the value of the discounting factor is maximised within the constraints of controlling Bayesian type I error. In other words, instead of treating a_0 as a parameter as is required with the NPP method, this new approach aims to borrow as much evidence from the historical data as possible through manual specification of fixed a_0, without inflating the Bayesian type I error above some pre-specified nominal level. As a result, in order to explore this approach it is necessary to utilise power priors that allow specification of fixed values of a_0, and such methods are outlined in §5.2.1 below.

***

5.2 Methods

5.2.1 Fixed Discounting Power Priors

Recall from Equation (1.15) in Chapter 1 that the FDPP is a power prior formulation in which the discounting parameter, a_0, is fixed (rather than treated as a parameter) and chosen to reflect the degree of evidence borrowed from the historical data. When $a_0 = 0$, all historical evidence is discounted, and when $a_0 = 1$ all historical evidence is incorporated. The FDPP can be written as $\pi(\theta|D_0, a_0) \propto L(\theta|D_0)^{a_0}\pi_0(\theta)$, where θ and D_0 denote the model parameters and the historical data, respectively.

In CRCTs, analysis often proceeds using a hierarchical modelling strategy, with random effects for each cluster or randomisation unit. In Chapter 3, a NPP was proposed to facilitate information borrowing under this hierarchical modelling framework where a_0 is treated as a parameter to be estimated, and reflects the degree of commensurability between the current and historical datasets. A similar methodology can be applied in order to implement the FDPP within the same hierarchical modelling framework, but without the need to specify a prior distribution for a_0. In comparison to the NPP, the FDPP is a less complex model, in particular because the normalising constant no longer depends on a_0 and so explicit calculation is not necessary. Specifically, using the notation introduced in §3.4, an FDPP for a hierarchical model with cluster-level random effects can be written as

$$\begin{aligned}&\pi(\theta,\beta,\sigma^2,\sigma_c^2,\mathbf{b}_0|D_0,a_0)\propto\\&\quad\prod_{\tilde{i}=1}^{m_0}\prod_{\tilde{j}=1}^{n_{0\tilde{i}}}\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y_{0\tilde{i},\tilde{j}}-\mathbf{X_0}\beta-\theta z_{0\tilde{i},\tilde{j}}-b_{0\tilde{i}})^2\right]\right)^{a_0}\pi(\mathbf{b}_0|\sigma_c^2)\\&\quad\times\pi(\beta)\pi(\theta)\pi(\sigma^2)\pi(\sigma_c^2)\end{aligned} \tag{5.1}$$

All formulations of the power prior discussed within this thesis so far have involved borrowing information from *all* parameters contained within the likelihood for the historical data. However, it is possible to focus only on certain parameters, in a formulation known as a partial borrowing power prior (PBPP). For example, Psioda and Ibrahim composed a PBPP to borrow information on the treatment effect parameter only [Psioda and Ibrahim, 2019]. Denoting parameters of interest shared across both the current and historical data by ψ, and nuisance parameters not of interest in the historical and current data by ξ_0 and ξ, respectively, a generic PBPP with fixed discounting parameter can be written as

$$\pi(\psi,\xi,\xi_0|D_0,a_0)\propto L(\psi,\xi_0|D_0)^{a_0}\pi(\psi)\pi(\xi),\pi(\xi_0) \tag{5.2}$$

and therefore the posterior distribution is

$$\pi(\psi,\xi,\xi_0|D,D_0,a_0)\propto L(\psi,\xi|D)L(\psi,\xi_0|D_0)^{a_0}\pi(\psi)\pi(\xi),\pi(\xi_0) \tag{5.3}$$

In this chapter, in order to mirror and extend the approach adopted by Psioda and Ibrahim [Psioda and Ibrahim, 2019] (who only considered borrowing from the treatment effect parameter), a PBPP with fixed a_0 was constructed to facilitate borrowing of information from the treatment effect parameter, θ, in the context of a hierarchical model as used in the analysis of CRCT data. By zero-subscripting parameters featuring only in the likelihood of the historical data, such a power prior can be expressed as

$$\begin{aligned}&\pi(\theta,\beta,\sigma^2,\sigma_c^2,\beta_0,\sigma_0^2,\sigma_{c0}^2,\mathbf{b}_0|D_0,a_0)\propto\\&\quad\prod_{\tilde{i}=1}^{m_0}\prod_{\tilde{j}=1}^{n_{0\tilde{i}}}\left(\frac{1}{\sigma_0\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma_0^2}(y_{0\tilde{i},\tilde{j}}-\mathbf{X_0}\beta_0-\theta z_{0\tilde{i},\tilde{j}}-b_{0\tilde{i}})^2\right]\right)^{a_0}\pi(\mathbf{b}_0|\sigma_{c0}^2)\\&\quad\times\pi(\theta)\pi(\beta)\pi(\sigma^2)\pi(\sigma_c^2)\pi(\beta_0)\pi(\sigma_0^2)\pi(\sigma_{c0}^2)\end{aligned}\tag{5.4}$$

and so the full posterior distribution of all parameters is

$$\begin{aligned}&\pi(\theta,\beta_0,\sigma_0^2,\sigma_{c0}^2,\mathbf{b}_0,\beta,\sigma^2,\sigma_c^2,\mathbf{b}|D,D_0,a_0)\propto\\&\quad L(\theta,\beta,\sigma^2,\sigma_c^2|D)\pi(\mathbf{b}|\sigma_c^2)L(\theta,\beta_0,\sigma_0^2,\sigma_{c0}^2|D_0)^{a_0}\pi(\mathbf{b_0}|\sigma_{c0}^2)\\&\quad\times\pi(\theta)\pi(\beta)\pi(\sigma^2)\pi(\sigma_c^2)\pi(\beta_0)\pi(\sigma_0^2)\pi(\sigma_{c0}^2)\end{aligned}\tag{5.5}$$

with the posterior distribution of the treatment effect, θ, being of primary interest, and reflecting the culmination of evidence from both the historical and the current trial data.

Within this chapter, interest lies in examining the study design operating characteristics associated with the FDPPs with information borrowed from all parameters (denoted simply FDPP), and from only the treatment effect parameter (denoted PBPP), as outlined in Equations (5.1) and (5.4), respectively.

5.2.2 Null and Alternative Sampling Priors

In §4.2, the concept of a *sampling prior* was introduced (denoted $\pi^{(s)}(\cdot)$), where prior distributions are placed upon parameters involved in a power or sample size calculation, such as a SD when considering a continuous outcome, and the ICC in the context of CRCTs. However, in Chapter 4, the target effect size was specified as a fixed value (e.g. defined as the minimum clinically important difference). In contrast, in this chapter, the specification of a *sampling prior* for the treatment effect parameter is explored. Adopting such an approach inherently facilitates the interpretation of statistical power and type I error in a Bayesian manner. This is in contrast to the frequentist interpretation which seeks to control type I error at a pre-specified level (usually 5%, two-sided) under the assumption of a *fixed* value of the treatment effect consistent with a null hypothesis, and to achieve a level of statistical power at a *fixed* value of the treatment effect consistent with an alternative hypothesis. Put simply, Bayesian interpretations of these operating characteristics can be thought of as power or type I error *averaged* over the range of plausible values implied by the prior distributions of the design parameters. Bayesian power is often referred to in the literature as Bayesian assurance [O'Hagan and Stevens, 2001]. The concept of Bayesian power and type I error rates in the context of trial design is not new; Spiegelhalter and Freedman discussed methodology to specify a prior distribution for the alternative hypothesis in power calculations

[Spiegelhalter and Freedman, 1986], and Rubin and Stern similarly proposed using posterior predictive distributions in sample size calculations [Rubin and Stern, 1998]. More recently, Bayesian type I error has been considered alongside the use of power priors [Chen et al., 2014c].

Psioda and Ibrahim also recently considered Bayesian power and type I error alongside the use of power priors in clinical trial design [Psioda and Ibrahim, 2019], introducing the concept of null and alternative sampling priors, which express prior information in support of the null and alternative hypotheses, respectively. This is in contrast to previous work, which only specified point mass sampling priors for the null hypothesis [Ibrahim et al., 2012], which facilitates a balance between controlling type I error at a pre-specified nominal level, and allowing for information to be borrowed through methods such as the power prior. As demonstrated in Chapter 4, under the frequentist framework, or when using a point mass prior for the null hypothesis, information borrowing results in inflated type I error whenever the historical information is not in support of the null hypothesis. By specifying a null sampling prior that places weight at non-zero values (as opposed to placing all mass at zero), information borrowing can occur whilst also controlling Bayesian type I error at some nominal level.

In application to the HeLP trial, suppose interest lies in testing a one-sided hypothesis (which is more appropriate than a two-sided hypothesis when considering information borrowing methods, for the reasons outlined in §4.3.3), with null hypothesis $H_0 : \theta > 0$, and alternative hypothesis $H_1 : \theta \leq 0$. The posterior distribution of the treatment effect obtained through analysis of the HeLP pilot data is shown in Figure 5.1a. The null and alternative sampling priors for the treatment effect informed by the pilot data are calculated by truncating values of the posterior distribution above and below zero, shown in Figure 5.1b and Figure 5.1c, respectively. The former is referred to as the Default Null (DN) sampling prior, and reflects the full range of values of θ in support of H_0 (i.e. all values above zero) according to the posterior treatment effect obtained from analysis of the pilot data. Similarly, the latter is referred to as the Default Alternative (DA) sampling prior, and reflects the full range of values of θ in support of H_1 (i.e. all values below zero).

The DN and DA sampling priors illustrated in Figure 5.1 are intuitive in expressing support for null and alternative hypotheses based on existing evidence on a treatment effect. However, Psioda and Ibrahim also discussed the possibility of making modifications to these sampling priors, for example if expert opinion suggests that the tails are too heavy [Psioda and Ibrahim, 2019]. They proposed truncating the DN and DA sampling priors by constraining the values of θ to be at least $1/K$ times as likely as the modal value, with K some integer value.

In this chapter, a range of modifications to the DN and DA sampling priors are considered. For the null sampling prior, the DN (obtained using posterior samples obtained from analysis of the HeLP pilot study data) sampling prior was modified to include: (i) the lower tertile of the DN and (ii) the lower quintile of the DN, by truncating the DN sampling prior at the 33rd and the 20th percentiles, respectively (and the points of truncation are discussed). These null sampling priors are hereafter referred to as the tertile-truncated null (TTN) and the quintile-truncated null (QTN) sampling priors, and are illustrated in Figure 5.2b and Figure 5.2c, respectively, with the DN illustrated in Figure 5.2a. For the alternative sampling prior when considering the HeLP study, two modifications are proposed: (i) truncation of the DA at $\theta = -0.25$, which was the pre-specified target effect size in the original HeLP sample size calculation [Wyatt et al., 2013] (referred to as the truncated alternative (TA) sampling prior), and (ii) the 50% credible interval of the DA, which clips the DA at the 25th and 75th percentiles (referred to as the clipped alternative (CA) sampling prior). The TA and the CA sampling priors from the HeLP pilot data are illustrated in Figure 5.3b) and Figure 5.3c), respectively, with the DA shown in Figure 5.3a. For completeness, point mass sampling priors for the null and alternative sampling priors are also explored, in a similar way to a frequentist design. Specifically, for the null sampling prior, all mass is placed at zero, and for the alternative sampling prior, all mass is placed at the pre-specified target effect size of -0.25.

In Chapter 4, a range of sampling prior strategies were considered for the standard deviation of the outcome (σ), the ICC (ρ) and the intercept term (β). In this chapter, point mass priors will be placed on these parameters, as per Sampling Strategy I detailed in §4.3.1, in order to facilitate simplicity of interpretation and to ease the computational burden of the simulation-based power calculations.

5.2.3 Choosing the Discounting Factor

By specifying null and alternative sampling priors for the treatment effect, it is possible to use information borrowing methods whilst also controlling Bayesian type I error at some nominal level. However, it is still necessary to have an approach to determine the amount of evidence that can be borrowed from historical data whilst appropriately controlling the type I error. In the context of using a power prior to facilitate information borrowing, the process of determining the amount of evidence to borrow from the historical data manifests itself in the choice of discounting factor, a_0.

Psioda and Ibrahim [Psioda and Ibrahim, 2019] suggest an approach through which they seek to maximise the amount of information borrowed from the historical data. They do this using an iterative procedure that begins with $a_0 = 0$ (i.e. borrowing no historical evidence, or equivalently analysing the current data, $\mathbf{D}$, alone) and increases

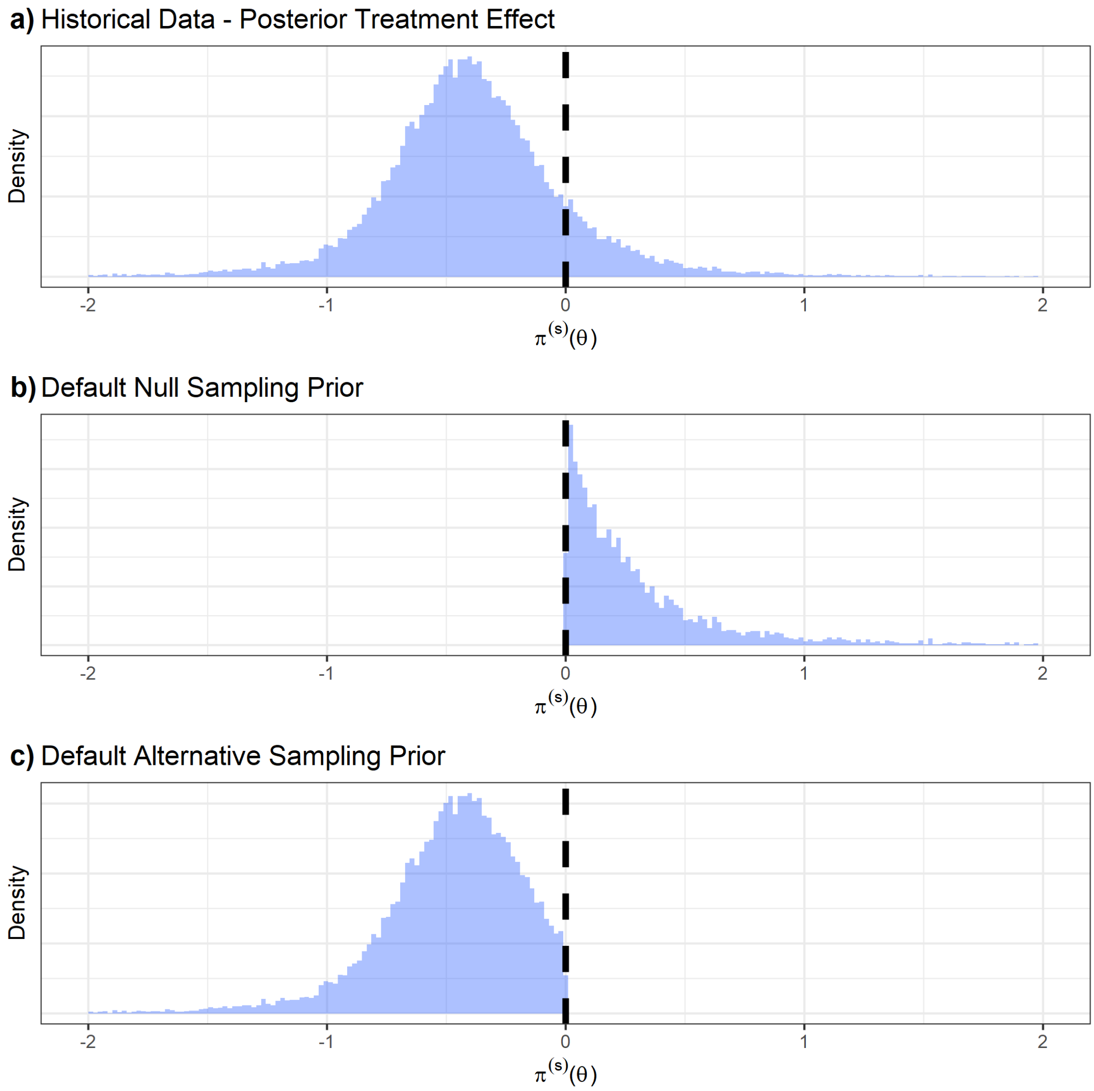
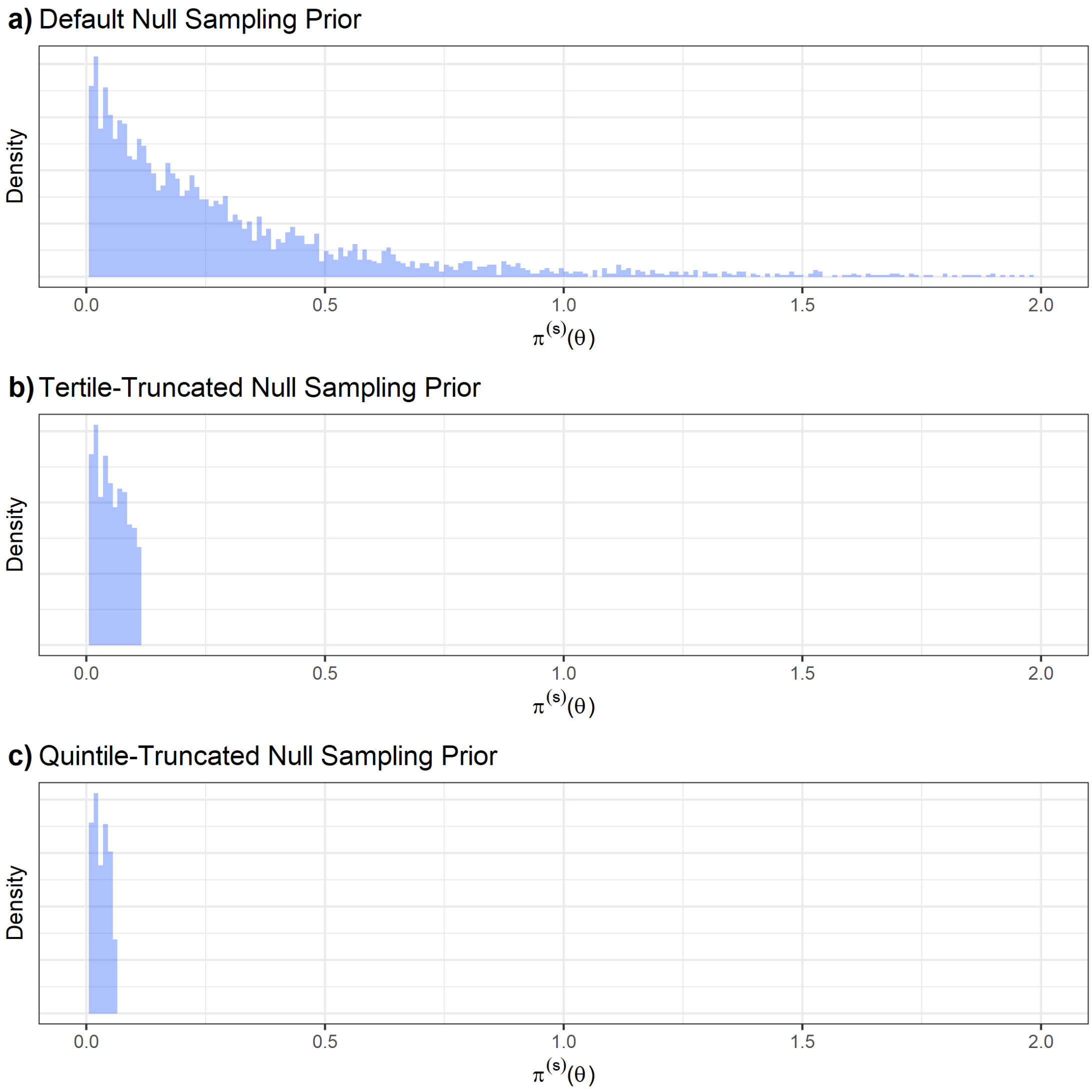


Figure 5.1: Posterior distribution of the treatment effect through analysis of the HeLP pilot data (a), and the associated null (b) and alternative (c) sampling priors.

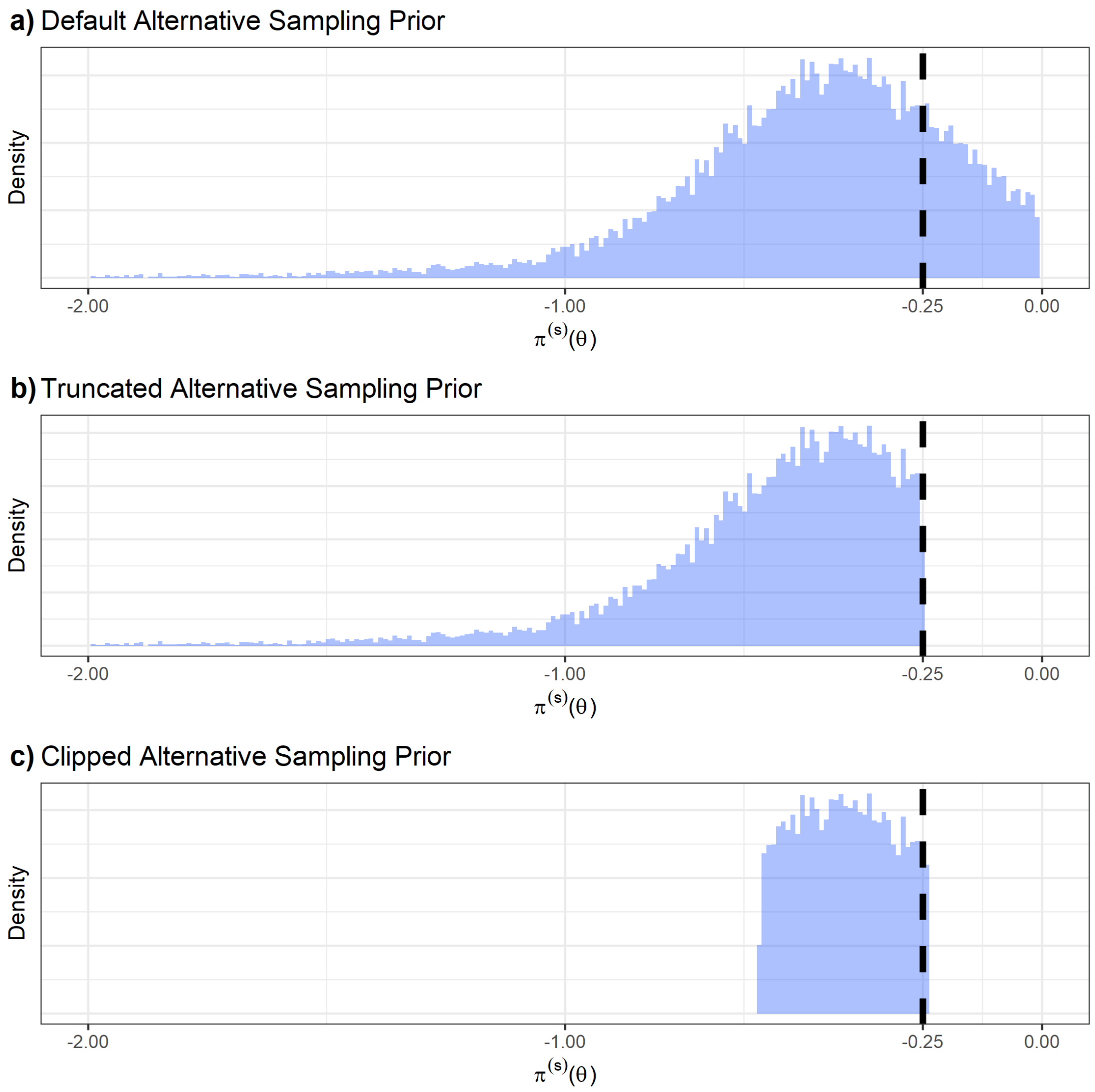


a) Default Null Sampling Prior

b) Tertile-Truncated Null Sampling Prior

c) Quintile-Truncated Null Sampling Prior

Figure 5.2: Distribution of the Default Null sampling prior (a), the Tertile-Truncated Null sampling prior (b) and the Quintile-Truncated Null sampling prior (c) from the HeLP pilot data.



a) Default Alternative Sampling Prior

b) Truncated Alternative Sampling Prior

c) Clipped Alternative Sampling Prior

Figure 5.3: Distribution of the Default Alternative sampling prior (a), the Truncated Alternative sampling prior (b) and the Clipped Alternative sampling prior (c) from the HeLP pilot data.

a_0 incrementally until the Bayesian type I error rate exceeds the pre-specified nominal level. The Bayesian probability of rejecting the null hypothesis without information borrowing (i.e. $P(\theta \leq 0|D, D_0, a_0 = 0)$) is equivalent to the frequentist p-value when $\theta = 0$. By acknowledging this fact, the approach becomes intuitive and can be thought of as beginning with a scenario which controls the frequentist type I error, and then maximising the amount of borrowed information (a_0) permissible before the Bayesian type I error is inflated above the nominal level. Once this value of a_0 has been determined, further simulations using this value can be undertaken to obtain an estimate of statistical power. Repeating this procedure for each proposed sample size (or number of clusters in the case of CRCTs), the smallest sample size which achieves both the desired level of statistical power whilst controlling the Bayesian type I error can be taken forward as the recruitment target.

This approach for determination of a_0 can be formalised as follows, for some given number of clusters per arm, k. Suppose that interest lies in testing a one-sided hypothesis for the treatment effect, θ, with support for the null hypothesis at values of $\theta > 0$, implying that a reduction in the outcome constitutes a favourable result. To initiate the algorithm, choose some small value of Δ (perhaps 0.05) to represent the incremental increases in a_0, and specify a total of N iterations. Furthermore, assume that historical data D_0 has already been obtained, and therefore interest lies in determining how much evidence from D_0 can be borrowed during the analysis of (yet to be collected) D without inflating the one-sided Bayesian type I error beyond some nominal level, α. Recall that $\pi^{(s)}(\cdot)$ and $\pi^{(f)}(\cdot)$ denote sampling and fitting priors, respectively, and let $\pi_N^{(s)}(\theta)$ and $\pi_A^{(s)}(\theta)$ denote the null and alternative sampling priors for the treatment effect. Let ξ represent the remaining parameters, and let μ_m and σ_m denote the expected mean and standard deviation of the cluster sizes. Then to determine a_0:

1. Generate N datasets, $\mathbf{D}_i, i = 1, \ldots, N$, comprising $2k$ clusters of size $m_j, j = 1, \ldots, 2k$ where $m_j \sim N(\mu_m, \sigma_m^2)$, using a sample from the sampling prior, $\pi^{(s)}(\xi)$ and the null sampling prior, $\pi_N^{(s)}(\theta)$ to generate each dataset.

2. For each value of a_0, where $a_0 = 0, \Delta, 2\Delta, \ldots, 1$, obtain posterior samples of the treatment effect, using the desired FDPP:

$$\pi(\theta, \xi|D_i, D_0, a_0) \propto L(\theta, \xi|D_i) L(\theta, \xi|D_0)^{a_0} \pi^{(f)}(\xi) \pi^{(f)}(\theta).$$

3. Using the sets of posterior samples, calculate the one-sided type I error, $P(\theta < 0|D, D_0, a_0)$ for each value of a_0, denoted $\hat{\alpha}_{a_0}$.

4. Using the values of $\hat{\alpha}_{a_0}$, determine the largest value of a_0 (constrained to be

no larger than 1) with which the type I error is controlled at some pre-specified nominal level, α.

This approach can be embedded within the wider problem of sample size determination, using a similar simulation-based method as outlined in §4.2. To begin, let $\hat{\Pi}_k$ denote the estimated statistical power simulated using k clusters per arm (assuming a two-arm trial), and let Π denote the (pre-specified) desired level of statistical power. To initiate the algorithm, choose some value of k (denoting number of clusters per arm) such that $\hat{\Pi}_{k-1} < \Pi$. That is, that the study is underpowered with $k-1$ clusters. Furthermore, specify a total of M iterations chosen to be large enough to ensure that power can be estimated with a suitable degree of precision, and let $h = 1, \ldots, M$. Then proceed according to the following steps:

1. Generate N datasets, $\mathbf{D}_i, i = 1, \ldots, N$, comprising $2k$ clusters of size $m_j, j = 1, \ldots, 2k$ where $m_j \sim N(\mu_m, \sigma_m^2)$, using a sample from the sampling prior, $\pi^{(s)}(\xi)$ and the null sampling prior, $\pi_N^{(s)}(\theta)$ to generate each dataset.

2. For each value of a_0, where $a_0 = 0, \Delta, 2\Delta, \ldots, 1$, obtain posterior samples of the treatment effect, using the desired FDPP:

$$\pi(\theta, \xi | D_i, D_0, a_0) \propto L(\theta, \xi | D_i) L(\theta, \xi | D_0)^{a_0} \pi^{(f)}(\xi) \pi^{(f)}(\theta).$$

3. Using the sets of posterior samples, calculate the type I error, $P(\theta < 0 | D, D_0, a_0)$ for each value of a_0, denoted $\hat{\alpha}_{a_0}$.

4. Using the values of $\hat{\alpha}_{a_0}$, determine the largest value of a_0 (constrained to be no larger than 1) with which the type I error is controlled at some pre-specified nominal level, α, and denote this value $\bar{a_0}$.

5. Simulate M datasets, $\tilde{\mathbf{D}}_h, h = 1, \ldots, M$, comprising k clusters of size $m_j, j = 1, \ldots, k$ where $m_j \sim N(\mu_m, \sigma_m^2)$, using a sample from sampling prior $\pi^{(s)}(\xi)$ and the *alternative* sampling prior, $\pi_A^{(s)}(\theta)$

6. Fit the analysis model with the FDPP to each $\tilde{D}_h$ to obtain samples from the posterior distribution of the treatment effect using $a_0 = \bar{a_0}$:

$$\pi(\theta, \xi | \tilde{D}_h, D_0, \bar{a_0}) \propto L(\theta, \xi | \tilde{D}_h) L(\theta, \xi | D_0)^{\bar{a_0}} \pi^{(f)}(\xi) \pi^{(f)}(\theta)$$

7. For each $\tilde{D}_h$, calculate $P(\theta < 0 | \tilde{D}_h, D_0, \bar{a_0})$

8. Calculate $\hat{\Pi}_k = \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}\{P(\theta < 0 | \tilde{D}_h, D_0, \bar{a_0}) \geq \alpha\}$

9. If $\hat{\Pi}_k < \Pi$, let $k = k + 1$ and return to step 1. Else, if $\hat{\Pi}_k \geq \Pi$, terminate the algorithm and declare the minimum study size required to achieve $\Pi\%$ power to be k clusters per arm.

The result of the implementation of this algorithm is the determination of the smallest number of clusters per arm required to achieve a desired level of Bayesian power, Π, whilst also maximising the amount of information borrowed from the historical data *and* controlling Bayesian type I error at some nominal level, α.

A worked example to elucidate the technicalities of this approach is provided in §5.3.2.

***

5.3 Results

5.3.1 Extending Previous Results - Bayesian Type I Error when Fitting a Normalised Power Prior

In Chapter 4, simulation-based calculations were undertaken to determine the minimum number of clusters required to achieve 80% or 90% power under each of sampling-prior strategies I - IV as outlined in §4.3.1, when analysis proceeds using an NPP. In addition, the one-sided frequentist type I error was presented. Here, these previously calculated results (i.e. when analysed using the NPP approach, as shown in Table 4.2) are extended to also report the one-sided Bayesian type I error under each sampling strategy when using the DN, the TTN and the QTN sampling priors, the results of which are shown in Table 5.1. Note that the frequentist type I error is calculated in the same way as the Bayesian type I error when using the Fixed Null (FN) sampling prior (and interpretation is similar), and so the results are identical. As can be seen in Table 5.1, Bayesian type I error is substantially smaller than frequentist type I error across all four sampling prior strategies. Whilst all non-point-mass sampling priors result in considerably smaller type I error rates compared to the FN, the smallest is observed when using the DN, and the largest when using the QTN.

5.3.2 Application to the HeLP Study

A Worked Example

Suppose that interest lies in determining the required number of clusters (k) per arm to achieve 80% Bayesian power whilst controlling one-sided Bayesian type I error at 2.5%. Furthermore, suppose that fixed sampling priors are specified for the intercept term, the ICC and the SD, so that $\pi^{(s)}(\beta) = 0.5$, $\pi^{(s)}(\rho) = 0.02$ and $\pi^{(s)}(\sigma) = 1.3$, and

Table 5.1: Estimated one-sided Bayesian type I error under sampling-prior strategies I - IV, when considering the null sampling priors outlined in §5.2.2. This table extends the results presented in Table 4.2

		80% Power		90% Power	
Sampling Strategy	Null Sampling Prior	k	One-sided Bayesian error	k	One-sided Bayesian error
I	FN[a]	19	4.9%	26	3.9%
I	DN	19	0.8%	26	0.7%
I	TTN	19	1.8%	26	1.3%
I	QTN	19	2.7%	26	1.7%
II	FN[a]	19	4.6%	26	4.0%
II	DN	19	0.6%	26	0.5%
II	TTN	19	1.1%	26	1.3%
II	QTN	19	1.8%	26	1.9%
III	FN[a]	32	3.8%	$>$42	-
III	DN	32	0.5%	$>$42	-
III	TTN	32	1.1%	$>$42	-
III	QTN	32	2.0%	$>$42	-
IV	FN[a]	24	3.7%	38	4.7%
IV	DN	24	0.9%	38	0.3%
IV	TTN	24	1.5%	38	0.9%
IV	QTN	24	1.9%	38	1.4%

[a] Bayesian and frequentist type I error under the FN sampling prior are equivalent.

that an average cluster size of 35 and a coefficient of variation in cluster size of 0.5 are assumed, in line with the assumptions made during the original HeLP power calculation [Wyatt et al., 2013]. Finally, suppose that the QTN and the DA sampling priors have been selected as fair representations of the null and alternative hypotheses, respectively, and that the FDPP will be used for the final analysis (i.e. that information borrowing will occur from *all* parameters, rather than just the treatment effect parameter).

Next, it is necessary to initialise the simulation-based power calculation by choosing some value of k (clusters per arm) such that the study will be underpowered. In this case, the algorithm will be initialised with $k = 10$. The next step is to determine the largest value of a_0 that does not result in inflation of the Bayesian type I error above 2.5%. In order to do so, Bayesian type I error is calculated, using simulated datasets according to the QTN sampling prior, for each value of a_0 between zero and one (inclusive) in increments of 0.05. The results of these calculations are shown in Figure 5.4, overlaid with a Locally Weighted Smoothing (LOESS) line.

As can be seen in Figure 5.4, the maximum permissable value of a_0 when using the QTN with $k = 10$ is 0.19; anything larger would result in inflation of the one-sided Bayesian type I error above the pre-specified 2.5%.

The next step in the process is to determine whether $k = 10$ clusters per arm, analysed with a FDPP using $a_0 = 0.19$ is sufficient to achieve 80% Bayesian power. Bayesian power can be estimated via simulation, with datasets simulated in this case according to the DA sampling prior. In this scenario, the Bayesian power was estimated to be 79.1%, below the required 80%, and as a result the algorithm is re-initialised at $k = 11$.

Further iterations of the algorithm estimated the maximum a_0 associated with $k = 11$ as 0.30, and Bayesian power as 81.2%. Given Bayesian power is now above the target power of 80%, the algorithm is terminated and the minimum number of clusters required is declared to be $k = 11$ per arm, which allows borrowing of the evidence from the historical data with $a_0 = 0.30$, whilst controlling the one-sided Bayesian type I error at 2.5%.

Determining a_0

As demonstrated above, the maximum permissable value of a_0 which allows for control of Bayesian Type I error varies according to the number of clusters per arm. For the HeLP study, this relationship is illustrated in Figure 5.5, and pertains to the TTN (Figure 5.5a) and the QTN (Figure 5.5b) for the FDPP and the PBPP. For the DN sampling prior, it was found that all evidence from the historical data could be incorporated (i.e. $a_0 = 1$) whilst controlling Bayesian type I error across all relevant values of k. Further-

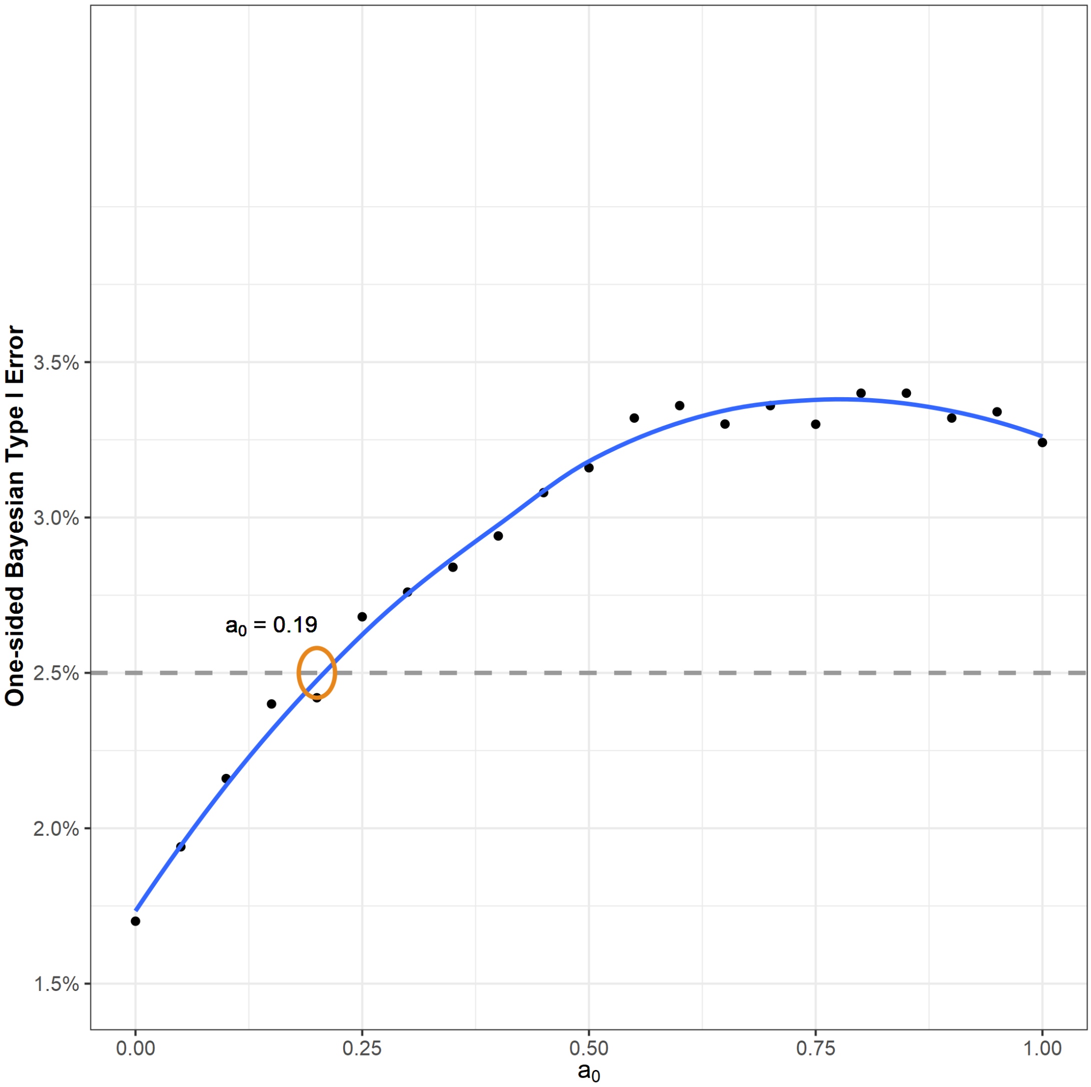


Figure 5.4: A scatterplot of a_0 against one-sided Bayesian type I error, overlaid with a LOESS line, calculated using data simulated with $k = 10$ clusters per arm, and a QTN sampling prior, and analysed using the FDPP.

more, in Chapter 4, it was established that when using a FN sampling prior (equivalent to frequentist type I error), it was not possible to borrow any historical evidence whilst controlling type I error. This result was verified again in this chapter for a selection of study sizes ($k_0 = 12, 17, 22$), chosen randomly simply for validation purposes (instead of validating for all values of k_0, in order to minimise unnecessary computational cost).

Across both Figure 5.5a) and Figure 5.5b), a clear relationship can be observed; as the number of clusters increases, so to does the amount of evidence that can be borrowed from the historical data without excessive inflation of the one-sided Bayesian type I error. Furthermore, in comparing the FDPP (in blue) with the PBPP (in grey), there is evidence that the former allows for a greater degree of information borrowing than the latter, particularly for smaller values of k. Finally, in comparing Figure 5.5a) with Figure 5.5b), it can be seen that the TTN sampling prior allows for a greater degree of information borrowing than the QTN sampling prior, although they both plateau at $a_0 = 1$ for larger values of k.

Sample Size and Power

The simulation-based sample size calculation methodology outlined in §5.2.3 was applied to each combination of null and alternative sampling priors (as per §5.2.2) in the context of the redesign of the HeLP study, with both FDPPs and PBPPs constructed using the HeLP pilot study data. For each combination of sampling priors and power prior formulation, the sample size requirements, alongside the associated maximal values of a_0, are shown in Table 5.2. A total of $5,000$ simulations was undertaken, for calculation of both the Bayesian one-sided type I error and the statistical power. Whilst the number of iterations was chosen to ensure manageable computation time, with $5,000$, a type I error rate of 2.5% can be estimated with precision of $\pm 0.4\%$, or power of 80% with precision of $\pm 1.1\%$.

As shown in Table 5.2, when using a FN sampling prior, it was not possible to allow any evidence to be borrowed from the pilot data without a resulting inflation in Bayesian type I error, consistent with the findings in Chapter 4. As a result, this means that using a FN sampling prior in the trial design will result in analysis of the definitive trial data alone. Using a Fixed Alternative (FA) sampling prior to calculate power, alongside a FN sampling prior, is similar to a frequentist approach to study design, with treatment effect fixed at zero under the null hypothesis and at the MCID of -0.25 under the alternative hypothesis. Implementing instead the DA sampling prior results in a substantial reduction in the number of clusters required to achieve 80% power. Using the TA or the CA sampling priors together with the FN results in further reductions in the number of clusters required, both to achieve 80% and 90% power.

The DN sampling prior allows for borrowing of *all* evidence from the historical data,

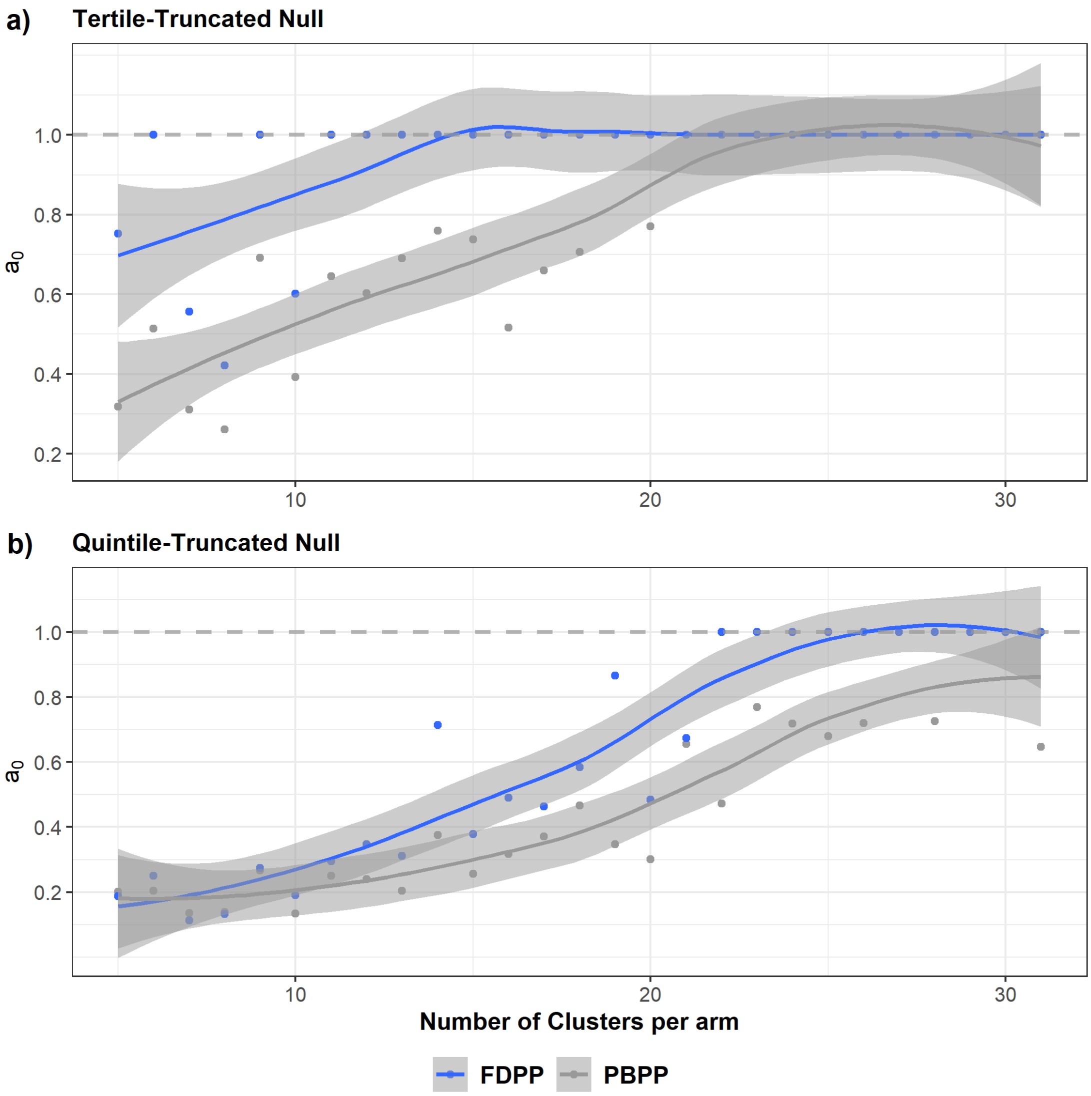


Figure 5.5: A scatterplot of the number of clusters per arm against the maximum value of a_0 which controls one-sided Bayesian type I error at 2.5%, for the TTN (a) and the QTN (b), overlaid with a LOESS line, in the context of the HeLP study. Shading represents 95% Confidence Intervals for the LOESS line.

regardless of choice of alternative sampling prior. This indicates that the evidence from the historical data incorporated through the power prior was insufficient to overwhelm the evidence within the DN to the extent that it resulted in inflation of one-sided type I error. In facilitating this information borrowing, the required number of clusters is reduced in comparison to using the FN sampling prior across all alternative sampling priors, for both 80% and 90% power.

When implementing the TTN and the QTN sampling priors, the amount of information borrowing varies according to the number of clusters and choice of alternative sampling prior. Across all alternative sampling priors, the amount of information that can be borrowed is larger under the TTN compared to the QTN. In addition, the FA allows the most borrowing, followed by the DA, then the TA then the CA. In terms of study size, the TTN and QTN require fewer clusters than the FN, but more than the DN, reflecting the fact that more information borrowing is facilitated compared to the FN, but not full borrowing as with the DN sampling prior.

In comparing the FDPP, which borrows information across all parameters, with the PBPP, which borrows information only from the treatment effect parameter, there are no differences in the required number of clusters when using the FN sampling prior. In fact, given $a_0 = 0$ when using the FN sampling prior, the power prior formulation is completely discounted and the two approaches become equivalent. Across the remaining three null sampling priors, similar patterns emerge in comparing required study sizes. Specifically, under the DA, TA and CA, fewer or the same number of) clusters are required when using an FDPP compared with a PBPP, although this relationship appears weakest under the CA. Conversely, when using the FA, fewer (or the same number of) clusters are required when using the PBPP at 90% power, although the same number are required at 80% power. In addition, the FDPP consistently facilitates a greater degree of information borrowing.

Table 5.2: The maximum value of a_0 that controls one-sided Bayesian type I error at 2.5% and the required number of clusters per arm (k) to achieve 80% and 90% power for each combination of null and alternative sampling priors, for each power prior Formulation.

			80% Power		90% Power	
Null Sampling Prior	Alternative Sampling Prior	Power Prior Formulation	a_0	k	a_0	k
FN	FA	FDPP	0	20	0	>32
FN	FA	PBPP	0	20	0	>32
FN	DA	FDPP	0	14	0	>32
FN	DA	PBPP	0	14	0	>32
FN	TA	FDPP	0	9	0	11
FN	TA	PBPP	0	9	0	11
FN	CA	FDPP	0	10	0	15
FN	CA	PBPP	0	10	0	15
DN	FA	FDPP	1	15	1	27
DN	FA	PBPP	1	15	1	26
DN	DA	FDPP	1	9	1	26
DN	DA	PBPP	1	12	-	>31
DN	TA	FDPP	1	5	1	7
DN	TA	PBPP	1	6	1	9
DN	CA	FDPP	1	6	1	11
DN	CA	PBPP	1	6	1	11
TTN	FA	FDPP	1	15	1	27
TTN	FA	PBPP	0.74	15	1	24
TTN	DA	FDPP	1	9	1	26
TTN	DA	PBPP	0.39	10	-	>31
TTN	TA	FDPP	0.75	5	0.56	7
TTN	TA	PBPP	0.32	5	0.69	9
TTN	CA	FDPP	1	6	1	11
TTN	CA	PBPP	0.31	7	0.64	11
QTN	FA	FDPP	0.46	17	1	28
QTN	FA	PBPP	0.37	17	0.77	23
QTN	DA	FDPP	0.30	11	1	26
QTN	DA	PBPP	0.25	11	0.65	31
QTN	TA	FDPP	0.19	5	0.19	10
QTN	TA	PBPP	0.20	6	0.13	10
QTN	CA	FDPP	0.13	8	0.31	13
QTN	CA	PBPP	0.27	9	0.38	14

Abbreviations: FN denotes Fixed Null; DN denotes Default Null; TTN denotes Tertile-Truncated Null; QTN denotes Quintile-Truncated Null; FA denotes Fixed Alternative; DA denotes Default Alternative; TA denotes Truncated Alternative; CA denotes Clipped Alternative; FDPP denotes Fixed Discounting Power Prior; and PBPP denotes Partial Borrowing Power Prior.

***

5.4 Discussion

Earlier work within this thesis has demonstrated that borrowing information from historical data, such as a pilot or feasibility study, within the context of CRCTs has the potential to justify smaller sample sizes, both as a result of increased precision, and, in cases where the historical data provides evidence in favour of the alternative hypothesis, by shifting the point estimate of the treatment effect. However, in addition to this potential increase in statistical efficiency, inflation in one-sided type I error rates occurs whenever the historical evidence contradicts the null hypothesis. In this chapter, a recently proposed approach by Psioda and Ibrahim [Psioda and Ibrahim, 2019] has been extended to CRCTs and applied to a hypothetical redesign of the HeLP study. Specifically, by adopting a Bayesian interpretation of type I error and statistical power, it has been demonstrated that it is possible to both borrow historical information (thus reducing sample size requirements) and control the Bayesian one-sided type I error.

The simulation-based approach to sample size calculation outlined within this chapter can be thought of as a two-stage process, the first of which involves determining the maximum value of a_0 that controls Bayesian one-sided type I error. During this stage, the determined value of a_0 is dependent on the choice of null sampling prior. As previously outlined, when a FN sampling prior is used, the interpretation is similar to the frequentist approach, and as a result, no information borrowing is permitted without inflation of Bayesian one-sided type I error. When there is support under the null sampling prior for a wide range of values that differ substantially from zero, as in the DN sampling prior, more information borrowing is allowed; this is an intuitive result, as the support for more extreme values under the null is able to overwhelm the evidence from the historical data which would otherwise inflate the type I error. Conversely, the QTN allows the least information to be borrowed, as it is the most heavily truncated of the three (non-fixed) sampling priors considered, and therefore has little support for values that lie far from the null hypothesis.

The second stage of the simulation-based sample size calculation involves determining the smallest number of clusters required to achieve a desired level of statistical power. This stage of the process is dependent on both the value of a_0 determined in the first stage, and on the choice of DN sampling prior. Recall that the HeLP pilot data showed some evidence of a treatment effect (as detailed in §3.5). As a result, larger values of a_0, and thus more information borrowing, can be expected to lead to smaller required sample sizes/number of clusters for a definitive trial. This was confirmed in the results within this chapter with, for example, a decrease from $k = 20$ when $a_0 = 0$, to $k = 15$

when $a_0 = 1$ when using a FA sampling prior. Furthermore, alternative sampling priors with support for larger treatment effects (e.g. the TA sampling prior) again resulted in fewer clusters being required. Given larger treatment effects require fewer participants or randomisation units (clusters) to detect, this is once again an intuitive result.

There is evidence in the results of the simulations that borrowing information from *all* parameters using the FDPP, allowed for a greater degree of information borrowing (i.e. larger values of a_0) in comparison to borrowing information *only* from the treatment effect parameter using the PBPP. Recall that values of a_0 are smaller when there is *stronger* evidence from the historical data in contradiction to the null sampling prior. This indicates that the prior information on the treatment effect may be more precise, or of a greater magnitude, under the PBPP compared to the FDPP.

When using a FA sampling prior, the results suggested that more clusters were required to achieve a desired level statistical power when using the FDPP in comparison to the PBPP. Recall from §3.5.3 that, in the context of the analysis of the HeLP data, incorporation of the pilot data through the power prior resulted in larger estimated values of the ICC. As such, the increase in required number of clusters may be due to the increased ICC and therefore reduced precision in treatment effect estimation under the FDPP, which is less likely to be apparent in the PBPP which estimates the variance components from the definitive trial data alone. However, under the DA and the TA sampling priors, this trend is reversed, and in fact it is the FDPP which requires fewer clusters. Given the DA and TA both support large treatment effects, it may be that the increased amount of information borrowing facilitated through the FDPP in comparison to the PBPP drives this reduction.

Evidently, the adoption of a Bayesian interpretation of type I error and statistical power, coupled with information borrowing techniques, can result in increased statistical efficiency and therefore reduced study sizes (and costs). However, it is important to acknowledge that the methods presented in this chapter are not merely a means of justifying smaller CRCTs. Rather, these methods represent a principled approach that both maximises value from hard-won pilot data that has already been collected (both through information borrowing and expression of data-driven null and alternative sampling priors), and facilitates a more intuitive, probabilistic interpretation of power and type I error, metrics that require a somewhat obtuse interpretation in the frequentist framework.

Despite the potential benefits of adopting the methods outlined in this chapter (and elsewhere), there remains a considerable barrier to practical implementation. Namely, the two-stage simulation-based approach to determining both a_0 and k requires significant computational cost and time, and is likely impractical to undertake without access to a high performance computing cluster.

It could also be argued that the modifications made to the DN sampling prior in order to create the TTN and the QTN sampling priors are somewhat arbitrary. Evidently, without these modifications, the DN expresses support for perhaps unrealistically large values of the treatment effect, resulting in borrowing all historical information. As such, at least in some situations, this modification is likely appropriate, but attempts should be made to justify these, or other, modifications, either empirically or through elicitation of expert opinion. Similar issues may occur in choosing alternative sampling priors, although similar justification to the choice of target difference in a traditional frequentist trial design could be considered. These issues are likely exacerbated when the methods are applied to CRCTs in comparison to individually randomised trials due to the substantial uncertainty associated with estimating the treatment effect using a hierarchical model with a small number of clusters, as is typical with a pilot study. It is likely that, if the pilot study is individually randomised, less substantial (or perhaps no) modifications to the default sampling priors would be required. However, an individually randomised pilot study followed by a cluster randomised definitive trial is not common.

To summarise, the methods to determine the largest value of a_0 which controls Bayesian type I error, proposed by [Psioda and Ibrahim, 2019], have been extended to CRCTs and applied to the HeLP study. The findings show that by adopting a Bayesian interpretation of power and type I error, information borrowing from the HeLP pilot study can be undertaken without excessive inflation of type I error. The result is a reduction in the required number of clusters to achieve a desired level of power. In some scenarios, the FDPP facilitates more efficient study design, and in others the PBPP is preferable. It may be difficult to justify *not* borrowing information from all parameters if the methodology allows, and so the FDPP may be the preferable power prior formulation. The computational burden of designing a CRCT using these novel methods is high. Furthermore, elements of these methods may require expert input in practice, and further research may be required to ensure modifications to sampling priors are appropriate. This is likely of greater importance in CRCTs compared with individually randomised trials. Finally, caution must be exercised in extending the results presented within this chapter to other scenarios, given the methodology was applied only to one CRCT, rather than explored more thoroughly through a simulation study.

Chapter 6

PPCRCT: Power Priors in Cluster Randomised Controlled Trials - An R Package

Within this chapter, an R software package is introduced which facilitates straightforward implementation of the Bayesian analysis methodology proposed within this thesis. Specifically, the package allows a user to fit a normalised power prior, a fixed discounting power prior or a partial borrowing power prior to clustered data. The structure of the package, as well as details of its functionality, is outlined.

***

6.1 Introduction

EARLIER chapters within this thesis have emphasised the importance of the development of statistical software packages alongside the introduction of novel methodology to ease the challenge of practical implementation and ultimately make such modern, often complex, methods more accessible, thus improving take-up. This chapter introduces an R package, `PPCRCT`, which facilitates implementation of the novel methodology proposed and developed in this thesis to incorporate historical data into the analysis of CRCTs. Specifically, the R package fits NPPs, FDPPs and PBPPs to data with a continuous outcome and clustering in both the current and historical datasets. The package can be used to run such analyses in their own right, or can be embedded within a simulation study framework to explore the impact of adopting these methods on trial design characteristics such as statistical power or type I error. At the time of submission of this thesis, `PPCRCT` was available to download from GitHub (https://github.com/benjones13/PPCRCT), with an ambition to upload to the Comprehensive R Archive Network (CRAN) at a later date. The package can be installed within R by running a single line of code:

```
devtools::install_github("benjones13/PPCRCT").
```

***

6.2 The R Programming Language

R is a free to use, open-source, object-oriented statistical programming language [R Core Team, 2019]. It effectively and efficiently facilitates data manipulation, data visualisation and a vast array of statistical analysis methods. One of the key strengths of R is its large and active community of users and developers, who together provide an ongoing platform for support and troubleshooting as well as ensuring an ever-growing suite of packages are developed and available to implement the latest statistical methodology. In comparison to commercial software, such as Stata [StataCorp, 2021], this active community and open-source structure makes both troubleshooting and access to tools which can implement new methodology much easier.

A number of tools have been developed in recent years which make R package development more straightforward and accessible for applied statisticians, including: `devtools` which is an R package itself that contains a collection of package development tools; version control software such as Git and GitHub (`https://github.com`); and books such as “R Packages” by Hadley Wickham [Wickham, 2015].

***

6.3 Stan - a Probabilistic Programming Language

Stan is a state-of-the-art, open-source probabilistic programming language, and perhaps the gold-standard software used for Bayesian statistical modelling, inference and prediction [Carpenter et al., 2017]. Stan provides a powerful yet flexible and intuitive platform, and interfaces with most popular statistical software packages, including R, Python and Stata. It implements HMC methods for posterior sampling (see §1.3.2), and therefore offers significant advantages in computational efficiency compared to other Bayesian software packages such as BUGS and JAGS, both of which implement Gibbs sampling.

Stan offers a range of features which allow for easier implementation of Bayesian inference in practice. In R, two key packages have been developed to support straightforward implementation of Bayesian regression methods using Stan - `rstanarm` [Goodrich et al., 2020] and `brms` [Bürkner, 2018]. Both packages have similar functionality, adopt syntax consistent with standard R packages for regression modelling such as the `lme4` [Bates et al., 2015], and can facilitate hierarchical modelling. Users who wish to fit more complex, bespoke models, can do so by writing their own `.stan` files. Once written, these bespoke models can be compiled, fitted and the results examined and manipulated using the `rstan` package [Stan Development Team, 2022], which offers an interface between R and Stan. Models fitted using the suite of functions in `rstan` output objects of type `"stanfit"`, for which further visualisation and summarisation are also supported.

***

6.4 The PPCRCT R Package

The new `PPCRCT` package provides a user-friendly means to fit the various power prior models outlined in this thesis within R. Specifically, the package allows the user to fit the NPP, introduced in Chapter 3 in the context of CRCTs, the FDPP and the PBPP, both introduced in Chapter 5 in the context of CRCTs, to data with a continuous outcome and clustering present in both the historical and the current datasets. The `PPCRCT` package gives the user some flexibility over choice of prior distributions, including allowing specification of either a Half-Cauchy or a Half-Normal prior distribution for the between-cluster standard deviation parameter, in line with recommendations by Gelman [Gelman, 2006]. The package also offers an automated, data-driven approach to specifying priors in the event that the user does not specify their own. Further detail on the prior distributions is provided in §6.4.2. The package contains two front-end (i.e. for use by the end-user) functions: `NPP` allows the user to fit an NPP, and `FDPP` allows the user to fit either a FDPP or a PBPP, depending on the chosen function parameter inputs.

6.4.1 Data Validation

Both `NPP` and `FDPP` take the following as data input objects:

- Two matrices (`X0` and `X`) which contain the design matrix (excluding the column of 1s representing the intercept term) for the historical and the current data, respectively, where each row represents a participant, the first column is a column of 1s and 0s representing intervention and control arms, respectively, and each subsequent column represents a covariate.

- Two vectors (`Y0` and `Y`) which contain the continuous outcome data for the historical and current datasets, respectively.

- Two vectors (`Z0` and `Z`) which contain consecutive, numerical cluster labels for each participant in the historical and current datasets, respectively.

Within both front-end functions, automated checks have been programmed to ensure that the data passed to the function by the user is in the correct format. If any issues are identified, the function is terminated and an informative error message is returned.

Specifically, data checks are undertaken upon `X` and `X0` to confirm that: (i) each is a matrix; and (ii) that the first column of each contains only 1s and 0s, which reflects the treatment group indicator variable. Similarly, for `Y` and `Y0`, checks are undertaken to ensure that both are vectors. For `Z` and `Z0`, checks are undertaken to ensure that (i) both are vectors; (ii) all values are integers; and (iii) that the cluster labels are numerically consecutive. Finally, checks are undertaken to ensure that the number of rows of `X0` is the same as the number of elements of `Y0` and `Z0`, and similarly for `X`, `Y` and `Z`. Missing data is not compatible with `PPCRCT` and if any datasets containing missing data are passed in to either function, the function is terminated with a warning to the user to remove missing data before proceeding.

6.4.2 Prior Distributions

The `PPCRCT` package provides some flexibility for specification of prior distributions. For the regression parameters, including the intercept term, normal prior distributions are implemented. However, the user has the ability to specify the means and standard deviations (SDs) for each of these prior distributions. Similarly, the within-cluster SD parameter is allocated an exponential prior distribution, but the user is able to specify the value of the rate parameter for this distribution. For the between-cluster SD, two prior distributions are possible; a Half-Cauchy prior, or a Half-Normal prior. In addition to specifying which to implement, the user is able to specify the value of the scale parameter (if using a Half-Cauchy prior), or the SD parameter (if using a Half-Normal prior).

Furthermore, if the user does not specify values for the parameters of the prior distributions, `PPCRCT` calculates and specifies them automatically. Let S_{Y_0} denote the SD of `Y0`, and let $S_{X_0}^{(i)}$ denote the SD of the i^{th} column of `X0`. For the intercept term, a N$(0, 2.5S_{Y_0})$ prior distribution is fitted. For the i^{th} regression coefficient, a N$(0, 2.5S_{Y_0}/S_{X_0}^{(i)})$ prior distribution is fitted. For the within-cluster SD, an Exponential$(1/S_{Y_0})$ prior distribution is fitted. These choices represent weakly informative prior distributions which can help to stabilise computations, and are specified in line with the choices used in `rstanarm` [Goodrich et al., 2020], a well-established R package used to fit Bayesian regression models.

For the final parameter, the between-cluster SD, if the user fails to specify a choice between the Half-Cauchy or the Half-Normal prior distribution, a default position will be adopted. If the total number of clusters in the historical data is fewer than five, a Half-Cauchy prior will be specified, with location parameter equal to zero and scale parameter equal to half of the observed between-cluster SD for the outcome calculated from the historical data. This is in line with Gelman's recommendation to use an informative prior such as the Half-Cauchy when fewer than five clusters in total are present, using a scale parameter that is "high but not off the scale" [Gelman, 2006]. When the historical data contains five or more clusters, a Half-Normal prior distribution will be specified as default, with mean parameter equal to zero, and SD parameter equal to ten times the observed between-cluster SD. This represents a non-informative prior distribution, once again in line with Gelman's recommendation [Gelman, 2006].

6.4.3 Stan Models

The `PPCRCT` package contains a total of eight bespoke Stan models, which are called as required within the `NPP` and `FDPP` functions. Each of these models sits as a standalone `.stan` file within the package file structure. Of the bespoke models, two are used in approximation of the normalising constant using the methodology outlined in §3.3 (phases 3a and 3b, Figure 6.1); two are used in fitting the NPPs (phases 5a, 5b, Figure 6.1); two are used in fitting the PBPPs (phases 3a and 3b, Figure 6.2) and two are used in fitting the FDPPs (phases 3c and 3d, Figure 6.2).

6.4.4 **NPP**

The `NPP` function is one of two front-end functions contained within the `PPCRCT` package. A schematic illustrating the structure of the function is shown in Figure 6.1. The blue box indicates the front-end function designed for direct use by the end user. The grey boxes detail the different phases of the functions, and orange boxes denote back-end functions created to modularise the code contained within the front-end function, but not designed for direct use by the end user.

After passing the required information via the `NPP` function arguments, the initial data checking is undertaken, as detailed in §6.4.1, followed by the default specification of the prior distributions if not provided by the user, as detailed in §6.4.2.

Next, the data and the information regarding the prior distributions are passed to the back-end function `Ca0_fun`, which implements the methodology outlined in §3.3 to generate a fine grid of the values of the normalising constant ($C(a_0)$) and associated discounting parameter values (a_0). Separate models are programmed to approximate these values depending on whether a Half-Cauchy (phase 3a) or a Half-Normal (phase 3b) prior distribution has been chosen (either by the user or according to the default position) for the between-cluster SD, represented by two separate `.stan` files within the package file structure. After generating the grid of approximations, the result, along with the data, is passed to a second back-end function, `NPP_modelfit`, within which the NPP model (as in §3.4) is fitted. The NPP model is fitted using one of two `.stan` functions, depending on the choice of prior distribution for the between-cluster SD (phases 5a and 5b). Finally, the results of the NPP model are returned to the user as an object of class `"stanfit"` for inspection, visualisation or further manipulation. The documentation for the `NPP` function contained within the R package is shown in Appendix F, and an example script is shown in Appendix G.

6.4.5 FDPP

The `FDPP` function is the second front-end function contained within the `PPCRCT` package. A key additional argument (compared to `NPP`) that the user is required to specify when using `FDPP` is the `partial.borrowing` argument. If `TRUE`, a PBPP is fitted (as in Equation (5.4)), and if `FALSE` a FDPP is fitted (as in Equation (5.1)). Phases 1 and 2 of the `FDPP` function are the same as for the `NPP` function; namely data checking, followed by the default specification of the prior distributions. Next, the data are passed to one of two back-end functions depending on whether a PBPP or a FDPP is to be fitted; `PBPP_modelfit` for the former and `FDPP_modelfit` for the latter. In phase 3, the back-end function fits the model using the chosen prior distribution for the between-cluster SD, with each of phase 3a - 3d calling a separate `.stan` file for execution depending on the choice of model and prior distribution. After fitting the required model, the output is returned to the user as an object of class `"stanfit"`. The documentation for the `FDPP` function contained within the R package is shown in Appendix H, and an example script is shown in Appendix I.

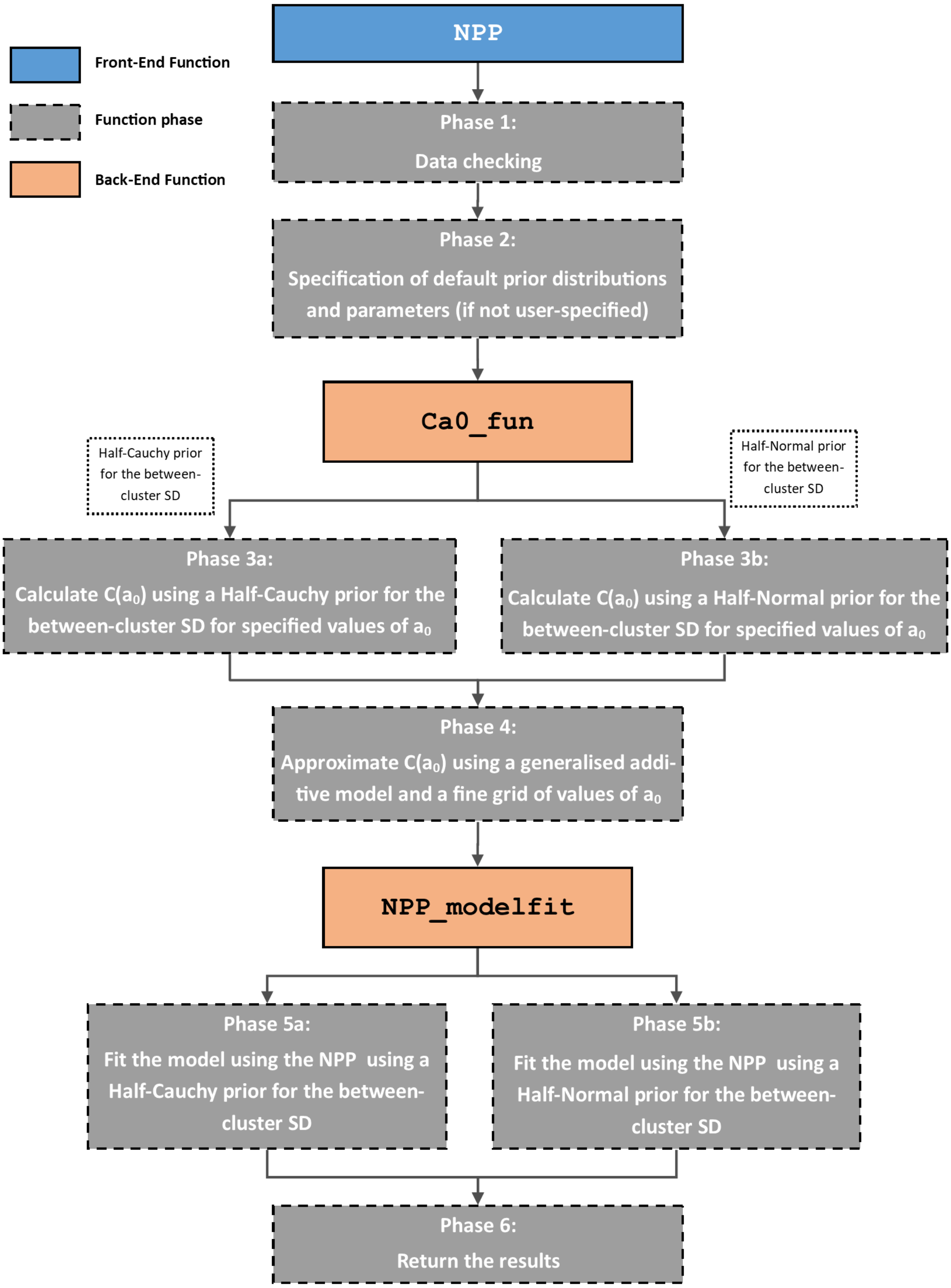

Front-End Function
Function phase
Back-End Function
NPP
Phase 1:
Data checking
Phase 2:
Specification of default prior distributions and parameters (if not user-specified)
Ca0_fun
Half-Cauchy prior for the between-cluster SD
Half-Normal prior for the between-cluster SD
Phase 3a:
Calculate C(a₀) using a Half-Cauchy prior for the between-cluster SD for specified values of a₀
Phase 3b:
Calculate C(a₀) using a Half-Normal prior for the between-cluster SD for specified values of a₀
Phase 4:
Approximate C(a₀) using a generalised additive model and a fine grid of values of a₀
NPP_modelfit
Phase 5a:
Fit the model using the NPP using a Half-Cauchy prior for the between-cluster SD
Phase 5b:
Fit the model using the NPP using a Half-Normal prior for the between-cluster SD
Phase 6:
Return the results

Figure 6.1: Schematic illustrating the structure of the `NPP` function.

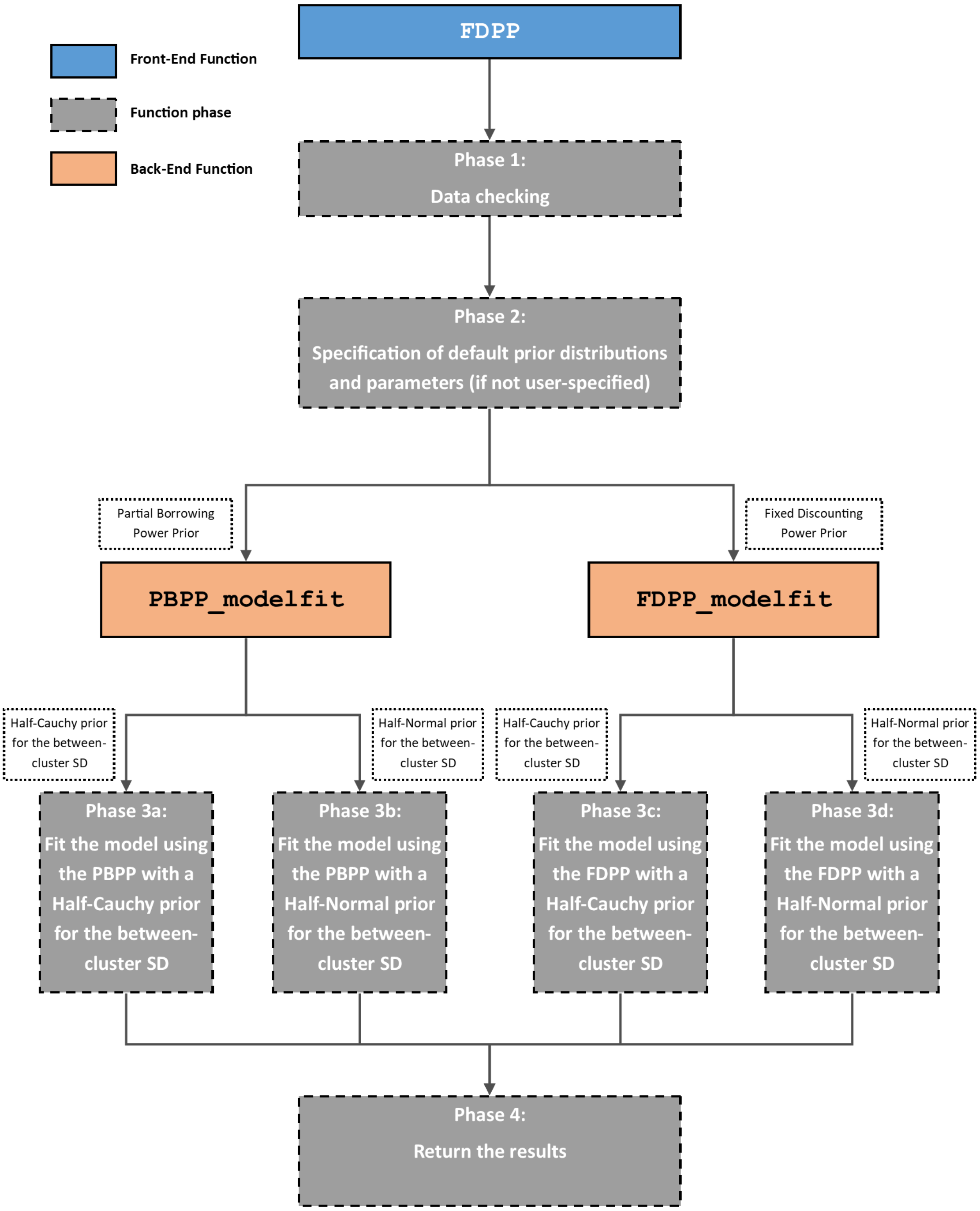

Front-End Function
Function phase
Back-End Function
FDPP
Phase 1:
Data checking
Phase 2:
Specification of default prior distributions and parameters (if not user-specified)
Partial Borrowing Power Prior
Fixed Discounting Power Prior
PBPP_modelfit
FDPP_modelfit
Half-Cauchy prior for the between-cluster SD
Half-Normal prior for the between-cluster SD
Half-Cauchy prior for the between-cluster SD
Half-Normal prior for the between-cluster SD
Phase 3a:
Fit the model using the PBPP with a Half-Cauchy prior for the between-cluster SD
Phase 3b:
Fit the model using the PBPP with a Half-Normal prior for the between-cluster SD
Phase 3c:
Fit the model using the FDPP with a Half-Cauchy prior for the between-cluster SD
Phase 3d:
Fit the model using the FDPP with a Half-Normal prior for the between-cluster SD
Phase 4:
Return the results

Figure 6.2: Schematic illustrating the structure of the `FDPP` function.

***

6.5 Discussion

Chapter 2 of this thesis presented a methodological systematic review, exploring both the use, and development, of Bayesian methodology in Cluster Randomised Controlled Trials. The review identified minimal use of Bayesian methods in practice, but some efforts to develop methodology. However, no papers were identified in which statistical software was developed/presented to support the practical use of the newly developed methodology, and this was identified as a potential barrier to the practical uptake of these novel methods. In response, the R package `PPCRCT` has been developed in order to increase the accessibility of the non-standard Bayesian methodology proposed within this thesis. `PPCRCT` has been designed and developed to provide a straightforward, user-friendly interface with which to fit power prior models to clustered trial data. In practice, it is envisaged that this package will not only help users to extract more value from their historical data, but also make the process of exploring the novel methods at the study design stage more accessible, by embedding the `PPCRCT` functionality within a simulation-based approach to sample size calculation.

Whilst `PPCRCT` was designed as a user-friendly tool to increase the accessibility of the relatively complex statistical methodology presented within this thesis, attention has also been given to trying to facilitate as much flexibility as possible, for example by allowing the user to specify values for the prior distribution parameters. Despite this, the level of flexibility currently provided by this package is limited when compared to the development of bespoke `.stan` files. For example, a user may wish to specify a more informative prior distribution for the discounting parameter; they may wish to include different regression coefficients in each dataset; or they may wish to borrow information from unclustered historical data, features which are not currently supported by `PPCRCT`. This reduction in flexibility is a necessary consequence of providing an accessible, user-friendly tool. However, like most modern software development projects, the development of the `PPCRCT` package is an ongoing, iterative process, and more features will be added at a later date to enhance usability further.

The version of `PPCRCT` presented within this thesis represents the first functional version of a tool that will continue to evolve as further methodological developments unfold. A key feature that would enhance the package would be the inclusion of functions to aid in study design and sample size calculation, such as the methodology presented in Chapter 4 and Chapter 5. However, at present, the limiting factor to adding such tools is the computational cost. The simulation-based approaches to sample size calculation employed within this thesis relied heavily on a high performance computing cluster to obtain results within a reasonable timescale (although still measured in days and weeks, rather than hours). Other features, such as increased flexibility as mentioned previously, or the extension of these methods to other types of outcome data (e.g. binary, count), may also be added to the package. Furthermore, GitHub provides a means through which users can provide feedback and report issues with the package encountered during use. These will be regularly monitored and issues and suggestions will be considered during the development and release of future versions.

To summarise, `PPCRCT` represents an easy-to-use, relatively flexible tool which can be used to fit complex Bayesian power prior models to clustered datasets such as those collected during CRCTs. It has been developed in response to earlier findings within this thesis that highlighted the importance of software development to support the application and uptake of novel statistical methodology. The version of the R package presented in this thesis represents an early version of an evolving tool that will be expanded in the future both in terms of its flexibility and its breadth of application to reflect further methodological developments.

Chapter 7

Discussion

The final chapter summarises the findings and newly developed methodology contained within this thesis. The relative strengths and drawbacks of the proposed methodologies are outlined, and opportunities for further development in the field are discussed.

CLUSTER Randomised Controlled Trials (CRCTs) have become increasingly commonplace in recent years, and the statistical methodology underpinning the design is now well-established in the literature. Furthermore, Bayesian statistical methods are growing in popularity the context of RCTs, with much research focused on the development of novel methodology in the fields of efficient designs, precision medicine, early phase studies and adaptive trials. Despite this, there remains only limited research undertaken to date which has focused on the development or application of Bayesian methods within the context of CRCT design and analysis. As such, this thesis explores just some of the ways in which Bayesian methodology can offer practical advantages over the traditional frequentist approach in CRCT design and analysis, focusing particularly on the construction of informative, data-driven prior distributions.

During the early stages of this project, an informal scoping review of the literature suggested that whilst there had been some methodological work undertaken in the use of Bayesian methods in CRCTs, the application of such methods remained uncommon. This finding motivated a methodological systematic review, which is presented in Chapter 2. The searches underpinning the review were undertaken in 2018, and subsequently updated in September 2021. The results of this methodological systematic review drew two key conclusions. Firstly, that the use of Bayesian methods in the analysis of CRCTs is rare, and even more so in the design of CRCTs, where no examples were identified. Secondly, that there is an opportunity for further methodological development in the field, in particular in the development of methods for specification of informative prior distributions, in statistical software development and in the application of Bayesian adaptive designs to CRCTs. The systematic review also sought to explore whether reporting quality of Bayesian CRCTs differed to the wider literature, but due to the small number of studies identified, the drawing of firm conclusions was not possible.

Within Chapter 3, the use of the power prior was explored in the context of the analysis of continuous CRCT data. Specifically, a novel power prior was proposed, which facilitates information borrowing from clustered historical trial data (e.g. pilot or feasibility study data) in the analysis of CRCT data, whilst automatically calibrating the strength of the information borrowed through estimation of the discounting factor. This methodology was first explored in the context of a re-analysis of the data from the definitive HeLP trial, where the HeLP pilot study data was used to construct the power prior. In comparison to analysis of the definitive trial data alone, use of the NPP resulted in a modest shift of the treatment effect estimate away from zero, and a modest increase in the estimate of the ICC. The methodology was subsequently evaluated through an extensive simulation study of different study sizes, treatment effects and ICCs. This simulation study demonstrated the sensitivity of the estimated discounting factor to varying agreement between the current and historical datasets, where more information borrowing was facilitated when the two datasets were more similar. It also showed that borrowing information through the NPP results in a reduced mean squared error, more precise estimation of the treatment effect and greater statistical power, although unsurprisingly it also introduced bias when the data generating mechanisms underpinning the current and historical datasets were not the same. Similarly, use of the NPP facilitated more precise estimation of the ICC, but also introduced some bias.

Chapter 4 outlined a simulation-based approach for determining the number of clusters required to achieve a desired level of statistical power, which allowed for the exploration of the NPP in the context of CRCT study design and sample size calculation. Underpinning this approach, the concept of a sampling prior was introduced as a means of expressing uncertainty in key design parameters such as the SD and the ICC. The impact of the use of the NPP on statistical power and type I error was quantified with application to a hypothetical redesign of the HeLP study and generalised through a simulation study. The results from the redesign of the HeLP study demonstrated that, had the study been designed to facilitate information borrowing from the pilot data using the NPP approach, it would have been possible to justify recruitment of fewer clusters in comparison to the frequentist approach to sample size determination. This conclusion remained true across all sampling-prior strategies. However, it was shown that placing a (non point-mass) sampling prior on the ICC had a detrimental effect on the required number of clusters, due to the substantial degree of uncertainty in estimating the ICC from historical data sources. This was true both for the NPP analyses and the approaches analysing the definitive trial data alone. The results also identified inflation of the one-sided type I error rate above the nominal rate. Similar findings were observed from the simulation study. Namely, increases in statistical power were demonstrated in scenarios where the treatment effect in the pilot data was of the same or greater magnitude (in the same direction) than the target effect size, with larger differences driving larger gains in power. In some scenarios, increases in power were shown even when the magnitude of the treatment effect in the pilot data was smaller than the target effect size. These gains in power were driven by either increased precision in the estimated treatment effect, a shifting of the treatment effect estimate away from the null hypothesis or, in some scenarios, a combination of both. The simulation study also showed that inflated type I error rates were induced whenever a non-null treatment effect was used to generate the historical data.

Chapter 5 extended a recently proposed approach for sample size calculation when using power priors for information borrowing to CRCTs. This approach involves the adoption of a Bayesian interpretation of statistical power and type I error, both traditionally frequentist concepts. By placing a sampling prior on the treatment effect parameter, a Bayesian interpretation of the study operating characteristics becomes inherent. A two-stage approach was outlined: in the first stage, a fixed value for the discounting parameter was chosen to maximise the amount of information borrowed from the historical data whilst maintaining Bayesian type I error at some nominal level. In the second stage, a power prior with fixed discounting parameter determined in the first stage was specified, and a simulation-based approach adopted to determine the required number of clusters to achieve the desired level of statistical power. This method was applied to a redesign of the HeLP study, using a range of null and alternative sampling priors constructed from the posterior distribution of the treatment effect from the historical data, and using two types of power prior: the FDPP, which borrows information from all parameters, and the PBPP, which borrows information only from the treatment effect parameter. The results showed that when the null sampling prior was based on the posterior treatment effect without truncation, the entirety of the historical evidence could be incorporated without inflation of the Bayesian type I error. The amount of information borrowing allowed was reduced as the truncation point became closer to zero. The results also showed that statistical power was increased when alternative sampling priors expressed greater support for treatment effects of larger magnitude. In some scenarios, more efficient study design was possible using the PBPP, but in others the FDPP led to reduced sample size requirements. However, it may be difficult to justify borrowing information only from the treatment effect parameter, and so use of the FDPP is generally recommended where possible.

In the final chapter, Chapter 6, an R package, `PPCRCT`, was introduced. `PPCRCT` is a statistical tool, readily available for download and use from GitHub (`https://github.com/benjones13/PPCRCT`) which allows for straightforward implementation of the three key power priors methods outlined within this thesis: (i) the NPP; (ii) the FDPP; and (iii) the PBPP. The package provides the user with flexibility in terms of specifying parameters for the prior distributions. Whilst the intention is to add further functionality in future, the existing tool is functional and user friendly and allows for use of the three power prior models without the need for any bespoke Stan programming. As a result, this tool should aid in the accessibility and uptake of the novel methodology outlined within this thesis.

***

7.1 Strengths and Limitations

A common criticism of the use of Bayesian methodology is the perceived potential to incorporate subjectivity into an analysis through specification of prior distributions. A key strength of the methodology outlined within this thesis is that it addresses this very concern; the power prior allows construction of a highly informative, yet entirely evidence-based, data-driven prior distribution. Furthermore, given the methodology proposes to use historical information such as that collected from pilot or feasibility studies, the data underpinning the construction of these prior distributions are likely to be both highly relevant and of high quality. As a result, this work provides a framework for incorporating existing evidence into CRCTs, which could be used for primary analyses, sensitivity analyses or even during the design and sample size calculation.

Furthermore, the results presented within this thesis clearly demonstrate the potential of the novel methods to facilitate more efficient study design by making maximal use of high quality data from pilot and feasibility studies, which themselves are often large, time-consuming and expensive bodies of research. Given the importance of minimising research waste and maximising the value of research efforts, the potential efficiency gains available through the use of this methodology are a significant strength.

One of the key barriers associated with the uptake of novel, modern statistical methodology, such as that presented within this thesis, is the time and effort often required to implement the methodology. In order to address this barrier, the `PPCRCT` package has been developed to allow straightforward implementation of the complex approaches proposed. As such, this thesis has not only outlined novel methodology, but also provided a means through which it can be applied in practice.

More generally, the use of Bayesian methodology offers an advantage over the frequentist approach through the ease of interpretation of results. Specifically, the probabilistic interpretation of results facilitated through Bayesian inference is more natural and intuitive compared to the hypothetical, long-run average interpretation offered through the frequentist approach. This probabilistic interpretation has the potential to encourage more meaningful discussions with clinicians and non-statistical triallists, and ultimately more robust and well-informed decision making. If properly communicated, this general advantage associated with Bayesian methodology could become a key justification for the adoption of methodology such as that presented within this thesis.

Despite the clear strengths associated with the novel methodology proposed within this thesis, there remain barriers to practical uptake, perhaps the most significant of which is the computational intensiveness associated with the use of such methods. Whilst MCMC procedures can themselves be computationally intensive, state-of-the-art statistical software, such as Stan, can often help to ease this burden. For this reason, the use of the power priors in analysis alone is unlikely to pose a significant challenge in terms of computational cost, with such models typically running within minutes on standard hardware. However, when considering these approaches in the context of study design and sample size calculation, which rely on extensive simulation-based approaches, the computational cost becomes exceptionally high. Indeed, all of the simulation-based sample size calculations presented within this thesis relied upon a high performance computing cluster, with run times taking days, or in some cases even weeks, rather than hours. These challenges are yet more pronounced when attempting to achieve 90%, rather than 80% statistical power (as is now commonplace for definitive trial design). This is as a result of power curves beginning to level off as values get closer to 100%, meaning the addition of clusters represents incrementally smaller power gains, thus requiring the simulation of a greater number of study scenarios. The computational cost is particularly high when implementing the methodology presented in Chapter 5, given the two-stage simulation based approach where both the discounting parameter and the study size are to be determined. In an attempt to address these computational challenges, a weighted maximum likelihood approach was explored as an approximation of the power prior with fixed discounting factor. However, this approach relied on optimisation methodology, which is itself computationally intensive, and appeared not to offer any notable advantage in computation time.

Whilst Bayesian methodology is becoming increasingly common in the design, conduct and analysis of clinical trials, it is still typically considered a non-standard approach, particularly in late phase trials, pragmatic trials, and trials of complex interventions, which are typical of CRCTs. As a result, there remain barriers to overcome before Bayesian methods become more widely accepted and used within CRCTs, as demonstrated by the small number of studies identified within the systematic review presented in Chapter 2. For the methodology presented within this thesis to be adopted in the design of large, definitive CRCTs, a significant step change in attitude to novel Bayesian methods would, at the time of writing, be required. However, such methods are becoming more commonplace and accepted, for example in adaptive designs and platform trials, with this change perhaps being accelerated by the novel trial designs used in response to the COVID-19 pandemic. Perhaps alongside this increasing acceptability, the novel methodology proposed within this thesis could initially be used to undertake supplementary or sensitivity analyses, before becoming a more feasible approach for primary analysis and/or study design in parallel with increases in uptake of Bayesian methodology within trial design and analysis in the future.

***

7.2 Future Work

All of the methodology proposed within this thesis has focused on application to continuous outcome data. Whilst continuous outcomes are perhaps the most frequently used in CRCTs, the use of binary and, to a lesser extent, count outcomes are common. As such, a natural area of focus for future work would be the extension of the power prior methodology to the analysis of binary data (via hierarchical logistic regression models) and count data (via hierarchical Poisson or negative-binomial regression).

A further opportunity for future work in this area is the extension of the proposed methodology to facilitate information borrowing from multiple historical datasets, rather than just a single source. Such methodology has already been proposed in the wider power prior literature [Ibrahim et al., 2015], but has not been discussed in the context of hierarchical models or CRCTs. Whilst the narrative within this thesis has suggested borrowing from a study's associated pilot or feasibility study, borrowing from multiple data sources has the potential to further enhance the value of this methodology. For example, if data from multiple relevant historical trials (with either cluster or individual-level randomisation, or a combination of both) assessing similar interventions was available, all of this evidence could be incorporated within the power prior. Such a model could be thought of as being similar to an individual patient data meta-analysis-style approach, with weighting of each study according to its own discounting factor, to specify an informative prior distribution. However, in exploring or adopting such an approach, care must be taken that the multiple sources of evidence are of sufficient quality and relevance to be included within the power prior analysis.

The problem of study design and sample size calculation for a CRCT is complex, not least because sample size can be determined not only through cluster-level recruitment targets, but also through individual-level recruitment targets. That is, in scenarios where cluster sizes are not fixed, CRCTs can increase statistical power not only by increasing the number of clusters recruited, but also by increasing the number of participants within each cluster. Within this thesis, the simulation studies pertaining to study design and sample size calculation have focused on scenarios in which the cluster size is fixed, simplifying the calculation to one of determining the required number of clusters (and hence sample size). However, in many cases, this may be an oversimplification. Future work could address this by considering increases in cluster size instead of increases in the number of clusters, or indeed a combination of both. Of course, regardless of evidence provided by simulation studies, the practical application of the power prior methods in study design would always need to be appropriately tailored to accommodate the design characteristics of the study in question.

Alongside any further methodological developments in the use of power priors in CRCTs, there is clear value in updating and increasing the functionality of `PPCRCT` to ensure that it is able to implement such developments. Notwithstanding the addition of new methodology to the package, there is potential for future work to increase the functionality, for example by allowing further flexibility in model fitting and prior specification, or by adding functions to visualise the results of the power prior analyses.

More generally, there remains significant opportunity for further work in the use of Bayesian methods in the design and analysis of CRCTs. A key area which could be the focus of future research is in Bayesian adaptive CRCT designs. Such methodology is becoming a focus of increased research in the context of individually randomised RCTs, but does not yet appear to be a topic of significant interest in cluster-randomised designs. Indeed there may be potential to combine the power prior approaches outlined within this thesis with adaptive design features, such as through interim analyses incorporating information borrowing.

***

7.3 Concluding Remarks

The novel methodology presented in this thesis has the potential to facilitate more efficient CRCT design by making maximal use of high quality, highly relevant pilot or feasibility study data to construct informative prior distributions, thus reducing research waste. Despite this, there remain significant barriers to practical uptake, namely the computational cost associated with simulation-based power calculations, and the acceptability of novel Bayesian methodology underpinning study design and primary analyses. There remains opportunity for further work in this area, both in the context of power priors, and the wider development of Bayesian methodology with application to CRCTs.

List of references

[Adab et al., 2018] Adab, P., Pallan, M. J., Lancashire, E. R., Hemming, K., Frew, E., Barrett, T., Bhopal, R., Cade, J. E., Canaway, A., Clarke, J. L., Daley, A., Deeks, J. J., Duda, J. L., Ekelund, U., Gill, P., Griffin, T., Mcgee, E., Hurley, K., Martin, J., Parry, J., Passmore, S., and Cheng, K. K. (2018). Effectiveness of a childhood obesity prevention programme delivered through schools, targeting 6 and 7 year olds: cluster randomised controlled trial (WAVES study). *BMJ*, 360:211.

[Adams et al., 2004] Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., and Campbell, M. J. (2004). Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, 57(8):785–794.

[Andrieu and Thoms, 2008] Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373.

[Arnold et al., 2011] Arnold, B. F., Hogan, D. R., Colford, J. M., and Hubbard, A. E. (2011). Simulation methods to estimate design power: An overview for applied research. *BMC Medical Research Methodology*, 11(1):1–10.

[Bates et al., 2015] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

[Betancourt, 2018] Betancourt, M. (2018). A conceptual introduction to hamiltonian monte carlo. arXiv1701.02434.

[Biro and Wien, 2010] Biro, F. M. and Wien, M. (2010). Childhood obesity and adult morbidities. *The American Journal of Clinical Nutrition*, 91(5):1499–1505.

[Bland, 2004] Bland, J. M. (2004). Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Medical Research Methodology*, 4(1):21.

[Bland and Altman, 1998] Bland, J. M. and Altman, D. G. (1998). Bayesians and frequentists. *BMJ*, 317(7166):1151.

[Breheny et al., 2020] Breheny, K., Passmore, S., Adab, P., Martin, J., Hemming, K., Lancashire, E. R., and Frew, E. (2020). Effectiveness and cost-effectiveness of The Daily Mile on childhood weight outcomes and wellbeing: a cluster randomised controlled trial. *International Journal of Obesity*, 44(4):812–822.

[Brown et al., 2016] Brown, A. R., Gajewski, B. J., Aaronson, L. S., Mudaranthakam, D. P., Hunt, S. L., Berry, S. M., Quintana, M., Pasnoor, M., Dimachkie, M. M., Jawdat, O., Herbelin, L., and Barohn, R. J. (2016). A Bayesian comparative effectiveness trial in action: Developing a platform for multisite study adaptive randomization. *Trials*, 17(1):428.

[Bürkner, 2018] Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411.

[Campbell et al., 2004] Campbell, M. K., Elbourne, D. R., and Altman, D. G. (2004). CONSORT statement: extension to cluster randomised trials. *BMJ*, 328(7441):702–708.

[Campbell et al., 2012] Campbell, M. K., Piaggio, G., Elbourne, D. R., and Altman, D. G. (2012). Consort 2010 statement: Extension to cluster randomised trials. *BMJ*, 345(7881).

[Carlin et al., 1992] Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). Hierarchical Bayesian Analysis of Changepoint Problems. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):389.

[Carpenter et al., 2017] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

[Carvalho and Ibrahim, 2021] Carvalho, L. M. and Ibrahim, J. G. (2021). On the normalized power prior. *Statistics in Medicine*, 40(24):5251–5275.

[Chen et al., 2014a] Chen, M. H., Ibrahim, J. G., Amy Xia, H., Liu, T., and Hennessey, V. (2014a). Bayesian sequential meta-analysis design in evaluating cardiovascular risk in a new antidiabetic drug development program. *Statistics in Medicine*, 33(9):1600–1618.

[Chen et al., 2011] Chen, M. H., Ibrahim, J. G., Lam, P., Yu, A., and Zhang, Y. (2011). Bayesian Design of Noninferiority Trials for Medical Devices Using Historical Data. *Biometrics*, 67(3):1163–1170.

[Chen et al., 2014b] Chen, M.-H., Ibrahim, J. G., Zeng, D., Hu, K., and Jia, C. (2014b). Bayesian design of superiority clinical trials for recurrent events data with applications to bleeding and transfusion events in myelodyplastic syndrome. *Biometrics*, 70(4):1003–1013.

[Chen et al., 2014c] Chen, M. H., Ibrahim, J. G., Zeng, D., Hu, K., and Jia, C. (2014c). Bayesian design of superiority clinical trials for recurrent events data with applications to bleeding and transfusion events in myelodyplastic syndrome. *Biometrics*, 70(4):1003–1013.

[Clark, 2014] Clark, M. (2014). Getting Started with Additive Models in R. Technical report, University of Notre Dam. Accessed: 05-08-2021.

[De Santis, 2006] De Santis, F. (2006). Power Priors and Their Use in Clinical Trials. *The American Statistician*, 60(2):122–129.

[Dechartres et al., 2011] Dechartres, A., Charles, P., Hopewell, S., Ravaud, P., and Altman, D. G. (2011). Reviews assessing the quality or the reporting of randomized controlled trials are increasing over time but raised questions about how quality is assessed. *Journal of Clinical Epidemiology*, 64(2):136–144.

[Delgado-Rodriguez et al., 2001] Delgado-Rodriguez, M., Ruiz-Canela, M., De Irala-Estevez, J., Llorca, J., and Martinez-Gonzalez, A. (2001). Participation of epidemiologists and/or biostatisticians and methodological quality of published controlled clinical trials. *Journal of Epidemiology and Community Health*, 55(8):569–72.

[Diaz-Ordaz et al., 2013] Diaz-Ordaz, K., Froud, R., Sheehan, B., and Eldridge, S. (2013). A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. *BMC Medical Research Methodology*, 13(1):127.

[Dienes, 2014] Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5:781.

[Donner et al., 1981] Donner, A., Birkett, N., and Buck, C. (1981). Randomization by cluster: Sample size requirements and analysis. *American Journal of Epidemiology*, 114(6):906–914.

[Donner et al., 1990] Donner, A., Brown, K. S., and Brasher, P. (1990). A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *International Journal of Epidemiology*, 19(4):795–800.

[Duan et al., 2006] Duan, Y., Ye, K., and Smith, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1):95–106.

[Eldridge and Kerry, 2012] Eldridge, S. and Kerry, S. (2012). *A Practical Guide to Cluster Randomised Trials in Health Services Research*. John Wiley & Sons, Inc.

[Eldridge et al., 2006] Eldridge, S. M., Ashby, D., and Kerry, S. (2006). Sample size for cluster randomized trials: Effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology*, 35(5):1292–1300.

[Eldridge et al., 2016] Eldridge, S. M., Chan, C. L., Campbell, M. J., Bond, C. M., Hopewell, S., Thabane, L., Lancaster, G. A., Altman, D., Bretz, F., Campbell, M., Cobo, E., Craig, P., Davidson, P., Groves, T., Gumedze, F., Hewison, J., Hirst, A., Hoddinott, P., Lamb, S. E., Lang, T., McColl, E., O'Cathain, A., Shanahan, D. R., Sutton, C., and Tugwell, P. (2016). CONSORT 2010 statement: Extension to randomised pilot and feasibility trials. *BMJ*, 355.

[Eldridge et al., 2015] Eldridge, S. M., Costelloe, C. E., Kahan, B. C., Lancaster, G. A., and Kerry, S. M. (2015). How big should the pilot study for my cluster randomised trial be? *Statistical Methods in Medical Research*, 25(3):1039–1056.

[Elsevier, 2020] Elsevier (2020). Mendeley.

[Froud et al., 2012] Froud, R., Eldridge, S., Diaz Ordaz, K., Marinho, V. C. C., and Donner, A. (2012). Quality of cluster randomized controlled trials in oral health: A systematic review of reports published between 2005 and 2009. *Community Dentistry and Oral Epidemiology*, 40(SUPPL. 1):3–14.

[Fuglstad et al., 2020] Fuglstad, G. A., Hem, I. G., Knight, A., Rue, H., and Riebler, A. (2020). Intuitive Joint Priors for Variance Parameters. *Bayesian Analysis*, 15(4):1109–1137.

[Gelfand and Smith, 1990] Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

[Gelman, 2006] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

[Gelman et al., 2013] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis, third edition*.

[George and McCulloch, 1993] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

[Gittins and Pezeshk, 2000a] Gittins, J. and Pezeshk, H. (2000a). A behavioral bayes method for determining the size of a clinical trial. *Therapeutic Innovation & Regulatory Science*, 34(2):355–363.

[Gittins and Pezeshk, 2000b] Gittins, J. and Pezeshk, H. (2000b). How large should a clinical trial be? *Journal of the Royal Statistical Society Series D: The Statistician*, 49(2):177–187.

[Gittins and Pezeshk, 2002] Gittins, J. C. and Pezeshk, H. (2002). A decision theoretic approach to sample size determination in clinical trials. *Journal of Biopharmaceutical Statistics*, 12(4):535–551.

[Golchi, 2020] Golchi, S. (2020). Use of historical individual patient data in analysis of clinical trials. arXiv2002.09910.

[Goodrich et al., 2020] Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.19.3.

[Gordon et al., 2021] Gordon, A. C., Mouncey, P. R., Al-Beidh, F., Rowan, K. M., Nichol, A. D., Arabi, Y. M., Annane, D., Beane, A., Van Bentum-Puijk, W., Berry, L. R., Bhimani, Z., Bonten, M. J. M., Bradbury, C. A., Brunkhorst, F. M., Buzgau, A., Cheng, A. C., Detry, M. A., Duffy, E. J., Est-Court, L. J., Fitzgerald, M., Goossens, H., Haniffa, R., Higgins, A. M., Hills, T. E., Horvat, C. M., Lamontagne, F., Lawler, P. R., Leavis, H. L., Linstrum, K. M., Litton, E., Lo-Renzi, E., Marshall, J. C., Mayr, F. B., Mcauley, D. F., Mcglothlin, A., Mcguin-Ness, S. P., Mcverry, B. J., Montgomery, S. K., Morpeth, S. C., Murthy, S., Orr, K., Parke, R. L., Parker, J. C., Patanwala, A. E., Pet-Tilä, V., Rademaker, E., Santos, C. T., Saunders, C. W., Seymour, M., Shankar-Hari, W. I., Sligl, A. F., Turgeon, A. M., Turner, F. L., Van De Veerdonk, R., Zarychanski, C., Green, R. J., Lewis, D. C., Angus, C. J., Mc-Arthur, S., Berry, S. A., and Webb, L. P. G. (2021). Interleukin-6 Receptor Antagonists in Critically Ill Patients with Covid-19. *New England Journal of Medicine*, 384(16):1491–1502.

[Gravestock and Held, 2017] Gravestock, I. and Held, L. (2017). Adaptive power priors with empirical Bayes for clinical trials. *Pharmaceutical Statistics*, 16(5):349–360.

[Gronau et al., 2017] Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E. J., and Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81:80–97.

[Gronau et al., 2020] Gronau, Q. F., Singmann, H., and Wagenmakers, E. J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(1):1–29.

[Gronau et al., 2019] Gronau, Q. F., Wagenmakers, E. J., Heck, D. W., and Matzke, D. (2019). A Simple Method for Comparing Complex Models: Bayesian Model Comparison for Hierarchical Multinomial Processing Tree Models Using Warp-III Bridge Sampling. *Psychometrika*, 84(1):261–284.

[Grydeland et al., 2014] Grydeland, M., Bjelland, M., Anderssen, S. A., Klepp, K. I., Bergh, I. H., Andersen, L. F., Ommundsen, Y., and Lien, N. (2014). Effects of a 20-month cluster randomised controlled school-based intervention trial on BMI of school-aged boys and girls: The HEIA study. *British Journal of Sports Medicine*, 48(9):768–773.

[Hammersley and Handscomb, 1964] Hammersley, J. and Handscomb, D. (1964). *Monte Carlo Methods*. Springer Netherlands.

[Hastings, 1970] Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97.

[Hees and Kieser, 2017] Hees, K. and Kieser, M. (2017). Blinded sample size recalculation in clinical trials incorporating historical data. *Contemporary Clinical Trials*, 63:2–7.

[Hemming et al., 2018] Hemming, K., Taljaard, M., McKenzie, J. E., Hooper, R., Copas, A., Thompson, J. A., Dixon-Woods, M., Aldcroft, A., Doussau, A., Grayling, M., Kristunas, C., Goldstein, C. E., Campbell, M. K., Girling, A., Eldridge, S., Campbell, M. J., Lilford, R. J., Weijer, C., Forbes, A. B., and Grimshaw, J. M. (2018). Reporting of stepped wedge cluster randomised trials: Extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ*, 363:1614.

[Hobbs et al., 2011] Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., and Sargent, D. J. (2011). Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials. *Biometrics*, 67(3):1047–1056.

[Hobbs et al., 2012] Hobbs, B. P., Sargent, D. J., and Carlin, B. P. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*, 7(3):639–674.

[Hoeting et al., 1999] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401.

[Ibrahim and Chen, 2000] Ibrahim, J. G. and Chen, M.-H. (2000). Power Prior Distributions for Regression Models. *Statistical Science*, 15(1):46–60.

[Ibrahim et al., 2015] Ibrahim, J. G., Chen, M. H., Gwon, Y., and Chen, F. (2015). The power prior: Theory and applications. *Statistics in Medicine*, 34(28):3724–3749.

[Ibrahim et al., 2012] Ibrahim, J. G., Chen, M. H., Xia, H. A., and Liu, T. (2012). Bayesian Meta-Experimental Design: Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes. *Biometrics*, 68(2):578–586.

[Ivers et al., 2011] Ivers, N. M., Taljaard, M., Dixon, S., Bennett, C., McRae, A., Taleban, J., Skea, Z., Brehaut, J. C., Boruch, R. F., Eccles, M. P., Grimshaw, J. M., Weijer, C., Zwarenstein, M., and Donner, A. (2011). Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *BMJ*, 343:d5886.

[Jones, 2018] Jones, B. (2018). The use of Bayesian Statistics in the design and analysis of cluster randomised controlled trials and their methodological and reporting quality: a protocol for an international methodological review. Accessed: 05-03-2020.

[Jones et al., 2021] Jones, B. G., Streeter, A. J., Baker, A., Moyeed, R., and Creanor, S. (2021). Bayesian statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review. *Systematic Reviews*, 10(1).

[Kass and Raftery, 1995] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

[Lauer et al., 2015] Lauer, S. A., Kleinman, K. P., and Reich, N. G. (2015). The effect of cluster size variability on statistical power in cluster-randomized trials. *PLoS ONE*, 10(4).

[Lewis and Wears, 1993] Lewis, R. J. and Wears, R. L. (1993). An introduction to the Bayesian analysis of clinical trials. *Annals of Emergency Medicine*, 22(8):1328–1336.

[Li et al., 2019] Li, B., Pallan, M., Liu, W. J., Hemming, K., Frew, E., Lin, R., Liu, W., Martin, J., Zanganeh, M., Hurley, K., Cheng, K. K., and Adab, P. (2019). The CHIRPY DRAGON intervention in preventing obesity in Chinese primaryschool-aged children: A cluster-randomised controlled trial. *PLoS Medicine*, 16(11).

[Liang and Zeger, 1986] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

[Liu et al., 2019] Liu, Z., Li, Q., Maddison, R., Ni Mhurchu, C., Jiang, Y., Wei, D. M., Cheng, L., Cheng, Y., Wang, D., and Wang, H. J. (2019). A School-Based Comprehensive Intervention for Childhood Obesity in China: A Cluster Randomized Controlled Trial. *Childhood Obesity*, 15(2):105–115.

[Lloyd et al., 2018] Lloyd, J., Creanor, S., Logan, S., Green, C., Dean, S. G., Hillsdon, M., Abraham, C., Tomlinson, R., Pearson, V., Taylor, R. S., Ryan, E., Price, L., Streeter, A., and Wyatt, K. (2018). Effectiveness of the Healthy Lifestyles Programme (HeLP) to prevent obesity in UK primary-school children: a cluster randomised controlled trial. *The Lancet Child and Adolescent Health*, 2(1):35–45.

[Lloyd et al., 2012] Lloyd, J. J., Wyatt, K. M., and Creanor, S. (2012). Behavioural and weight status outcomes from an exploratory trial of the Healthy Lifestyles Programme (HeLP): A novel school-based obesity prevention programme. *BMJ Open*, 2(3).

[Lunn et al., 2000] Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

[Meng and Schilling, 2002] Meng, X. L. and Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586.

[Meng and Wong, 1996] Meng, X. L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6(4):831–860.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

[Mitchell and Beauchamp, 1988] Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

[Moberg and Kramer, 2015] Moberg, J. and Kramer, M. (2015). A brief history of the cluster randomised trial design. *Journal of the Royal Society of Medicine*, 108(5):192–198.

[Moher et al., 2010] Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., and Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340:869.

[Moher et al., 2009] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7):e1000097.

[Morris et al., 2019] Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

[Nanni et al., 2020] Nanni, O., Viale, P., Vertogen, B., Lilli, C., Zingaretti, C., Donati, C., Masini, C., Monti, M., Serra, P., Vespignani, R., Grossi, V., Biggeri, A., Scarpi, E., Galardi, F., Bertoni, L., Colamartini, A., Falcini, F., Altini, M., Massa, I., Gaggeri, R., and Martinelli, G. (2020). PROTECT Trial: A cluster-randomized study with hydroxychloroquine versus observational support for prevention or early-phase treatment of Coronavirus disease (COVID-19): A structured summary of a study protocol for a randomized controlled trial. *Trials*, 21(1):1–4.

[Neal, 2012] Neal, R. M. (2012). Mcmc using hamiltonian dynamics. arXiv1206.1901.

[Neelon and O'Malley, 2010] Neelon, B. and O'Malley, A. J. (2010). Bayesian Analysis Using Power Priors with Application to Pediatric Quality of Care. *Journal of Biometrics & Biostatistics*, 01(01).

[Neuenschwander et al., 2009] Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566.

[Neuenschwander et al., 2010] Neuenschwander, B., Capkun-Niggli, G., Branson, M., and Spiegelhalter, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18.

[O'Hagan and Stevens, 2001] O'Hagan, A. and Stevens, J. W. (2001). Bayesian Assessment of Sample Size for Clinical Trials of Cost-Effectiveness. *Medical Decision Making*, 21(3):219–230.

[Ouzzani et al., 2016] Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1):210.

[Pallmann et al., 2018] Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., Villar, S. S., Wason, J. M., Weir, C. J., Wheeler, G. M., Yap, C., and Jaki, T. (2018). Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Medicine*, 16(1):1–15.

[Pezeshk and Gittins, 2002] Pezeshk, H. and Gittins, J. (2002). A fully bayesian approach to calculating sample sizes for clinical trials with binary responses. *Therapeutic Innovation & Regulatory Science*, 36(1):143–150.

[Plummer, 2003] Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

[Pocock, 1976] Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29(3):175–188.

[Psioda and Ibrahim, 2019] Psioda, M. A. and Ibrahim, J. G. (2019). Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics (Oxford, England)*, 20(3):400–415.

[Public Health England, 2015] Public Health England (2015). Childhood obesity: applying all our health. https://www.gov.uk/government/publications/childhood-obesity-applying-all-our-health/childhood-obesity-applying-all-our-health. Accessed: 09-10-2021.

[R Core Team, 2019] R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Robert and Casella, 2011] Robert, C. and Casella, G. (2011). A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data 1. *Statistical Science*, 26(1):102–115.

[Roberts et al., 1997] Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1):110–120.

[Roberts and Rosenthal, 2001] Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367.

[Rubin and Stern, 1998] Rubin, D. B. and Stern, H. S. (1998). Sample Size Determination Using Posterior Predictive Distributions. *The Indian Journal of Statistics*, 60(1):161–175.

[Rutterford et al., 2015] Rutterford, C., Taljaard, M., Dixon, S., Copas, A., and Eldridge, S. (2015). Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: A review. *Journal of Clinical Epidemiology*, 68(6):716–723.

[Ryan et al., 2019] Ryan, E. G., Bruce, J., Metcalfe, A. J., Stallard, N., Lamb, S. E., Viele, K., Young, D., and Gates, S. (2019). Using Bayesian adaptive designs to improve phase III trials: A respiratory care example. *BMC Medical Research Methodology*, 19(1):99.

[Schmidli et al., 2014] Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032.

[Spiegelhalter, 2001] Spiegelhalter, D. J. (2001). Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine*, 20(3):435–452.

[Spiegelhalter et al., 2004] Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*.

[Spiegelhalter and Freedman, 1986] Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5(1):1–13.

[Stan Development Team., 2020] Stan Development Team. (2020). Prior Choice Recommendations. Accessed: 20-05-2021.

[Stan Development Team, 2022] Stan Development Team (2022). RStan: the R interface to Stan. R package version 2.21.5.

[StataCorp, 2021] StataCorp (2021). *Stata Statistical Software: Release 17*. StataCorp LLC, College Station, TX.

[Swiger et al., 1964] Swiger, L. A., Harvey, W. R., Everson, D. O., and Gregory, K. E. (1964). The Variance of Intraclass Correlation Involving Groups with One Observation. *Biometrics*, 20(4):818.

[Taljaard et al., 2010] Taljaard, M., McGowan, J., Grimshaw, J. M., Brehaut, J. C., McRae, A., Eccles, M. P., and Donner, A. (2010). Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: Low precision will improve with adherence to reporting standards. *BMC Medical Research Methodology*, 10(15).

[The NHS Information Centre, 2011] The NHS Information Centre (2011). National Child Measurement Programme - England, 2010-11, School year. `https://digital.nhs.uk/data-and-information/publications/statistical/national-child-measurement-programme/2010-11-school-year#summary`. Accessed: 15-12-2020.

[Tokolahi et al., 2016] Tokolahi, E., Hocking, C., Kersten, P., and Vandal, A. C. (2016). Quality and Reporting of Cluster Randomized Controlled Trials Evaluating Occupational Therapy Interventions: A Systematic Review. *OTJR: Occupation, Participation and Health*, 36(1):1539449215618625.

[Turner et al., 2004] Turner, R. M., Prevost, A. T., and Thompson, S. G. (2004). Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Statistics in Medicine*, 23(8):1195–1214.

[Turner et al., 2005] Turner, R. M., Thompson, S. G., and Spiegelhalter, D. J. (2005). Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical trials*, 2(2):108–118.

[Ukoumunne, 2002] Ukoumunne, O. C. (2002). A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine*, 21(24):3757–3774.

[Ukoumunne et al., 1999] Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, J. A., and Burney, P. G. (1999). Methods for evaluating area-wide and organisation-based interventions in health and health care: A systematic review. *Health Technology Assessment*, 3(5).

[van Rosmalen et al., 2018] van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B., and Lesaffre, E. (2018). Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research*, 27(10):3167–3182.

[Vehtari et al., 2021] Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Burkner, P. C. (2021). Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)*†. *Bayesian Analysis*, 16(2):667–718.

[Viele et al., 2014] Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., and Thompson, L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54.

[Wang et al., 1993] Wang, C., Rutledge, J., and Gianola, D. (1993). Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics Selection Evolution*, 25(1):41.

[Wang et al., 1994] Wang, C., Rutledge, J., and Gianola, D. (1994). Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genetics Selection Evolution*, 26(2):91.

[Wang and Gelfand, 2002] Wang, F. and Gelfand, A. E. (2002). A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models. *Statistical Science*, 17(2):193–208.

[Wasserstein et al., 2019] Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a World Beyond "$p < 0.05$". *American Statistician*, 73(sup1):1–19.

[Waters et al., 2017] Waters, E., Gibbs, L., Tadic, M., Ukoumunne, O. C., Magarey, A., Okely, A. D., De Silva, A., Armit, C., Green, J., O'Connor, T., Johnson, B., Swinburn, B., Carpenter, L., Moore, G., Littlecott, H., and Gold, L. (2017). Cluster randomised trial of a school-community child health promotion and obesity prevention intervention: Findings from the evaluation of fun 'n healthy in Moreland! *BMC Public Health*, 18(1):92.

[Weber et al., 2019] Weber, S., Li, Y., Seaman, J., Kakizume, T., and Schmidli, H. (2019). Applying meta-analytic-predictive priors with the r bayesian evidence synthesis tools. arXiv1907.00603.

[Wheeler et al., 2019] Wheeler, G. M., Mander, A. P., Bedding, A., Brock, K., Cornelius, V., Grieve, A. P., Jaki, T., Love, S. B., Odondi, L., Weir, C. J., Yap, C., and Bond, S. J. (2019). How to design a dose-finding study using the continual reassessment method. *BMC Medical Research Methodology*, 19(1):18.

[Wickham, 2015] Wickham, H. (2015). *R Packages*. O'Reilly.

[Wood, 2017] Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition.

[Wood, 2004] Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.

[World Health Organisation, 2020] World Health Organisation (2020). Childhood overweight and obesity. http://www.who.int/dietphysicalactivity/childhood/en/. Accessed: 09-10-2021.

[Wyatt et al., 2013] Wyatt, K. M., Lloyd, J. J., Abraham, C., Creanor, S., Dean, S., Densham, E., Daurge, W., Green, C., Hillsdon, M., Pearson, V., Taylor, R. S., Tomlinson, R., and Logan, S. (2013). The healthy lifestyles programme (help), a novel school-based intervention to prevent obesity in school children: study protocol for a randomised controlled trial. *Trials*, 14(1):95.

[Wyatt et al., 2011] Wyatt, K. M., Lloyd, J. J., Creanor, S., and Logan, S. (2011). The development, feasibility and acceptability of a school-based obesity prevention programme: results from three phases of piloting. *BMJ Open*, 1(1):e000026–e000026.

[Xiao, 2017] Xiao, S. (2017). *Bayesian Design and Analysis of Cluster Randomised Controlled Trials*. Phd thesis, Indiana University.

[Yu et al., 2021] Yu, L. M., Bafadhel, M., Dorward, J., Hayward, G., Saville, B. R., Gbinigie, O., van Hecke, O., Ogburn, E., Evans, P. H., Thomas, N. P., Patel, M. G., Richards, D., Berry, N., Detry, M. A., Saunders, C., Fitzgerald, M., Harris, V., Shanyinde, M., de Lusignan, S., Andersson, M. I., Barnes, P. J., Russell, R. E., Nicolau, D. V., Ramakrishnan, S., Hobbs, F. D., Butler, C. C., Thomas, N. P., Saunders, C. T., Russell, R. E., and Hobbs, F. R. (2021). Inhaled budesonide for COVID-19 in people at high risk of complications in the community in the UK (PRINCIPLE): a randomised, controlled, open-label, adaptive platform trial. *The Lancet*, 398(10303):843–855.

[Zeger and Karim, 1991] Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):79–86.

[Zheng and Wason, 2022] Zheng, H. and Wason, J. M. (2022). Borrowing of information across patient subgroups in a basket trial based on distributional discrepancy. *Biostatistics*, 23(1):120–135.

Appendix A

Systematic Review Protocol

The use of Bayesian Statistics in the design and analysis of cluster randomised controlled trials and their methodological and reporting quality: a protocol for an international methodological review.

B.G Jones

1. Background

In a Cluster Randomised Controlled Trial (CRCT), randomisation units are in the form of groups or "clusters" as opposed to randomising individuals as is typical in traditional Randomised Controlled Trials (RCTs). Examples of clusters include schools, communities or GP practices. Randomisation of clusters is conducted for a number of reasons: (i) when the intervention is to be delivered at the cluster level (e.g. to a whole school/class within a school); (ii) when there is a risk of contamination, either between subjects/participants or health professionals or (iii) when there is a clear administrative, logistic or cost-based rationale[1].

Cluster randomisation has methodological implications that go beyond merely the randomisation procedure itself. Measurements on individuals within the same cluster are likely to be more correlated to one another than measurements on individuals from different clusters. This correlation creates an additional level of complexity which must be accounted for in both the study design and sample size calculation, and the statistical analysis of the results. Failure to do so can result in an underpowered study and ultimately spurious conclusions of efficacy or effectiveness of the intervention or treatment under investigation.

CRCTs are a relatively novel study design, but the methodology is becoming increasingly well established in the literature. Prior to the 1980s, there was only sparse use of CRCTs[2], but they have become increasingly more common in the last 30 years, from just seven reported in 1990, to over 120 in 2008[3,4]. With such a rapid increase in the use of the CRCT design, there have been some attempts to develop new Bayesian methodology for the design and analysis of such trials. This includes relatively simple Bayesian Hierarchical modelling to handle the clustered nature of the data, through to more novel approaches to design and sample size calculation such as that developed by Turner et al.[5,6]. However, a brief scoping review suggests that the uptake of Bayesian methodology in CRCTs is limited.

Furthermore, with the increased use of CRCTs, the need for consistent, high quality reporting is crucial. In response to this need, the CONSORT extension to Cluster Randomised Trials was first published in 2004[7] and updated in 2012[8]. A recent review of the methodological quality of sample size calculations in a sample of 300 Cluster Randomised Trials published between 2000-2008 found that only 166 presented a sample size calculation, of which only 102 accounted appropriately for clustering[9]. A separate recent review of the same sample of CRCTs examined the impact of the 2004 CONSORT extension on more general methodological quality and concluded that adherence to reporting guidelines and quality remains low[10]. Similar reviews of CRCT reporting quality have been conducted and produced similar conclusions[11,12]. However, to our knowledge, none have focussed specifically on CRCTs which used Bayesian techniques.

As such, this review aims to:

(i) Quantify and explore the use of Bayesian methodology in the design or analysis of CRCTs.

(ii) Appraise the quality of reporting of CRCTs conducted in a Bayesian framework against the current relevant CONSORT guidelines and identify whether the reporting quality differs from those using a frequentist approach. The impact of the introduction of the CONSORT guidelines in 2004 and 2012 on reporting quality will also be appraised.

Methods

2.1. Inclusion Criteria

We will seek to identify all reported/published CRCTs in which Bayesian methodology was used, or as a minimum considered. We will include references that discuss Bayesian methodology, even in cases where Bayesian approaches were not actually implemented, whilst acknowledging that, in the majority of cases, only methodology that has been used will be described or discussed.

We will not restrict our search on the basis of publication date, location, intervention type or population in any way, provided the relevant paper was published in the English language.

In order to be included in the review, it must be clear that randomisation occurred at a group level, as per the definition of a CRCT. If this is not the case, the study will be excluded, but will not be excluded for any other reporting or methodological shortcomings.

We aim to include not only publications reporting primary results of efficacy or effectiveness, but also protocol papers, papers reporting secondary analyses and papers reporting the results of feasibility/pilot studies. If, however, both a protocol and a results paper are identified from the search, a single entry will be recorded, using both sources if appropriate for data extraction. In this scenario, headline data (e.g. year of publication, country) will be taken from the results paper as opposed to the protocol. We will also seek to obtain additional detail from published protocols or monographs if deemed useful or necessary for data collection. Appropriate systematic reviews and meta-analyses will also be considered, with the view to identifying additional primary studies.

We will exclude papers that only report on cost-effectiveness, as well as trials that used a stepped-wedge design.

2.2 Data Collection

We will collect data on a selection of the quality reporting standards as outlined in the 2012 CONSORT extension to Cluster Randomised Trials[13], which will allow us to measure the reporting and methodological quality of each of the included studies.

In addition, we will collect information on journal endorsement of the CONSORT statement and reported statistician involvement in the design and/or analysis of the study. In the same way as defined by Diaz-Ordaz et al.[11], we will classify a journal as a strong endorser if the words “required”, ”must”, ”should” or “strongly recommended” are used in their author instructions, a medium endorser if words “encouraged”, ”recommended”, “advised” or “please” were used, and a low endorser if “may wish to consider” or ”see CONSORT” is used. We include a fourth category, “none”, if the journal includes no mention of the CONSORT statement in its guidelines to authors. We will discern whether a statistician was involved in the trial via previously used criteria[14] – a statistician will be deemed to be involved if at least one co-author belonged to a department of epidemiology, clinical epidemiology and/or Medical Statistics/biostatistics. We will also seek to identify Clinical Trial Unit involvement where possible by examination of the author list and note whether the statistician involved was associated with a Clinical Trials Unit.

We will collect descriptive information on each study to be included in the review, including location(s) of the study, the location of the institution and the institution name within which the first author belongs (UK, US/Canada, Europe, Australia/New Zealand, Africa, Asia, Other), year of publication, sample size, number of clusters, type of primary outcome (binary, categorical, continuous), and whether the publication reported the intraclass correlation coefficient (ICC) for the primary outcome and any secondary outcomes.

We will quantify the use of Bayesian methodology in the design and analysis of cluster randomised trials. In particular, we will note whether methods of Bayesian sample size calculation and analysis have been used or even simply discussed. We will also include a description of the Bayesian methodology used, with collection of further details, particularly if the reported methods are non-standard.

In any cases where information has been omitted but indicated as present elsewhere, such as details of a sample size calculation referred to in a published protocol, we will seek to obtain this information and include it in our data collection and analyses without penalty.

A full specification of the data to be collected are included in the supplementary material.

2.3 Search Strategy

Taljaard et al.[15] presented a search strategy to identify cluster randomised controlled trials. We will adapt this strategy to include only publications with the word "Bayes" (with appropriate truncation) included in the title, abstract or text (Table 1).

Table 1: Search Strategy used to search Medline and Embase within Ovid

#	Search
Existing published strategy for randomized controlled trials	
1	(article OR randomized controlled trials).pt.
2	Animals/
3	Humans/
4	#2 NOT (2 AND 3)
5	#1 NOT #4
Cluster-design related terms	
6	(cluster$ adj2 randomi$).tw.
7	((communit$ adj2 intervention$) or (communit$ adj2 randomi$)).tw.
8	group$ randomi$.tw.
9	#6 OR #7 OR #8
10	intervention?.tw.
11	Cluster Analysis/
12	Health Promotion/
13	Program Evaluation/
14	Health Education/
15	#10 OR #11 OR #12 OR #13 OR #14
16	#9 OR #15
17	bayes$.af.
18	#16 AND #17
Final Search	
19	#18 AND #5
20	limit #19 to (randomized controlled trial)

pt. represents publication type; / represents MeSH search; $ allows for truncation of words; adj allows for adjacency between search words; tw represents text words in abstract and/or title; af represents all fields; ? is a wildcard which retrieves one or 0 characters

We will use Ovid to conduct the search outlined in Table 1 on both MedLine and Embase databases. There are some minor differences in the search strategies for the two databases. Embase does not include “randomised controlled trials” as a publication type and so it is necessary to edit the search strategy to accommodate this, by restricting search term #1 in Table 1 to “article” or “randomized controlled trial” and limiting the final search to randomized controlled trials (#20). We will also search the Cochrane Library Central Register of Controlled Trials (CENTRAL) using an appropriate adaptation of the strategy outlined in Table 1.

2.4 Analysis

We will present summary statistics for the all data collected on reporting quality, analysis and sample size calculation methodology (including Bayesian) and demographic data.

As one of the research questions asks whether or not the publication of the CONSORT guidelines for cluster trials in 2004 and 2012 has improved reporting quality, we will define three time periods to be used for comparison: (i) pre 2004; (ii) 2005 – 2012 and (iii) 2013 – 2018. Summary statistics will be presented for each of these periods and overall.

Data analysis will be conducted using R[16].

2.5 Quality Control

We will conduct the initial search using the search strategy defined in section 2. Following export of references to Mendeley[17] and removal of duplicate studies, the process of selecting the final references for use in the review will be split in to three stages:

1) An initial sift using the titles and abstracts alone will be conducted twice, by independent reviewers, facilitated by the software package Rayyan[18]. Rayyan allows users to categorise each reference as "include" or "exclude" and subsequently collaborate with colleagues to discuss any disagreements flagged by the software during the process. Any disagreements will be resolved through discussion or, if an agreement cannot be reached, a final decision will be made by a third individual.
2) Full publications will be obtained for all remaining references. A final inclusion/exclusion decision will be made on the basis of the full text. If capacity allows this stage will again be conducted twice, independently. If capacity is insufficient, a minimum of 50% of the remaining references will be reviewed twice. As before, disagreements will be resolved in the first instance through discussion, and failing that through a third reviewer.
3) Data extraction will be conducted by means of populating an excel spreadsheet on the basis of the data to be collected, as outlined in the supplementary material. We will strive to conduct the data collection twice, independently, but if there is insufficient capacity, at least 50% of the data collection will be undertaken twice. Disagreement will be resolved as in stages 1) and 2).

References

1. Eldridge SM, Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research.* John Wiley & Sons; 2012.

2. Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Med Res Methodol*. 2004;4(1):21. doi:10.1186/1471-2288-4-21.

3. Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *Int J Epidemiol*. 1990;19(4):795-800. doi:10.1093/ije/19.4.795.

4. Moberg J, Kramer M. A brief history of the cluster randomised trial design. *J R Soc Med*. 2015;108(5):192-198. doi:10.1177/0141076815582303.

5. Turner RM, Prevost AT, Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med*. 2004;23(8):1195-1214. doi:10.1002/sim.1721.

6. Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clin Trials*. 2005;2(2):108-118. doi:10.1191/1740774505cn072oa.

7. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *Bmj*. 2004;328(7441):702-708. doi:10.1136/bmj.328.7441.702.

8. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: Extension to cluster randomised trials. *BMJ*. 2012;345(7881). doi:10.1136/bmj.e5661.

9. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: A review. *J Clin Epidemiol*. 2015;68(6):716-723. doi:10.1016/j.jclinepi.2014.10.006.

10. Ivers NM, Taljaard M, Dixon S, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *BMJ*. 2011;343:d5886. doi:10.1136/BMJ.D5886.

11. Diaz-Ordaz K, Froud R, Sheehan B, Eldridge S. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. *BMC Med Res Methodol*. 2013;13(1):127. doi:10.1186/1471-2288-13-127.

12. Tokolahi E, Hocking C, Kersten P, Vandal AC. Quality and Reporting of Cluster Randomized Controlled Trials Evaluating Occupational Therapy Interventions: A Systematic Review. *OTJR Occup Particip Heal*. 2016;36(1):1539449215618625. doi:10.1177/1539449215618625.

13. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ*. 2012;345(sep04 1):e5661-e5661. doi:10.1136/bmj.e5661.

14. Delgado-Rodriguez M, Ruiz-Canela M, De Irala-Estevez J, Llorca J, Martinez-Gonzalez A. Participation of epidemiologists and/or biostatisticians and methodological quality of published controlled clinical trials. *J Epidemiol Community Health*. 2001;55(8):569-572. doi:10.1136/JECH.55.8.569.

15. Taljaard M, McGowan J, Grimshaw JM, et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: Low precision will improve with adherence to reporting standards. *BMC Med Res Methodol*. 2010;10. doi:10.1186/1471-2288-10-15.

16. R Core Team. R: A Language and Environment for Statistical Computing. 2017. https://www.r-project.org/.

17. Elsevier. Mendeley. https://www.mendeley.com/.

18. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210. doi:10.1186/s13643-016-0384-4.

Appendix B

Systematic Review Data Collection Form - Primary Results Papers

The use of Bayesian Statistics in the design and analysis of cluster randomised controlled trials and their methodological and reporting quality: an international methodological review.

Data Collection Form – Primary Results Papers

Section A: Demographics

1) Title of Publication

……………………………………………………………………………………………………
……………………………………………………………………………………………………
……………………………………………………………………………………………………

2) Year of Publication

………….

3) Name of First Author

……………………………………………………

4) Specify the country in which the first author is based

□ UK
□ US/Canada
□ Europe excl. UK
□ Australia/New Zealand
□ Africa
□ Asia
□ Other

5) Country/ies in which the study was conducted

□ UK
□ US/Canada
□ Europe excl. UK
□ Australia/New Zealand
□ Africa
□ Asia
□ Other

6) Total Target Sample Size

……………………………………………………

7) Target Number of Clusters

……………………………………………………

Section A: Demographic Information

8) Total Recruited Sample Size

……………………………………………………

9) Number of Clusters Recruited

……………………………………………………

10) Specify nature of cluster (e.g. school, village)

……………………………………………………

11) Primary outcome type

□ Binary □ Categorical □ Continuous

□ Time-to-Event □ Ordinal □ Count

Section B: CONSORT Reporting Quality

12) Identification as cluster randomised trial in the title?

□ Yes □ No

13) Sample Size Calculation

a) Is there a description of the method of sample size calculation?

□ Yes □ No

If yes:
b) Was clustering accounted for in the sample size calculation?

□ Yes □ No

c) Is there a specification of the number of clusters required?

□ Yes □ No

d) Is there a specification of assumed cluster size?

□ Yes □ No

e) Is there a specification of whether equal or unequal cluster sizes are assumed?

□ Yes □ No

f) Was the variability in cluster size accounted for?

□ Yes □ No

g) Is there an estimation of a coefficient of intracluster correlation (usually ICC)?

□ Yes □ No

h) If Yes above, is there also an indication of its uncertainty?

□ Yes □ No

i) If Yes in h), was this uncertainty accounted for?

□ Yes □ No

14) Are there details of how clustering was accounted for in the analysis?

□ Yes □ No □ Unclear

15) Are there details of the number of clusters randomised?

□ Yes □ No

16) Are there details of the number of clusters receiving intended treatment?

□ Yes □ No

17) Are there details of the number of clusters analysed for the primary outcome at the primary endpoint?

□ Yes □ No

18) Are there details of cluster-level losses and exclusions?

□ Yes □ No

19) Are there details of individual-level losses and exclusions?

□ Yes □ No

20) Are baseline characteristics at the individual level provided?

□ Yes □ No

21) Are baseline characteristics at the cluster level provided?

□ Yes □ No

22) Are coefficients of intracluster correlation provided for each primary outcome?

□ All □ Some □ None

23) Are coefficients of intracluster correlation provided for each secondary outcome?

□ All □ Some □ None

24) Have p-values been calculated for baseline comparisons?

□ Yes □ No

25) If yes above, was clustering accounted for in the comparisons?

□ Yes □ No □ Unclear

Section C: Technical Information

26) Were Bayesian methods used in the design or sample size calculation?

□ Yes □ No

27) If yes, please describe the design/sample size calculation method used:

……………………………………………………………………………………………
……………………………………………………………………………………………
……………………………………………………………………………………………

28) If no, were Bayesian methods for design/sample size calculation discussed?

□ Yes □ No

29) Were Bayesian methods used in the analysis?

□ Yes □ No

If no, proceed to 31), If yes:
30) Please describe the analysis method used

……………………………………………………………………………………………
……………………………………………………………………………………………
……………………………………………………………………………………………

31) Were the priors

□ Informative □ Weakly Informative

□ Non-informative □ Unspecified

32) If no, were Bayesian methods for analysis discussed?

□ Yes □ No

Section D: Additional Information

33) How strong was the journal’s endorsement of the CONSORT guidelines? Please classify strength of endorsement as high if the words “required”, ”must”, ”should” or “strongly recommended” were used in their author instructions, a medium endorser if words “encouraged”, ”recommended”, “advised” or “please” were used, and a low endorser if “may wish to consider” or ”see CONSORT” were used.

□ None □ Low □ Medium

□ High

34) Was a Statistician involved in the study? Please select yes if there is a clearly designated statistician, or if at least one of the co-authors belonged to a department of epidemiology or biostatistics (online searching may be required if the paper does not contain sufficient detail).

□ Yes □ No

35) If yes, was the Statistician associated with any of the following according to the detail in the paper or online?

□ Clinical Trials Unit □ Academic Statistical Department □ Commercial Pharmaceutical Company

□ Clinical Research Organisation □ Other

36) If other, please specify

……………………………………………………………………………………………
……………………………………………………………………………………………

37) Other than the statistician, was a Clinical Trials Unit (CTU) involved in the Study? If there was no mention of CTU involvement in the study, and no co-authors were listed with CTU affiliations, please select no.

□ Yes □ No

Appendix C

Systematic Review Data Collection Form - Secondary Results Papers

The use of Bayesian Statistics in the design and analysis of cluster randomised controlled trials and their methodological and reporting quality: an international methodological review.

Data Collection Form – Secondary Results Papers

Section A: Demographics

1) Title of Publication

……………………………………………………………………………………………………
……………………………………………………………………………………………………
……………………………………………………………………………………………………

2) Year of Publication

………….

3) Name of First Author

…………………………………………………

4) Specify the country in which the first author is based

□ UK □ US/Canada □ Europe excl. UK

□ Australia/New Zealand □ Africa □ Asia

□ Other

5) Country/ies in which the study was conducted

□ UK □ US/Canada □ Europe excl. UK

□ Australia/New Zealand □ Africa □ Asia

□ Other

6) Total Recruited Sample Size

…………………………………………………

7) Number of Clusters Recruited

…………………………………………………

Section A: Demographic Information

8) Specify nature of cluster (e.g. school, village)

…………………………………………………………

9) Primary outcome type

□ Binary   □ Categorical   □ Continuous

□ Time-to-Event   □ Ordinal   □ Count

Section B: Technical Information

10) Were Bayesian methods used in the design or sample size calculation?

□ Yes □ No

11) If yes, please describe the design/sample size calculation method used:

…………………………………………………………………………………………………
…………………………………………………………………………………………………
…………………………………………………………………………………………………

12) If no, were Bayesian methods for design/sample size calculation discussed?

□ Yes □ No

13) Were Bayesian methods used in the analysis?

□ Yes □ No

If no, proceed to 16), If yes:
14) Please describe the analysis method used

…………………………………………………………………………………………………
…………………………………………………………………………………………………
…………………………………………………………………………………………………

15) Were the priors

□ Informative □ Weakly Informative

□ Non-informative □ Unspecified

16) If no, were Bayesian methods for analysis discussed?

□ Yes □ No

Section C: Additional Information

17) Was a Statistician involved in the study? Please select yes if there is a clearly designated statistician, or if at least one of the co-authors belonged to a department of epidemiology or biostatistics (online searching may be required if the paper does not contain sufficient detail).

□ Yes □ No

18) If yes, was the Statistician associated with any of the following according to the detail in the paper or online?

□ Clinical Trials Unit

□ Academic Statistical Department

□ Commercial Pharmaceutical Company

□ Clinical Research Organisation

□ Other

19) If other, please specify

……………………………………………………………………………………………
……………………………………………………………………………………………

Appendix D

A Supplementary Simulation Study

A second simulation study was undertaken for the purposes of examining the sensitivity of the discounting factor, a_0, to larger differences between pilot and definitive trial data sets, and more extreme values of the ICC than are typically observed in CRCTs. In this study only the NPP approach is used for analysis, and only data on the median value of a_0 at each iteration is captured and presented. Furthermore, only the parameters of the data generating mechanism for the pilot data are varied, with the definitive trial data being simulated with a treatment effect and individual level standard deviation of 1, an ICC of 0.05, and a total of 150 clusters, with 15 participants per cluster. For the pilot data, a total of 24 clusters are simulated, with 15 participants in each cluster. The treatment effect and the ICC for the pilot data are varied across scenarios according to Table D.1. A total of 1100 iterations were run for each scenario, with model formulation and posterior sampling undertaken for the NPP as outlined in §3.6.2.

Table D.1: A table of treatment effects and ICCs for simulation of pilot trial data

Scenario	Treatment Effect	ICC
1.1	1	0.05
1.2	1	0.25
1.3	1	0.5
2.1	0.75	0.05
2.2	0.75	0.25
2.3	0.75	0.5
3.1	0.5	0.05
3.2	0.5	0.25
3.3	0.5	0.5
4.1	1.25	0.05
4.2	1.25	0.25
4.3	1.25	0.5
5.1	1.5	0.05
5.2	1.5	0.25
5.3	1.5	0.5

Figure D.1 shows the posterior density of the median a_0 for each scenario outlined in Table D.1, and demonstrates that the value of a_0 is sensitive to differences between datasets, both in terms of the treatment effect and the ICC, and appear to discount more substantially as these differences grow larger. When the ICC in the pilot data deviates from that observed in the definitive data up to a value of 0.5, it can be seen that the posterior of a_0 becomes very close to zero. Similar results can be observed as the treatment effect deviates further from that observed in the definitive trial data.

Figure D.2 shows the number of iterations (out of a total of 1100 in each scenario) in which at least one divergent transition was observed for each of the scenarios. It can clearly be seen that, for scenarios in which the data generating mechanisms between the two datasets are similar, there are a large number of iterations with divergent transitions. This reduces rapidly as the differences between these datasets become more extreme.

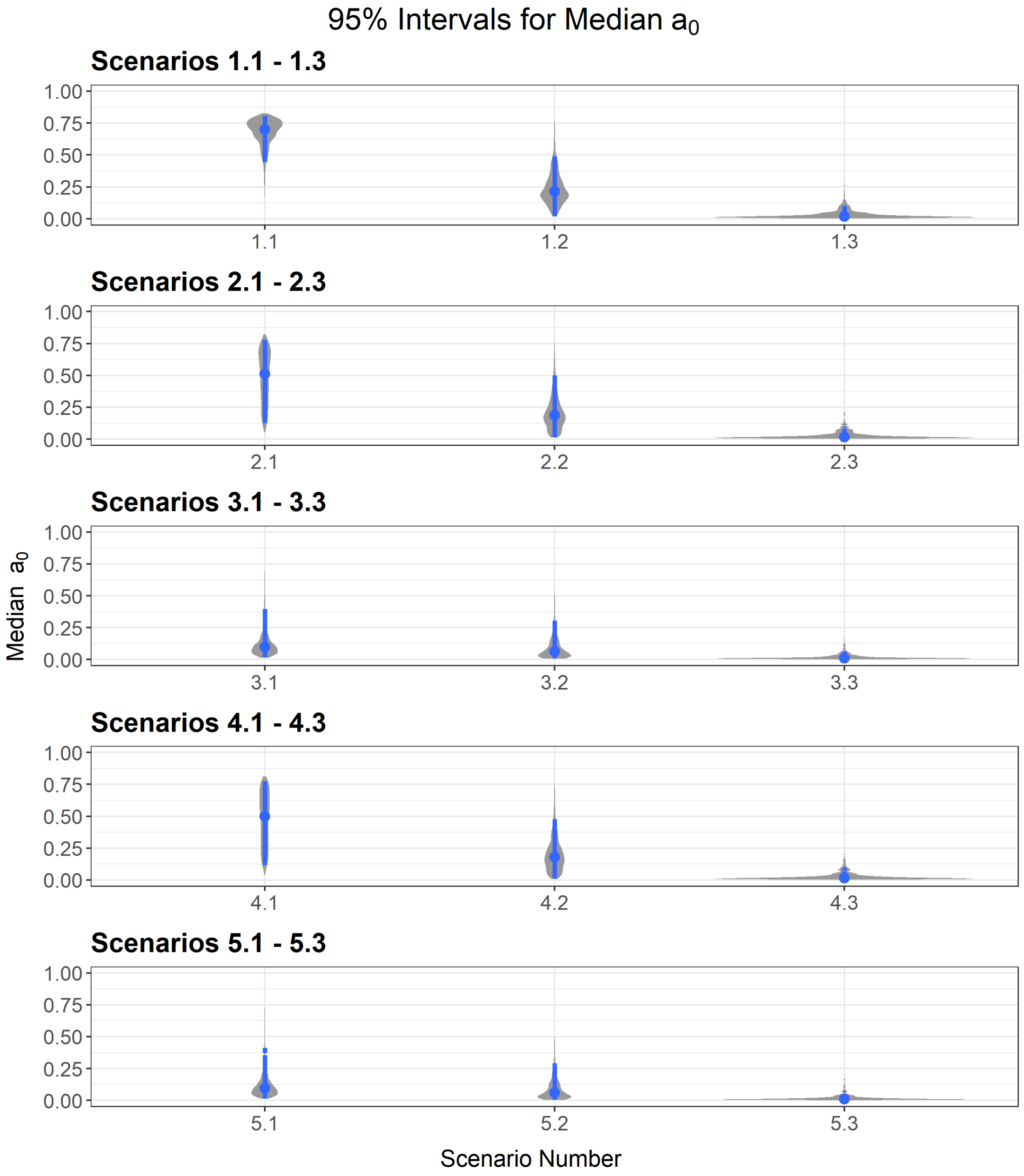


Figure D.1: Eye plots of the posterior density of the median a_0 by scenario, with 95% HPDIs.

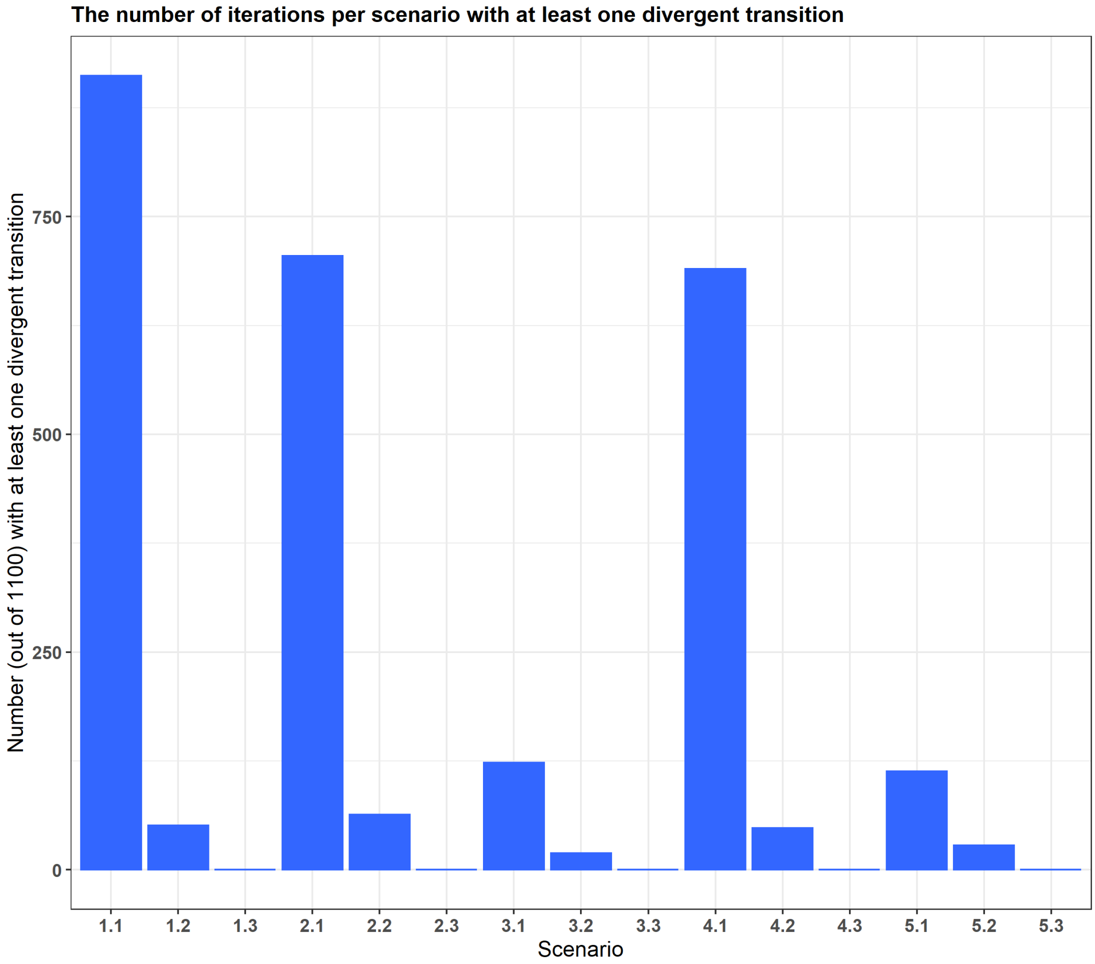


Figure D.2: Bar Graph of the the number of iterations (out of 1100) per scenario with at least one divergent transition.

Appendix E

Full Results Table for Simulation Study in Chapter 3

Table E.1: Simulation Study Results

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
1.1.1	DB	1079					0.0039	0.0041	0.064	87.3	95.7	0.26	0.0014	0.000093	0.0095	98.9	0.036
	DF						0.0037	0.0041	0.064	88.0	95.0	0.26	0.0015	0.00013	0.011	99.4	0.037
	PB						0.0036	0.0036	0.06	91.1	96.0	0.25	0.0012	0.000086	0.0092	98.5	0.034
	PF						0.0035	0.0036	0.06	91.3	95.8	0.24	0.0014	0.00012	0.011	99.1	0.036
	NPP		0.65	0.63	(0.19,0.98)	(0.24,0.99)	0.0037	0.0037	0.061	90.1	96	0.25	0.000033	0.000074	0.0086	99.0	0.033
1.1.2	DB	1072					0.0032	0.0047	0.068	84.8	94.2	0.27	0.00076	0.00027	0.016	92.0	0.062
	DF						0.0030	0.0047	0.068	85.6	93.8	0.27	0.00064	0.00025	0.016	93.1	0.061
	PB						0.0032	0.0043	0.065	87.7	93.7	0.25	-0.0022	0.00024	0.016	91.0	0.059
	PF						0.0031	0.0043	0.065	88.1	93.3	0.25	-0.0021	0.00023	0.015	92.6	0.057
	NPP		0.66	0.63	(0.17,0.98)	(0.22,1.0)	0.0031	0.0044	0.066	87.6	93.7	0.26	-0.0028	0.00025	0.016	89.9	0.059
1.1.3	DB	1083					0.001	0.0048	0.069	83.2	94.9	0.28	0.00042	0.00031	0.018	94.2	0.069
	DF						0.00093	0.0048	0.069	83.4	94.9	0.27	-0.00059	0.0003	0.017	94.5	0.068
	PB						0.00011	0.0044	0.066	84.8	94.6	0.26	-0.0040	0.00031	0.017	93.4	0.066
	PF						0.000092	0.0044	0.066	85.5	94.6	0.26	-0.0048	0.00031	0.017	92.5	0.065
	NPP		0.65	0.62	(0.15,0.98)	(0.2,1.0)	0.00029	0.0044	0.067	84.7	94.7	0.27	-0.0038	0.00031	0.017	93.6	0.066
1.1.4	DB	1082					0.0038	0.0041	0.064	87.2	95.7	0.26	0.0015	0.000094	0.0096	98.8	0.036
	DF						0.0037	0.0041	0.064	87.7	95.0	0.26	0.0016	0.00013	0.011	99.3	0.037
	PB						0.0036	0.0038	0.062	89.3	96.0	0.25	0.0044	0.00014	0.011	96.8	0.039
	PF						0.0035	0.0038	0.062	89.8	95.5	0.25	0.0050	0.00017	0.012	97.9	0.04
	NPP		0.58	0.58	(0.16,0.96)	(0.19,0.96)	0.0036	0.0038	0.061	89.9	95.7	0.25	0.0010	0.000086	0.0092	98.5	0.034
1.1.5	DB	1078					0.0035	0.0048	0.069	84.9	93.9	0.27	0.00069	0.00027	0.016	91.9	0.062
	DF						0.0033	0.0048	0.069	85.7	93.5	0.27	0.00057	0.00025	0.016	93.0	0.061
	PB						0.0034	0.0044	0.066	87.4	93.6	0.26	0.00066	0.00025	0.016	92.9	0.06
	PF						0.0033	0.0044	0.066	87.1	93.5	0.26	0.00060	0.00023	0.015	93.9	0.059
	NPP		0.63	0.61	(0.16,0.98)	(0.21,0.99)	0.0033	0.0045	0.067	86.7	93.1	0.26	-0.0014	0.00025	0.016	90.8	0.06

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
1.1.6	DB	1082					0.00086	0.0048	0.069	83.1	94.9	0.28	0.00049	0.00031	0.018	94.2	0.069
	DF						0.00079	0.0048	0.069	83.3	94.9	0.27	-0.00053	0.00031	0.017	94.5	0.068
	PB						-0.00013	0.0045	0.067	84.8	94.6	0.27	-0.0022	0.00030	0.017	94.4	0.067
	PF						0.00010	0.0045	0.067	85.1	94.7	0.26	-0.0030	0.00030	0.017	93.2	0.066
	NPP		0.65	0.63	(0.16,0.98)	(0.21,1.0)	0.00012	0.0045	0.067	84.7	94.7	0.27	-0.0027	0.00030	0.017	94.4	0.067
1.1.7	DB	1083					0.0037	0.0041	0.064	87.3	95.8	0.26	0.0014	0.000093	0.0096	98.9	0.036
	DF						0.0036	0.0041	0.064	87.9	95.0	0.26	0.0015	0.00013	0.011	99.3	0.037
	PB						0.0037	0.0041	0.064	88.7	95.8	0.26	0.00912	0.00026	0.013	92.5	0.045
	PF						0.0036	0.0041	0.064	88.7	95.7	0.26	0.010	0.00029	0.014	93.4	0.045
	NPP		0.49	0.5	(0.13,0.91)	(0.14,0.9)	0.0036	0.0038	0.062	88.8	95.8	0.26	0.0017	0.000097	0.0097	98.9	0.036
1.1.8	DB	1082					0.0036	0.0048	0.069	84.8	93.9	0.27	0.00064	0.00027	0.016	92.1	0.062
	DF						0.0034	0.0048	0.069	85.7	93.5	0.27	0.00052	0.00025	0.016	93.2	0.061
	PB						0.0037	0.0046	0.067	85.7	93.8	0.26	0.0045	0.00028	0.016	92.5	0.062
	PF						0.0035	0.0046	0.067	86.3	93.5	0.26	0.0044	0.00026	0.016	94.6	0.06
	NPP		0.57	0.57	(0.14,0.96)	(0.17,0.96)	0.0035	0.0045	0.067	86.4	93.6	0.26	0.00012	0.00026	0.016	91.8	0.061
1.1.9	DB	1086					0.00087	0.0048	0.069	82.9	95.0	0.28	0.00044	0.00032	0.018	94.1	0.069
	DF						0.00080	0.0048	0.069	83.1	95.0	0.27	-0.00058	0.00031	0.018	94.5	0.068
	PB						-0.00019	0.0046	0.068	83.5	94.9	0.27	0.00025	0.0003	0.017	94.2	0.067
	PF						-0.00012	0.0046	0.068	84.6	94.8	0.27	-0.00059	0.00030	0.017	94.1	0.066
	NPP		0.64	0.62	(0.15,0.98)	(0.20,0.99)	0.00011	0.0046	0.068	84.2	94.9	0.27	-0.0013	0.00030	0.017	94.4	0.067
1.2.1	DB	1090					0.003	0.014	0.12	85.4	97.7	0.54	0.0095	0.00042	0.018	99.4	0.08
	DF						0.0028	0.014	0.12	89.9	95.8	0.5	0.0058	0.00044	0.020	99.9	0.068
	PB						-0.056	0.013	0.10	87.3	94.3	0.45	0.0095	0.00037	0.017	98.8	0.068
	PF						-0.056	0.013	0.10	89.2	92.3	0.43	0.0079	0.00041	0.019	99.5	0.063
	NPP		0.59	0.58	(0.15,0.97)	(0.19,0.98)	-0.035	0.012	0.1	89.3	96.7	0.48	0.0049	0.00021	0.014	99.8	0.063

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
1.2.2	DB	1068					0.013	0.018	0.13	84.9	94.9	0.53	0.00071	0.00099	0.031	93.0	0.12
	DF						0.013	0.018	0.13	87.5	93.8	0.51	-0.00027	0.00090	0.03	91.1	0.10
	PB						-0.036	0.014	0.11	88.1	93.4	0.45	-0.0049	0.00076	0.027	90.2	0.096
	PF						-0.036	0.014	0.11	89.7	92.6	0.44	-0.0047	0.00070	0.026	90.0	0.091
	NPP		0.6	0.59	(0.15,0.97)	(0.19,0.99)	-0.02	0.015	0.12	88.1	94.7	0.47	-0.0084	0.00078	0.027	89.7	0.095
1.2.3	DB	1080					-0.0017	0.019	0.14	79.5	95.4	0.56	0.0033	0.0014	0.037	92.6	0.14
	DF						-0.0017	0.019	0.14	82.0	94.6	0.53	0.00029	0.0013	0.035	92.1	0.13
	PB						-0.037	0.016	0.12	84.4	94.0	0.48	-0.0094	0.0011	0.033	90.1	0.12
	PF						-0.037	0.016	0.12	85.6	93.3	0.47	-0.011	0.0011	0.031	87.4	0.12
	NPP		0.6	0.59	(0.13,0.98)	(0.17,0.99)	-0.028	0.016	0.12	84.2	94.6	0.49	-0.0097	0.0012	0.033	89.3	0.12
1.2.4	DB	1088					0.0037	0.014	0.12	85.5	97.7	0.54	0.0097	0.00044	0.019	99.3	0.08
	DF						0.0036	0.014	0.12	89.9	95.8	0.50	0.0060	0.00045	0.02	99.9	0.068
	PB						-0.056	0.015	0.11	82.4	93.7	0.48	0.019	0.00087	0.023	94.5	0.084
	PF						-0.057	0.015	0.11	84.8	93.0	0.45	0.018	0.00087	0.023	97.2	0.077
	NPP		0.54	0.55	(0.14,0.96)	(0.16,0.96)	-0.033	0.012	0.11	87.8	96.2	0.49	0.0076	0.00031	0.016	99.4	0.07
1.2.5	DB	1072					0.015	0.018	0.13	85.1	95.0	0.53	0.00074	0.00098	0.031	93.1	0.12
	DF						0.014	0.018	0.13	87.8	93.8	0.51	-0.00026	0.00089	0.03	91.2	0.1
	PB						-0.036	0.015	0.12	84.0	93.3	0.47	0.0037	0.00088	0.029	92.7	0.11
	PF						-0.036	0.015	0.12	85.2	92.4	0.46	0.0035	0.00080	0.028	93.1	0.099
	NPP		0.58	0.57	(0.14,0.97)	(0.17,0.98)	-0.018	0.015	0.12	86.5	94.4	0.48	-0.0044	0.00081	0.028	90.9	0.10
1.2.6	DB	1076					-0.0015	0.019	0.14	79.6	95.4	0.56	0.0034	0.0014	0.037	92.7	0.14
	DF						-0.0016	0.019	0.14	82.2	94.8	0.53	0.00039	0.0013	0.035	92.2	0.13
	PB						-0.036	0.016	0.12	82.2	94.4	0.49	-0.0029	0.0011	0.033	92.8	0.12
	PF						-0.036	0.016	0.12	83.9	93.2	0.48	-0.0051	0.001	0.032	91.6	0.12
	NPP		0.61	0.59	(0.13,0.98)	(0.18,0.99)	-0.027	0.016	0.12	82.9	94.6	0.50	-0.0058	0.0012	0.034	91.3	0.13

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
1.2.7	DB	1090					0.0034	0.014	0.12	85.4	97.7	0.54	0.0097	0.00044	0.019	99.3	0.08
	DF						0.0033	0.014	0.12	89.9	95.8	0.5	0.0060	0.00045	0.02	99.9	0.068
	PB						-0.058	0.018	0.12	76.1	94.2	0.51	0.033	0.002	0.03	84.9	0.11
	PF						-0.058	0.018	0.12	78.4	92.9	0.49	0.032	0.0019	0.029	90.9	0.096
	NPP		0.48	0.49	(0.11,0.93)	(0.13,0.92)	-0.03	0.013	0.11	84.5	96.7	0.5	0.011	0.00045	0.018	98.8	0.079
1.2.8	DB	1079					0.013	0.018	0.13	84.7	94.8	0.53	0.00091	0.00098	0.031	93.2	0.12
	DF						0.013	0.018	0.13	87.4	93.8	0.51	-0.000066	0.00089	0.03	91.4	0.11
	PB						-0.038	0.017	0.12	80.0	93.2	0.5	0.016	0.0013	0.033	91.2	0.12
	PF						-0.038	0.017	0.13	82	92.6	0.48	0.015	0.0012	0.031	94.2	0.11
	NPP		0.53	0.54	(0.12,0.95)	(0.15,0.96)	-0.018	0.016	0.12	84.1	94.9	0.49	0.00042	0.00090	0.03	93.1	0.11
1.2.9	DB	1073					-0.0022	0.019	0.14	79.6	95.4	0.56	0.0034	0.0014	0.037	92.6	0.14
	DF						-0.0022	0.019	0.14	82.0	94.8	0.53	0.00040	0.0013	0.035	92.1	0.13
	PB						-0.037	0.017	0.13	80.1	94.7	0.51	0.0057	0.0012	0.034	93.6	0.13
	PF						-0.036	0.017	0.13	81.5	93.1	0.49	0.0031	0.0011	0.033	93.4	0.12
	NPP		0.6	0.58	(0.13,0.97)	(0.17,0.98)	-0.027	0.017	0.13	81.6	94.7	0.51	-0.0010	0.0012	0.034	92.5	0.13
1.3.1	DB	1080					0.0036	0.0041	0.064	87.1	95.8	0.26	0.0015	0.000094	0.0096	98.8	0.036
	DF						0.0035	0.0041	0.064	87.7	95.1	0.26	0.0015	0.00013	0.011	99.3	0.037
	PB						0.023	0.0042	0.06	95.1	94.4	0.25	0.0025	0.00011	0.01	98.1	0.036
	PF						0.023	0.0042	0.06	95.4	93.9	0.25	0.0029	0.00014	0.011	98.6	0.037
	NPP		0.59	0.59	(0.16,0.97)	(0.2,0.98)	0.015	0.0039	0.061	93.0	95.8	0.25	0.00043	0.000079	0.0089	99.0	0.033
1.3.2	DB	1075					0.0033	0.0048	0.069	84.7	93.8	0.27	0.00057	0.00027	0.016	91.9	0.062
	DF						0.0031	0.0048	0.069	85.6	93.4	0.27	0.00046	0.00025	0.016	93.0	0.061
	PB						0.018	0.0047	0.066	91.2	93.2	0.26	-0.0012	0.00024	0.016	91.3	0.059
	PF						0.017	0.0047	0.066	91.3	93.0	0.25	-0.0011	0.00023	0.015	92.7	0.058
	NPP		0.62	0.6	(0.16,0.98)	(0.2,0.99)	0.013	0.0046	0.067	89.7	93.1	0.26	-0.0024	0.00025	0.016	90.1	0.059

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
1.3.3	DB	1084					0.00071	0.0048	0.070	82.8	94.8	0.28	0.00048	0.00031	0.018	94.3	0.069
	DF						0.00064	0.0048	0.07	83.0	94.8	0.27	-0.00053	0.00030	0.017	94.6	0.068
	PB						0.010	0.0045	0.066	88.3	94.7	0.26	-0.0032	0.00030	0.017	94.4	0.066
	PF						0.01	0.0045	0.066	88.5	94.6	0.26	-0.0040	0.00030	0.017	93.2	0.065
	NPP		0.62	0.6	(0.14,0.98)	(0.19,0.99)	0.0078	0.0046	0.067	86.7	94.7	0.27	-0.0033	0.00030	0.017	94.4	0.067
1.3.4	DB	1076					0.0036	0.004	0.063	87.5	95.9	0.26	0.0015	0.000094	0.0096	98.8	0.036
	DF						0.0035	0.004	0.063	88.0	95.3	0.26	0.0015	0.00013	0.011	99.3	0.037
	PB						0.023	0.0043	0.061	94.1	94.2	0.26	0.0058	0.00017	0.012	95.5	0.04
	PF						0.023	0.0043	0.061	94.0	93.9	0.25	0.0065	0.00020	0.013	96.7	0.042
	NPP		0.53	0.54	(0.14,0.94)	(0.16,0.94)	0.014	0.0039	0.061	91.6	95.9	0.26	0.0012	0.000090	0.0094	98.3	0.035
1.3.5	DB	1072					0.0036	0.0047	0.069	84.7	93.9	0.27	0.00070	0.00027	0.016	92.0	0.062
	DF						0.0035	0.0048	0.069	85.6	93.6	0.27	0.00058	0.00025	0.016	93.1	0.061
	PB						0.018	0.0047	0.066	90.7	93.6	0.26	0.0019	0.00025	0.016	92.4	0.061
	PF						0.018	0.0047	0.066	90.7	93.3	0.26	0.0019	0.00023	0.015	94.1	0.059
	NPP		0.59	0.58	(0.14,0.97)	(0.18,0.98)	0.013	0.0046	0.067	89.4	93.8	0.26	-0.00088	0.00025	0.016	91.4	0.06
1.3.6	DB	1085					0.00042	0.0048	0.07	82.7	94.8	0.28	0.00050	0.00031	0.018	94.2	0.069
	DF						0.00034	0.0048	0.07	82.9	94.8	0.27	-0.00051	0.00031	0.017	94.6	0.068
	PB						0.0098	0.0046	0.067	87.6	94.7	0.27	-0.0014	0.00029	0.017	94.1	0.067
	PF						0.0099	0.0046	0.067	87.9	94.7	0.26	-0.0022	0.00029	0.017	93.9	0.066
	NPP		0.63	0.61	(0.14,0.98)	(0.19,0.99)	0.0074	0.0046	0.068	86.4	94.5	0.27	-0.0022	0.00030	0.017	94.2	0.067
1.3.7	DB	1080					0.0039	0.0041	0.064	87.4	95.7	0.26	0.0015	0.000095	0.0096	98.8	0.036
	DF						0.0038	0.0041	0.064	88.0	95.0	0.26	0.0016	0.00013	0.011	99.3	0.037
	PB						0.024	0.0046	0.064	92.4	94.4	0.26	0.011	0.00031	0.014	90.4	0.046
	PF						0.024	0.0046	0.064	92.4	93.9	0.26	0.012	0.00034	0.014	92.2	0.047
	NPP		0.46	0.47	(0.12,0.88)	(0.12,0.86)	0.013	0.004	0.062	90.7	95.5	0.26	0.0019	0.00010	0.0099	98.4	0.036

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
1.3.8	DB	1081					0.0035	0.0048	0.069	84.8	93.8	0.27	0.00056	0.00027	0.016	92.0	0.062
	DF						0.0033	0.0048	0.069	85.6	93.5	0.27	0.00045	0.00025	0.016	93.2	0.061
	PB						0.018	0.0049	0.067	90.0	93.4	0.26	0.0057	0.00030	0.016	92.3	0.062
	PF						0.018	0.0049	0.067	89.9	93.3	0.26	0.0055	0.00028	0.016	94.1	0.061
	NPP		0.53	0.54	(0.13,0.94)	(0.15,0.94)	0.012	0.0047	0.067	88.6	93.4	0.26	0.00032	0.00026	0.016	91.5	0.062
1.3.9	DB	1090					0.00047	0.0048	0.069	82.8	94.9	0.28	0.00043	0.00031	0.018	94.1	0.069
	DF						0.00041	0.0048	0.069	82.9	94.9	0.27	-0.00058	0.00030	0.017	94.5	0.068
	PB						0.0098	0.0047	0.068	87.1	94.9	0.27	0.0010	0.00030	0.017	93.9	0.068
	PF						0.0099	0.0047	0.068	87.4	95.0	0.27	0.00017	0.00029	0.017	94.0	0.067
	NPP		0.61	0.6	(0.14,0.97)	(0.18,0.98)	0.0072	0.0046	0.068	86.1	95.0	0.27	-0.00095	0.00030	0.017	94.3	0.068
1.4.1	DB	1093					0.0028	0.014	0.12	85.3	97.7	0.54	0.0097	0.00044	0.019	99.3	0.081
	DF						0.0027	0.014	0.12	89.8	95.8	0.5	0.0060	0.00045	0.02	99.9	0.068
	PB						0.0011	0.01	0.10	95.8	97.0	0.44	0.0065	0.00027	0.015	99.1	0.063
	PF						0.0012	0.01	0.1	96.5	96.0	0.42	0.0044	0.00031	0.017	99.9	0.058
	NPP		0.64	0.62	(0.17,0.98)	(0.22,0.99)	0.0019	0.01	0.1	94.5	97.9	0.47	0.0039	0.00018	0.013	99.8	0.06
1.4.2	DB	1075					0.013	0.018	0.13	84.7	94.8	0.53	0.00073	0.00099	0.031	93.0	0.12
	DF						0.013	0.018	0.13	87.4	93.7	0.51	-0.00026	0.00090	0.030	91.2	0.10
	PB						0.0076	0.013	0.11	94.6	94.8	0.44	-0.0086	0.00075	0.026	88.0	0.092
	PF						0.0077	0.013	0.11	95.2	94.2	0.43	-0.0084	0.00071	0.025	86.6	0.087
	NPP		0.64	0.62	(0.16,0.98)	(0.21,0.99)	0.0096	0.014	0.12	94.2	94.9	0.46	-0.010	0.00078	0.026	87.9	0.092
1.4.3	DB	1076					-0.0019	0.019	0.14	79.5	95.4	0.56	0.0033	0.0014	0.037	92.5	0.14
	DF						-0.0019	0.019	0.14	82.0	94.7	0.53	0.00028	0.0013	0.036	92.0	0.13
	PB						-0.0021	0.014	0.12	90.0	95.4	0.48	-0.012	0.0012	0.032	89.0	0.12
	PF						-0.002	0.014	0.12	91.5	94.1	0.46	-0.014	0.0011	0.031	86.2	0.11
	NPP		0.63	0.61	(0.15,0.98)	(0.19,0.99)	-0.0022	0.015	0.12	88.4	95.3	0.49	-0.011	0.0012	0.033	88.4	0.12

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
1.4.4	DB	1090					0.003	0.014	0.12	85.4	97.8	0.54	0.0096	0.00044	0.019	99.3	0.08
	DF						0.0029	0.014	0.12	89.8	95.9	0.5	0.0059	0.00044	0.02	99.9	0.068
	PB						0.000031	0.012	0.11	92.6	96.2	0.47	0.015	0.00065	0.021	96.6	0.078
	PF						0.0000078	0.012	0.11	94.1	95.3	0.44	0.014	0.00067	0.022	98.2	0.071
	NPP		0.59	0.58	(0.15,0.97)	(0.19,0.98)	0.0013	0.011	0.11	92.8	97.2	0.48	0.0066	0.00027	0.015	99.4	0.067
1.4.5	DB	1071					0.013	0.018	0.13	85.1	94.8	0.53	0.00074	0.00097	0.031	93.1	0.12
	DF						0.013	0.018	0.13	87.8	93.7	0.51	-0.00025	0.00089	0.030	91.2	0.11
	PB						0.007	0.014	0.12	92.7	94.9	0.46	-0.0000094	0.00081	0.028	91.4	0.10
	PF						0.0069	0.014	0.12	93.7	93.7	0.45	-0.00010	0.00074	0.027	91	0.096
	NPP		0.61	0.6	(0.15,0.98)	(0.2,0.99)	0.0093	0.015	0.12	92.5	95.3	0.47	-0.0060	0.00078	0.027	89.5	0.099
1.4.6	DB	1082					-0.0021	0.019	0.14	79.5	95.4	0.56	0.0032	0.0014	0.037	92.5	0.14
	DF						-0.0021	0.019	0.14	82.0	94.6	0.53	0.00020	0.0013	0.036	92.0	0.13
	PB						-0.0018	0.015	0.12	88.7	95.6	0.49	-0.0057	0.0011	0.033	91.3	0.12
	PF						-0.0016	0.015	0.12	90.1	94.6	0.47	-0.0078	0.0010	0.031	90.7	0.12
	NPP		0.63	0.61	(0.15,0.98)	(0.19,0.99)	-0.0021	0.015	0.12	87.7	95.4	0.5	-0.0076	0.0012	0.033	90.3	0.13
1.4.7	DB	1089					0.0033	0.014	0.12	85.5	97.8	0.54	0.0096	0.00043	0.018	99.3	0.08
	DF						0.0032	0.014	0.12	90.0	95.9	0.5	0.0059	0.00044	0.020	99.9	0.068
	PB						-0.0013	0.014	0.12	88.0	95.8	0.5	0.029	0.0017	0.029	89.6	0.099
	PF						-0.0013	0.014	0.12	89.8	94.9	0.48	0.028	0.0016	0.028	92.1	0.09
	NPP		0.51	0.52	(0.13,0.94)	(0.15,0.94)	0.00074	0.012	0.11	90.9	97.0	0.50	0.0098	0.00041	0.018	99.2	0.077
1.4.8	DB	1084					0.014	0.018	0.13	85.1	94.8	0.53	0.00085	0.00098	0.031	93.1	0.12
	DF						0.013	0.018	0.13	87.6	93.7	0.51	-0.00014	0.00089	0.030	91.2	0.11
	PB						0.0067	0.016	0.12	89.1	94.2	0.49	0.012	0.0012	0.032	92.4	0.11
	PF						0.0068	0.016	0.12	90.5	93.5	0.47	0.011	0.001	0.030	93.5	0.11
	NPP		0.57	0.56	(0.10,0.96)	(0.17,0.97)	0.0093	0.015	0.12	89.9	94.8	0.49	-0.00088	0.00087	0.029	92.6	0.11

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
1.4.9	DB	1078					-0.0017	0.019	0.14	79.7	95.4	0.56	0.0032	0.0014	0.037	92.5	0.14
	DF						-0.0018	0.019	0.14	82.2	94.6	0.53	0.00016	0.0013	0.036	91.9	0.13
	PB						-0.0010	0.016	0.13	87.5	94.9	0.50	0.0030	0.0011	0.033	92.9	0.13
	PF						-0.00073	0.016	0.13	88.7	94.3	0.49	0.00046	0.0010	0.032	92.3	0.12
	NPP		0.62	0.60	(0.14,0.98)	(0.19,0.99)	-0.0015	0.016	0.13	86.8	95.5	0.51	-0.0026	0.0012	0.034	92.1	0.13
2.1.1	DB	1075					-0.00016	0.0042	0.065	85.9	94.8	0.26	0.0014	0.000095	0.0096	98.7	0.036
	DF						-0.00021	0.0042	0.065	86.2	94.6	0.26	0.0016	0.00013	0.011	99.1	0.037
	PB						-0.0015	0.0035	0.059	91.3	95.2	0.24	0.00010	0.000082	0.009	98.1	0.032
	PF						-0.0014	0.0035	0.059	91.5	94.6	0.23	0.0013	0.00011	0.010	98.8	0.034
	NPP		0.63	0.62	(0.17,0.98)	(0.22,0.99)	-0.00095	0.0035	0.06	90.3	95.0	0.24	-0.0010	0.000064	0.0079	99.1	0.03
2.1.2	DB	1057					-0.0037	0.0047	0.068	81.4	94.4	0.27	-0.00012	0.00026	0.016	93.6	0.062
	DF						-0.0037	0.0047	0.068	81.6	94.1	0.26	-0.00020	0.00024	0.015	95.0	0.06
	PB						-0.0037	0.0038	0.062	87.8	94.7	0.24	-0.0052	0.00024	0.015	90.6	0.056
	PF						-0.0036	0.0038	0.062	88.7	94.3	0.24	-0.0052	0.00022	0.014	90.5	0.054
	NPP		0.65	0.62	(0.15,0.98)	(0.2,1.0)	-0.0037	0.0040	0.063	87.1	94.8	0.25	-0.0060	0.00026	0.015	88.6	0.056
2.1.3	DB	1063					-0.005	0.005	0.071	79.4	94.6	0.28	0.0018	0.00033	0.018	94.0	0.070
	DF						-0.005	0.005	0.071	79.9	94.3	0.27	0.00074	0.00032	0.018	94.9	0.069
	PB						-0.0036	0.0042	0.065	87.3	94.4	0.25	-0.0068	0.00032	0.016	91.3	0.063
	PF						-0.0038	0.0042	0.065	87.2	94.7	0.25	-0.0075	0.00033	0.016	88.8	0.063
	NPP		0.59	0.57	(0.1,0.98)	(0.14,0.98)	-0.0041	0.0044	0.066	84.8	94.9	0.26	-0.0059	0.00032	0.017	92.0	0.065
2.1.4	DB	1073					0.00014	0.0042	0.064	86.0	94.8	0.26	0.0015	0.000098	0.0098	98.7	0.036
	DF						0.000091	0.0041	0.064	86.3	94.6	0.26	0.0016	0.00013	0.011	99.1	0.037
	PB						-0.0015	0.0037	0.061	89.6	94.7	0.24	0.0070	0.00020	0.012	93.8	0.041
	PF						-0.0014	0.0037	0.061	89.6	94.6	0.24	0.0079	0.00022	0.013	94.4	0.042
	NPP		0.56	0.55	(0.15,0.95)	(0.17,0.95)	-0.00096	0.0036	0.06	89.4	95.2	0.25	0.00068	0.000085	0.0092	98.9	0.033

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
2.1.5	DB	1063					-0.0038	0.0047	0.068	81.3	94.4	0.27	-0.00015	0.00025	0.016	93.7	0.062
	DF						-0.0039	0.0047	0.068	81.8	94.2	0.26	-0.00024	0.00023	0.015	95.1	0.060
	PB						-0.0038	0.0040	0.063	86.5	94.5	0.25	0.00015	0.00022	0.015	94.2	0.058
	PF						-0.0038	0.0040	0.063	87.0	94.5	0.25	-0.000086	0.00020	0.014	95.9	0.056
	NPP		0.65	0.62	(0.15,0.98)	(0.20,0.99)	-0.0038	0.0041	0.064	85.8	94.7	0.25	-0.0033	0.00023	0.015	91.1	0.058
2.1.6	DB	1044					-0.0049	0.0050	0.070	79.6	95.0	0.28	0.0013	0.00032	0.018	94.2	0.07
	DF						-0.0049	0.0050	0.070	80.0	94.6	0.27	0.00032	0.00031	0.018	94.9	0.069
	PB						-0.0036	0.0042	0.065	85.8	94.7	0.26	-0.0037	0.00028	0.016	93.3	0.064
	PF						-0.0037	0.0042	0.065	86.3	94.8	0.26	-0.0044	0.00028	0.016	92.5	0.064
	NPP		0.67	0.64	(0.14,0.98)	(0.19,0.99)	-0.004	0.0043	0.066	84.7	95.0	0.26	-0.0043	0.00030	0.017	93.1	0.065
2.1.7	DB	1075					0.00054	0.0042	0.065	86.0	94.9	0.26	0.0016	0.000099	0.0098	98.7	0.036
	DF						0.00048	0.0042	0.065	86.3	94.7	0.26	0.0018	0.00013	0.011	99.1	0.037
	PB						-0.0016	0.0042	0.065	85.6	94.8	0.26	0.017	0.00053	0.016	79.3	0.051
	PF						-0.0016	0.0042	0.065	85.9	94.7	0.26	0.018	0.00055	0.015	81.3	0.051
	NPP		0.44	0.46	(0.11,0.85)	(0.11,0.84)	-0.00074	0.0038	0.062	88.4	95.6	0.25	0.0020	0.00011	0.010	98.4	0.036
2.1.8	DB	1067					-0.0035	0.0047	0.068	81.4	94.4	0.27	-0.00013	0.00025	0.016	93.7	0.062
	DF						-0.0036	0.0047	0.068	81.9	94.1	0.26	-0.00021	0.00023	0.015	95.1	0.06
	PB						-0.0035	0.0043	0.065	84.4	94.4	0.26	0.0074	0.00030	0.016	92.1	0.061
	PF						-0.0035	0.0043	0.065	84.6	94.1	0.25	0.0069	0.00028	0.015	94.5	0.059
	NPP		0.58	0.57	(0.14,0.95)	(0.17,0.95)	-0.0035	0.0042	0.065	84.3	94.3	0.26	-0.00040	0.00024	0.016	93.3	0.061
2.1.9	DB	1061					-0.0049	0.0050	0.071	79.4	94.5	0.28	0.0015	0.00032	0.018	94.3	0.070
	DF						-0.0049	0.0051	0.071	79.8	94.2	0.27	0.00047	0.00031	0.018	94.9	0.069
	PB						-0.0037	0.0044	0.066	83.8	94.9	0.26	0.0013	0.00027	0.017	95.0	0.066
	PF						-0.0037	0.0044	0.066	84.4	94.8	0.26	0.00057	0.00027	0.016	94.8	0.065
	NPP		0.68	0.65	(0.16,0.98)	(0.22,0.99)	-0.004	0.0045	0.067	83.9	95.1	0.26	-0.0012	0.00029	0.017	94.3	0.066

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
2.2.1	DB	1093					0.0044	0.014	0.12	86.7	97.0	0.54	0.0089	0.00035	0.016	99.6	0.080
	DF						0.0043	0.014	0.12	91.1	95.6	0.50	0.0052	0.00037	0.019	100.0	0.067
	PB						-0.089	0.016	0.090	90.0	88.7	0.39	0.0077	0.00028	0.015	98.7	0.058
	PF						-0.089	0.016	0.090	91.1	86.1	0.38	0.0071	0.00033	0.017	99.5	0.056
	NPP		0.54	0.54	(0.12,0.96)	(0.15,0.97)	-0.054	0.012	0.097	91.7	95.4	0.44	0.0015	0.00010	0.010	99.9	0.052
2.2.2	DB	1083					0.0028	0.017	0.13	84.4	95.8	0.53	0.0016	0.0011	0.033	92.2	0.12
	DF						0.0027	0.017	0.13	86.3	94.8	0.51	0.00054	0.00097	0.031	91.1	0.11
	PB						-0.071	0.015	0.10	89.9	88.5	0.40	-0.0095	0.00066	0.024	87.9	0.084
	PF						-0.071	0.015	0.10	90.5	87.3	0.39	-0.0090	0.00061	0.023	88.0	0.080
	NPP		0.53	0.54	(0.10,0.96)	(0.13,0.97)	-0.044	0.014	0.11	90.1	93.4	0.44	-0.014	0.00078	0.024	84.4	0.084
2.2.3	DB	1082					-0.0016	0.019	0.14	80.9	95.0	0.56	0.0026	0.0014	0.037	93.0	0.14
	DF						-0.0014	0.019	0.14	83.0	94.6	0.53	-0.00028	0.0012	0.035	92.1	0.13
	PB						-0.06	0.015	0.10	89.1	91.2	0.42	-0.020	0.0012	0.029	85.0	0.11
	PF						-0.061	0.015	0.10	89.9	90.6	0.42	-0.021	0.0012	0.028	79.9	0.10
	NPP		0.52	0.52	(0.078,0.96)	(0.1,0.96)	-0.041	0.015	0.11	88.2	94.2	0.46	-0.019	0.0013	0.031	85.2	0.11
2.2.4	DB	1088					0.0049	0.014	0.12	86.9	97.0	0.54	0.0088	0.00034	0.016	99.6	0.079
	DF						0.0048	0.014	0.12	91.2	95.6	0.50	0.0051	0.00037	0.019	100.0	0.067
	PB						-0.088	0.018	0.10	80.5	88.6	0.43	0.023	0.0010	0.022	90.3	0.082
	PF						-0.088	0.018	0.10	82.4	87.8	0.41	0.023	0.0010	0.022	93.1	0.077
	NPP		0.49	0.51	(0.11,0.95)	(0.13,0.94)	-0.051	0.013	0.10	88.1	96.0	0.46	0.0054	0.00020	0.013	99.9	0.063
2.2.5	DB	1077					0.0047	0.017	0.13	84.8	95.6	0.53	0.0015	0.0011	0.033	92.1	0.12
	DF						0.0045	0.017	0.13	86.7	94.6	0.51	0.00048	0.00097	0.031	91.2	0.11
	PB						-0.07	0.017	0.11	85.4	89.5	0.43	0.0051	0.00076	0.027	91.7	0.098
	PF						-0.070	0.017	0.11	86.9	88.3	0.42	0.0049	0.00068	0.026	93.6	0.093
	NPP		0.54	0.54	(0.11,0.96)	(0.14,0.96)	-0.043	0.014	0.11	87.9	93.7	0.45	-0.0075	0.00076	0.027	89.2	0.093

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
2.2.6	DB	1081					-0.0021	0.019	0.14	80.8	95.0	0.55	0.0023	0.0014	0.037	92.7	0.14
	DF						-0.0019	0.019	0.14	82.8	94.6	0.53	-0.00061	0.0012	0.035	91.7	0.13
	PB						-0.061	0.015	0.11	87.1	91.8	0.44	-0.0088	0.00097	0.030	91.4	0.11
	PF						-0.061	0.016	0.11	88.2	90.4	0.43	-0.010	0.00094	0.029	88.2	0.11
	NPP		0.6	0.58	(0.10,0.98)	(0.14,0.98)	-0.045	0.015	0.11	87.2	94.1	0.47	-0.013	0.0011	0.031	88.3	0.12
2.2.7	DB	1086					0.0046	0.014	0.12	86.8	97.0	0.54	0.0089	0.00034	0.016	99.7	0.08
	DF						0.0045	0.014	0.12	91.3	95.6	0.50	0.0052	0.00037	0.018	100.0	0.067
	PB						-0.089	0.021	0.11	71.3	90.0	0.48	0.047	0.0031	0.030	60.6	0.11
	PF						-0.088	0.021	0.11	73.4	89.1	0.46	0.047	0.003	0.029	67.9	0.10
	NPP		0.42	0.45	(0.088,0.90)	(0.09,0.88)	-0.046	0.013	0.11	83.6	96.3	0.48	0.0099	0.00038	0.017	99.7	0.077
2.2.8	DB	1085					0.0029	0.017	0.13	84.5	95.7	0.53	0.0016	0.0011	0.033	92.2	0.12
	DF						0.0028	0.017	0.13	86.4	94.7	0.51	0.00056	0.00097	0.031	91.2	0.11
	PB						-0.071	0.019	0.12	78.7	90.4	0.46	0.025	0.0016	0.031	86.0	0.11
	PF						-0.071	0.019	0.12	81.2	88.6	0.45	0.023	0.0014	0.029	91.2	0.11
	NPP		0.51	0.51	(0.098,0.94)	(0.12,0.94)	-0.043	0.015	0.12	85.1	94.1	0.47	0.00043	0.00090	0.030	92.3	0.11
2.2.9	DB	1086					-0.0019	0.019	0.14	80.8	94.9	0.55	0.0024	0.0014	0.037	92.7	0.14
	DF						-0.0017	0.019	0.14	82.9	94.6	0.53	-0.00053	0.0012	0.035	91.7	0.13
	PB						-0.061	0.017	0.11	83.5	92.4	0.47	0.0058	0.0010	0.031	94.1	0.12
	PF						-0.062	0.017	0.11	85.0	91.2	0.45	0.0038	0.00095	0.031	93.9	0.11
	NPP		0.62	0.60	(0.11,0.98)	(0.16,0.98)	-0.046	0.016	0.12	84.9	94.0	0.48	-0.0043	0.0011	0.033	92.4	0.12
2.3.1	DB	1082					0.00017	0.0042	0.065	85.9	94.8	0.26	0.0015	0.000098	0.0098	98.7	0.036
	DF						0.00013	0.0042	0.065	86.2	94.6	0.26	0.0016	0.00013	0.011	99.1	0.037
	PB						0.035	0.0047	0.059	97.6	92.5	0.24	0.0031	0.00011	0.010	97.4	0.036
	PF						0.035	0.0047	0.059	97.7	91.7	0.24	0.0037	0.00014	0.011	98.2	0.037
	NPP		0.54	0.54	(0.13,0.95)	(0.16,0.95)	0.020	0.0040	0.060	93.9	94.8	0.25	-0.00043	0.000070	0.0084	99.0	0.031

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
2.3.2	DB	1062					-0.0038	0.0047	0.068	81.5	94.4	0.27	-0.000021	0.00026	0.016	93.7	0.062
	DF						-0.0039	0.0047	0.068	81.8	94.1	0.26	-0.00011	0.00024	0.015	95.1	0.06
	PB						0.023	0.0043	0.062	94.3	94.0	0.25	-0.0028	0.00022	0.015	92.7	0.057
	PF						0.023	0.0043	0.062	94.4	93.6	0.24	-0.0029	0.00021	0.014	93.7	0.055
	NPP		0.56	0.55	(0.11,0.97)	(0.15,0.97)	0.013	0.0043	0.064	90.3	94.5	0.25	-0.0048	0.00025	0.015	89.9	0.057
2.3.3	DB	1085					-0.0050	0.0050	0.070	79.5	94.7	0.28	0.0016	0.00033	0.018	93.9	0.07
	DF						-0.0050	0.0050	0.070	80.0	94.4	0.27	0.00060	0.00032	0.018	94.7	0.069
	PB						0.016	0.0044	0.064	92.3	94.5	0.25	-0.0054	0.00030	0.017	91.7	0.064
	PF						0.016	0.0044	0.064	92.4	94.4	0.25	-0.00601	0.00031	0.016	90.4	0.063
	NPP		0.52	0.52	(0.079,0.97)	(0.11,0.96)	0.0090	0.0045	0.066	89.1	94.6	0.26	-0.0050	0.00031	0.017	92.0	0.065
2.3.4	DB	1076					0.0036	0.0040	0.063	87.5	95.9	0.26	0.0015	0.000094	0.0096	98.8	0.036
	DF						0.0035	0.004	0.063	88.0	95.3	0.26	0.0015	0.00013	0.011	99.3	0.037
	PB						0.023	0.0043	0.061	94.1	94.2	0.26	0.0058	0.00017	0.012	95.5	0.040
	PF						0.023	0.0043	0.061	94.0	93.9	0.25	0.0065	0.00020	0.013	96.7	0.042
	NPP		0.53	0.54	(0.14,0.94)	(0.16,0.94)	0.014	0.0039	0.061	91.6	95.9	0.26	0.0012	0.000090	0.0094	98.3	0.035
2.3.5	DB	1073					-0.0038	0.0047	0.068	81.4	94.4	0.27	-0.000071	0.00026	0.016	93.6	0.062
	DF						-0.0039	0.0046	0.068	81.8	94.1	0.26	-0.00016	0.00024	0.015	95.0	0.060
	PB						0.023	0.0045	0.063	93.2	94.1	0.25	0.0024	0.00023	0.015	94.8	0.059
	PF						0.023	0.0045	0.063	93.7	93.9	0.25	0.0021	0.00022	0.015	95.6	0.057
	NPP		0.57	0.56	(0.12,0.96)	(0.15,0.96)	0.013	0.0043	0.065	90.1	94.6	0.25	-0.0024	0.00024	0.015	92.1	0.059
2.3.6	DB	1066					-0.0058	0.0050	0.070	79.2	94.7	0.28	0.0015	0.00033	0.018	93.8	0.070
	DF						-0.0058	0.0050	0.070	79.6	94.4	0.27	0.00050	0.00032	0.018	94.7	0.069
	PB						0.015	0.0045	0.065	91.5	94.7	0.26	-0.0020	0.00028	0.017	94.1	0.065
	PF						0.015	0.0044	0.065	91.7	94.7	0.26	-0.0027	0.00028	0.016	93.1	0.064
	NPP		0.62	0.59	(0.11,0.98)	(0.15,0.98)	0.0093	0.0045	0.066	89.3	94.7	0.26	-0.0032	0.00030	0.017	93.3	0.066

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
2.3.7	DB	1078					-0.00033	0.0042	0.065	85.7	94.7	0.26	0.0016	0.000098	0.0098	98.7	0.036
	DF						-0.00038	0.0042	0.065	86.1	94.5	0.26	0.0017	0.00013	0.011	99.1	0.037
	PB						0.034	0.0054	0.065	94.9	92.9	0.26	0.019	0.00065	0.017	74.0	0.053
	PF						0.034	0.0054	0.065	94.7	92.8	0.26	0.020	0.00067	0.016	76.4	0.053
	NPP		0.39	0.41	(0.093,0.81)	(0.086,0.78)	0.014	0.004	0.062	91.2	95.5	0.25	0.0019	0.00011	0.01	98.5	0.036
2.3.8	DB	1079					-0.0036	0.0047	0.068	81.4	94.4	0.27	-0.00017	0.00026	0.016	93.5	0.062
	DF						-0.0036	0.0047	0.068	81.9	94.2	0.26	-0.00026	0.00024	0.015	94.9	0.06
	PB						0.023	0.0048	0.065	91.7	94.0	0.26	0.0095	0.00035	0.016	90.5	0.061
	PF						0.023	0.0048	0.065	91.9	94.0	0.26	0.0090	0.00032	0.016	93.3	0.060
	NPP		0.51	0.52	(0.11,0.93)	(0.13,0.92)	0.012	0.0044	0.066	89.3	94.4	0.26	-0.00014	0.00025	0.016	93.0	0.061
2.3.9	DB	1069					-0.0047	0.0050	0.071	79.7	94.6	0.28	0.0014	0.00033	0.018	94.0	0.070
	DF						-0.0047	0.0050	0.071	80.1	94.2	0.27	0.00033	0.00032	0.018	94.8	0.069
	PB						0.016	0.0046	0.066	91.1	94.6	0.26	0.0026	0.00029	0.017	94.9	0.066
	PF						0.016	0.0046	0.066	91.6	94.8	0.26	0.0019	0.00028	0.017	94.8	0.066
	NPP		0.63	0.61	(0.13,0.97)	(0.18,0.98)	0.010	0.0045	0.067	89.2	94.7	0.27	-0.00069	0.00029	0.017	94.5	0.067
2.4.1	DB	1092					0.0044	0.014	0.12	86.7	97.0	0.54	0.0087	0.00034	0.016	99.6	0.079
	DF						0.0043	0.014	0.12	91.1	95.6	0.5	0.0050	0.00037	0.018	100.0	0.067
	PB						-0.00012	0.0082	0.090	98.7	96.4	0.38	0.0041	0.00017	0.012	99.5	0.052
	PF						-0.00014	0.0082	0.090	99.1	95.3	0.37	0.0029	0.00022	0.015	99.9	0.05
	NPP		0.61	0.60	(0.15,0.98)	(0.19,0.99)	0.0018	0.0087	0.094	97.9	97.3	0.42	0.00014	0.000078	0.0088	100.0	0.048
2.4.2	DB	1075					0.0041	0.017	0.13	84.8	95.8	0.53	0.0014	0.0011	0.033	92.1	0.12
	DF						0.0039	0.017	0.13	86.6	94.8	0.51	0.00040	0.00097	0.031	91.2	0.11
	PB						0.0028	0.010	0.10	98.1	93.5	0.39	-0.014	0.00073	0.023	82.8	0.078
	PF						0.0026	0.010	0.10	98.2	92.8	0.38	-0.014	0.00069	0.022	82.5	0.075
	NPP		0.61	0.59	(0.13,0.98)	(0.17,0.99)	0.0031	0.011	0.11	96.2	95.0	0.42	-0.017	0.00083	0.023	80.7	0.079

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
2.4.3	DB	1088					-0.0015	0.019	0.14	80.9	94.9	0.56	0.0028	0.0014	0.037	92.8	0.14
	DF						-0.0013	0.019	0.14	82.9	94.6	0.53	-0.00017	0.0012	0.035	91.8	0.13
	PB						-0.0014	0.011	0.11	96.9	94.7	0.42	-0.023	0.0014	0.028	82.4	0.10
	PF						-0.0015	0.011	0.11	97.2	93.8	0.41	-0.024	0.0013	0.027	76.6	0.099
	NPP		0.58	0.57	(0.098,0.98)	(0.13,0.98)	-0.0014	0.013	0.11	93.6	94.9	0.45	-0.021	0.0014	0.030	83.9	0.11
2.4.4	DB	1095					0.0048	0.014	0.12	86.8	97.0	0.54	0.0089	0.00035	0.016	99.6	0.080
	DF						0.0047	0.014	0.12	91.2	95.6	0.5	0.0052	0.00037	0.019	100.0	0.067
	PB						0.00042	0.010	0.10	96.7	96.2	0.42	0.018	0.00074	0.020	93.6	0.075
	PF						0.00046	0.010	0.10	97.3	95.5	0.40	0.019	0.00079	0.021	95.9	0.071
	NPP		0.56	0.56	(0.13,0.96)	(0.16,0.97)	0.0025	0.0097	0.099	96.8	97.2	0.44	0.0044	0.00017	0.012	99.9	0.06
2.4.5	DB	1072					0.0033	0.017	0.13	84.6	95.8	0.53	0.0017	0.0011	0.033	92.3	0.12
	DF						0.0031	0.017	0.13	86.6	94.8	0.51	0.00063	0.00097	0.031	91.2	0.11
	PB						0.0027	0.012	0.11	96.2	93.8	0.42	0.00039	0.00070	0.026	91.0	0.094
	PF						0.0025	0.012	0.11	96.6	93.0	0.41	0.00041	0.00062	0.025	92.3	0.089
	NPP		0.60	0.59	(0.13,0.97)	(0.17,0.98)	0.0029	0.012	0.11	95.2	94.5	0.44	-0.0093	0.00075	0.026	88.3	0.090
2.4.6	DB	1085					-0.002	0.019	0.14	80.7	94.9	0.55	0.0023	0.0014	0.037	92.7	0.14
	DF						-0.0018	0.019	0.14	82.8	94.6	0.53	-0.00060	0.0012	0.035	91.7	0.13
	PB						-0.0020	0.012	0.11	95.5	95.5	0.44	-0.013	0.0010	0.029	89.7	0.11
	PF						-0.0023	0.012	0.11	95.9	95.4	0.43	-0.014	0.00099	0.028	85.6	0.10
	NPP		0.65	0.62	(0.12,0.98)	(0.17,0.99)	-0.0022	0.013	0.11	93.3	95.9	0.46	-0.015	0.0012	0.030	87.4	0.11
2.4.7	DB	1088					0.0056	0.014	0.12	87.0	97.0	0.54	0.0089	0.00035	0.016	99.6	0.080
	DF						0.0055	0.014	0.12	91.4	95.7	0.50	0.0052	0.00037	0.019	100.0	0.067
	PB						0.00098	0.013	0.11	92.8	96.0	0.47	0.042	0.0026	0.029	67.6	0.10
	PF						0.0012	0.013	0.11	93.5	95.0	0.45	0.042	0.0025	0.028	73.9	0.098
	NPP		0.47	0.48	(0.11,0.92)	(0.11,0.91)	0.0035	0.011	0.10	95.1	97.2	0.47	0.0093	0.00035	0.016	99.7	0.075

Table E.1: Simulation Study Results (continued)

Scenario	Model[a]	N[b]	Discounting Factor (a_0)				Treatment Effect						Intracluster Correlation Coefficient				
			Median	Mean	95% CrI[c]	95% HPDI[d]	Bias	MSE[e]	Emp. SE[f]	Power	Coverage	Interval Width[g]	Bias	MSE[e]	Emp. SE[f]	Coverage	Interval Width[h]
2.4.8	DB	1082					0.0033	0.017	0.13	84.6	95.7	0.53	0.0012	0.0011	0.032	92.2	0.12
	DF						0.0031	0.017	0.13	86.5	94.6	0.51	0.00022	0.00096	0.031	91.1	0.11
	PB						0.0022	0.014	0.12	92.1	94.4	0.45	0.020	0.0013	0.030	89.1	0.11
	PF						0.0021	0.014	0.12	93.3	93.1	0.44	0.019	0.0012	0.029	93.0	0.10
	NPP		0.56	0.55	(0.12,0.96)	(0.15,0.96)	0.0027	0.013	0.11	92.7	94.6	0.46	-0.0011	0.00086	0.029	91.7	0.10
2.4.9	DB	1086					-0.0018	0.019	0.14	80.8	94.9	0.55	0.0022	0.0014	0.037	93.0	0.14
	DF						-0.0017	0.019	0.14	82.9	94.6	0.53	-0.00066	0.0012	0.035	91.9	0.13
	PB						-0.0022	0.013	0.11	93.4	95.9	0.46	0.0019	0.00094	0.031	94.2	0.12
	PF						-0.0024	0.013	0.11	94.0	95.2	0.45	0.00015	0.00089	0.030	93.0	0.11
	NPP		0.66	0.63	(0.14,0.98)	(0.19,0.99)	-0.0021	0.013	0.11	92.2	95.9	0.47	-0.0064	0.0011	0.032	92.4	0.12

[a] DB: Definitive Data, Bayesian Analysis; DF: Definitive Data, Frequentist Analysis; PB: Pooled Data, Bayesian Analysis; PF = Pooled Data, Frequentist Analysis; NPP = Normalised Power Prior

[b] N denotes the number of iterations (out of 1100) for each scenario used in calculation of the metrics displayed in the table

[c] Credible Interval

[d] Highest Posterior Density Interval

[e] Mean Squared Error

[f] Empirical Standard Error

[g] Of the 95% Credible Interval

[h] Of the 95% Highest Posterior Density Interval

Appendix F

NPP Documentation

NPP is used to fit a Normalised Power Prior for analysis of a (current) dataset, using a second (historical) dataset to formulate the Power Prior, where both datasets contain clustering and have a continuous outcome.

```
NPP(X, X0, Y, Y0, Z, Z0, sigma.b.prior = c("hnormal", "hcauchy"),
intercept.prior.mean = 0, intercept.prior.sd = NULL,
reg.prior.mean = 0, reg.prior.sd = NULL,
sigma.b.prior.parm = NULL, sigma.prior.parm = NULL,
nits_normalise = 2000, burnin_normalise = NULL,
nchains_normalise = 4, max_treedepth_normalise = 10,
thin_normalise = 1, adapt_delta_normalise = 0.95,
nits_npp = 5000, burnin_npp = NULL, nchains_npp = 4,
max_treedepth_npp = 10, thin_npp = 1, adapt_delta_npp = 0.95,
a0_increment = 0.05, seed = 12345, parallel = F, ...)
```

X A matrix. The design matrix for the current dataset, excluding the intercept term. The first column must represent treatment allocation, where a 1 represents treatment and 0 represents control.

X0 A matrix. The design matrix for the historical dataset, excluding the intercept term. The first column must represent treatment allocation, where a 1 represents treatment and 0 represents control.

Y A vector containing the continuous outcome data for the current dataset.

Y0 A vector containing the continuous outcome data for the historical dataset.

Z A vector of consecutive integers containing cluster indices for the current dataset.

Z0 A vector of consecutive integers containing cluster indices for the historical dataset.

sigma.b.prior One of either “hnormal” or “hcauchy” to indicate whether a Half-Normal or a Half-Cauchy prior distribution should be fitted to the between-cluster SD parameter.

intercept.prior.mean The mean for the normal prior distribution for the intercept. Defaults to 0.

intercept.prior.sd The standard deviation for the normal prior distribution for the intercept.

reg.prior.mean A vector of means for the normal prior distribution for each of the regression coefficients (of length equal to the number of columns of X0).

reg.prior.sd A vector of standard deviations for the normal prior distribution for each of the regression coefficients (of length equal to the number of columns of `X0`).

sigma.b.prior.parm The parameter for the prior distribution for the between-cluster standard deviation. If `sigma.b.prior = "hcauchy"` this represents the scale parameter of the Half-Cauchy distribution. If `sigma.b.prior = "hnormal"` this represents the standard deviation of the Half-Normal distribution.

sigma.prior.parm The rate parameter for the exponential prior distribution for the residual standard deviation.

nits_normalise An integer. Number of iterations per chain used in the Markov Chain Monte Carlo procedure for estimating the normalising constant. Defaults to 2000. See `rstan::stan()` for further details.

burnin_normalise An integer. Number of iterations per chain to be discarded in the Markov Chain Monte Carlo procedure for estimating the normalising constant. Defaults to one half of `nits_normalise`. See `rstan::stan()` for further details.

nchains_normalise An integer. Number of chains to be used in the Markov Chain Monte Carlo procedure for estimating the normalising constant. Defaults to 4. See `rstan::stan()` for further details.

max_treedepth_normalise An integer. Maximum treedepth for the Markov Chain Monte Carlo procedure for estimating the normalising constant. Defaults to 10. See `rstan::stan()` for further details.

thin_normalise A positive integer specifying the period for saving Markov Chain Monte Carlo samples for the procedure estimating the normalising constant. Defaults to 1. See `rstan::stan()` for further details.

adapt_delta_normalise Value of adapt delta used in the Markov Chain Monte Carlo procedure for estimating the normalising constant. Defaults to 0.95. See `rstan::stan()` for further details.

nits_npp An integer. Number of iterations per chain used in the Markov Chain Monte Carlo procedure for fitting the NPP model. Defaults to 5000. See `rstan::stan()` for further details.

burnin_npp An integer. Number of iterations per chain to be discarded in the Markov Chain Monte Carlo procedure for fitting the NPP model. Defaults to one half of `nits_npp`. See `rstan::stan()` for further details.

nchains_npp An integer. Number of chains to be used in the Markov Chain Monte Carlo procedure for fitting the NPP model. Defaults to 4. See `rstan::stan()` for further details.

max_treedepth_npp An integer. Maximum treedepth for the Markov Chain Monte Carlo procedure for fitting the NPP model. Defaults to 10. See `rstan::stan()` for further details.

thin_npp A positive integer specifying the period for saving Markov Chain Monte Carlo samples for the procedure fitting the NPP model. Defaults to 1. See `rstan::stan()` for further details.

adapt_delta_npp Value of adapt delta used in the Markov Chain Monte Carlo procedure fitting the NPP model. Defaults to 0.95. See `rstan::stan()` for further details.

a0_increment Value of the increments by which `a0` is increased between each estimation of the normalising constant. Defaults to 0.05.

seed Set the seed.

parallel Logical. If `TRUE`, parallelisation of the MCMC chains is implemented using the number of cores available on the local machine.

... Further arguments passed to or from other methods.

An object of S4 class `stanfit` representing the fitted results. `beta[1]` represents the treatment effect parameter.

Appendix G

NPP Example

```
X <- as.matrix(currdat[,c("Group","T0SDS_BMI")])
X0 <- as.matrix(histdat[,c("Condition","BMI0sds")])
Y <- c(currdat[,c("BMI_Change")])
Y0 <- c(histdat[,c("BMI_Change")])
Z <- c(currdat[,c("SchoolCode")])
Z0 <- c(histdat[,c("School")])

C <- PPCRCT::NPP(X = X, X0 = X0, Y = Y, Y0 = Y0,
                 Z = Z, Z0 = Z0,
                 sigma.b.prior = "hcauchy",
                 sigma.b.prior = "hcauchy",
                 intercept.prior.mean = 0,
                 intercept.prior.sd = 5,
                 reg.prior.mean = c(0,0), reg.prior.sd = c(5,5),
                 sigma.b.prior.parm = 0.3, sigma.prior.parm = 1,
                 nits_normalise = 2000,
                 burnin_normalise = 1000,
                 nchains_normalise = 4,
                 max_treedepth_normalise = 10,
                 thin_normalise = 1,
                 adapt_delta_normalise = 0.99,
                 nits_npp = 3500,
                 burnin_npp = 1750,
                 nchains_npp = 4,
                 max_treedepth_npp = 10,
                 thin_npp = 1,
                 adapt_delta_npp = 0.99, seed = 201580,
                 parallel = FALSE)
```

Appendix H

FDPP Documentation

`FDPP` is used to fit a Fixed Discounting Power Prior (FDPP) to analysis of a (current) dataset, using a second (historical) dataset to formulate the Power Prior, where both datasets contain clustering and have a continuous outcome.

```
FDPP(X, X0, Y, Y0, Z, Z0, a0, partial.borrowing = F,
sigma.b.prior = c("hnormal", "hcauchy"),
intercept.prior.mean = 0, intercept.prior.sd = NULL,
reg.prior.mean = 0, reg.prior.sd = NULL,
sigma.b.prior.parm = NULL, sigma.prior.parm = NULL,
nits_fdpp = 5000, burnin_fdpp = NULL, nchains_fdpp = 4,
max_treedepth_fdpp = 10, thin_fdpp = 1,
adapt_delta_fdpp = 0.95, seed = 12345, parallel = F, ...)
```

`X` A matrix. The design matrix for the current dataset, excluding the intercept term. The first column must represent treatment allocation, where a 1 represents treatment and 0 represents control.

`X0` A matrix. The design matrix for the historical dataset, excluding the intercept term. The first column must represent treatment allocation, where a 1 represents treatment and 0 represents control.

`Y` A vector containing the continuous outcome data for the current dataset.

`Y0` A vector containing the continuous outcome data for the historical dataset.

`Z` A vector of consecutive integers containing cluster indices for the current dataset.

`Z0` A vector of consecutive integers containing cluster indices for the historical dataset.

`a0` The discounting factor. Must be a value between 0 and 1.

`partial.borrowing` logical. If `TRUE`, the Partial Borrowing Power Prior is used (borrowing information from the treatment effect parameter only). If `FALSE`, the Fixed Discounting Power Prior (borrowing information from all parameters) is used. Defaults to `FALSE`.

`sigma.b.prior` One of either “hnormal” or “hcauchy” to indicate whether a Half-Normal or a Half-Cauchy prior distribution should be fitted to the between-cluster SD parameter.

`intercept.prior.mean` The mean for the normal prior distribution for the intercept. Defaults to 0.

intercept.prior.sd The standard deviation for the normal prior distribution for the intercept.

reg.prior.mean A vector of means for the normal prior distribution for each of the regression coefficients (of length equal to the number of columns of `X0`).

reg.prior.sd A vector of standard deviations for the normal prior distribution for each of the regression coefficients (of length equal to the number of columns of `X0`).

sigma.b.prior.parm The parameter for the prior distribution for the between-cluster standard deviation. If `sigma.b.prior = "hcauchy"` this represents the scale parameter of the Half-Cauchy distribution. If `sigma.b.prior = "hnormal"` this represents the standard deviation of the Half-Normal distribution.

sigma.prior.parm The rate parameter for the exponential prior distribution for the residual standard deviation.

nits_fdpp An integer. Number of iterations per chain used in the Markov Chain Monte Carlo procedure for fitting the FDPP model. Defaults to 5000. See `rstan::stan()` for further details.

burnin_fdpp An integer. Number of iterations per chain to be discarded in the Markov Chain Monte Carlo procedure for fitting the FDPP model. Defaults to one half of `nits_fdpp`. See `rstan::stan()` for further details.

nchains_fdpp An integer. Number of chains to be used in the Markov Chain Monte Carlo procedure for fitting the FDPP model. Defaults to 4. See `rstan::stan()` for further details.

max_treedepth_fdpp An integer. Maximum treedepth for the Markov Chain Monte Carlo procedure for fitting the FDPP model. Defaults to 10. See `rstan::stan()` for further details.

thin_fdpp A positive integer specifying the period for saving Markov Chain Monte Carlo samples for the procedure fitting the FDPP model. Defaults to 1. See `rstan::stan()` for further details.

adapt_delta_fdpp Value of adapt delta used in the Markov Chain Monte Carlo procedure fitting the FDPP model. Defaults to 0.95. See `rstan::stan()` for further details.

seed Set the seed.

parallel logical. If `TRUE`, parallelisation of the MCMC chains is implemented using the number of cores available on the local machine.

`...` Further arguments passed to or from other methods.

An object of S4 class `stanfit` representing the fitted results. `beta[1]` represents the treatment effect parameter.

Appendix I

FDPP Example

```
X <- as.matrix(currdat[,c("Group","T0SDS_BMI")])
X0 <- as.matrix(histdat[,c("Condition","BMI0sds")])
Y <- c(currdat[,c("BMI_Change")])
Y0 <- c(histdat[,c("BMI_Change")])
Z <- c(currdat[,c("SchoolCode")])
Z0 <- c(histdat[,c("School")])

C_fixed = PPCRCT::FDPP(X = X, X0 = X0, Y = Y, Y0 = Y0,
                       Z = Z, Z0 = Z0, a0 = 0.5,
                       partial.borrowing = T,
                       sigma.b.prior =  "hcauchy",
                       intercept.prior.mean = 0,
                       intercept.prior.sd = 5,
                       reg.prior.mean = c(0,0),
                       reg.prior.sd = c(5,5),
                       sigma.b.prior.parm = 0.3,
                       sigma.prior.parm = 1,
                       nits_fdpp = 3500, burnin_fdpp = 1750,
                       nchains_fdpp = 4, max_treedepth_fdpp = 10,
                       thin_fdpp = 1, adapt_delta_fdpp = 0.9,
                       seed = 201580, parallel = FALSE)
```

Bound copies of published papers

Jones *et al. Systematic Reviews* (2021) 10:91
https://doi.org/10.1186/s13643-021-01637-1

Systematic Reviews

RESEARCH Open Access

Bayesian statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review

Benjamin G. Jones[1,2*], Adam J. Streeter[1,3], Amy Baker[1], Rana Moyeed[4] and Siobhan Creanor[1,5,6]

Abstract

Background: In a cluster randomised controlled trial (CRCT), randomisation units are "clusters" such as schools or GP practices. This has methodological implications for study design and statistical analysis, since clustering often leads to correlation between observations which, if not accounted for, can lead to spurious conclusions of efficacy/ effectiveness. Bayesian methodology offers a flexible, intuitive framework to deal with such issues, but its use within CRCT design and analysis appears limited. This review aims to explore and quantify the use of Bayesian methodology in the design and analysis of CRCTs, and appraise the quality of reporting against CONSORT guidelines.

Methods: We sought to identify all reported/published CRCTs that incorporated Bayesian methodology and papers reporting development of new Bayesian methodology in this context, without restriction on publication date or location. We searched Medline and Embase and the Cochrane Central Register of Controlled Trials (CENTRAL). Reporting quality metrics according to the CONSORT extension for CRCTs were collected, as well as demographic data, type and nature of Bayesian methodology used, journal endorsement of CONSORT guidelines, and statistician involvement.

Results: Twenty-seven publications were included, six from an additional hand search. Eleven (40.7%) were reports of CRCT results: seven (25.9%) were primary results papers and four (14.8%) reported secondary results. Thirteen papers (48.1%) reported Bayesian methodological developments, the remaining three (11.1%) compared different methods. Four (57.1%) of the primary results papers described the method of sample size calculation; none clearly accounted for clustering. Six (85.7%) clearly accounted for clustering in the analysis. All results papers reported use of Bayesian methods in the analysis but none in the design or sample size calculation.

(Continued on next page)

* Correspondence: b.g.jones@exeter.ac.uk
[1]Medical Statistics, Faculty of Health: Medicine, Dentistry and Human Sciences, University of Plymouth, Room N15, ITTC Building 1, Plymouth Science Park, Plymouth, Devon PL6 8BX, UK
[2]NIHR ARC South West Peninsula (PenARC), College of Medicine and Health, University of Exeter, Exeter, Devon, UK
Full list of author information is available at the end of the article

(Continued from previous page)

Conclusions: The popularity of the CRCT design has increased rapidly in the last twenty years but this has not been mirrored by an uptake of Bayesian methodology in this context. Of studies using Bayesian methodology, there were some differences in reporting quality compared to CRCTs in general, but this study provided insufficient data to draw firm conclusions. There is an opportunity to further develop Bayesian methodology for the design and analysis of CRCTs in order to expand the accessibility, availability, and, ultimately, use of this approach.

Keywords: Cluster randomised trial, Bayesian, CONSORT statement, Sample size, Statistical power, Hierarchical modelling

Background

In a cluster randomised controlled trial (CRCT), randomisation occurs at the group (or "cluster") level as opposed to the individual level that is typical in traditional Randomised Controlled Trials (RCTs). Examples of naturally-occurring clusters include schools, villages and GP practices. Randomisation of clusters, rather than individuals, is conducted for a number of reasons: (i) when the intervention is to be delivered at the cluster level (e.g. to a whole school/class within a school); (ii) when there is a risk of contamination, either between participants or those delivering the intervention; or (iii) when there is a clear administrative, logistic or cost-based rationale [1].

Cluster randomisation has methodological implications that go beyond merely the randomisation procedure itself. Measurements on individuals within the same cluster are likely to be more correlated to one another than measurements on individuals from different clusters. This correlation creates an additional level of complexity, which must be accounted for in both the study design and sample size calculation, and the statistical analysis. Failure to do so can result in an underpowered study and ultimately spurious conclusions about the efficacy or effectiveness of the intervention or treatment under investigation.

CRCTs are a relatively novel study design, but the methodology is now well established in the literature. Prior to the 1980s, there was only sparse use of CRCTs [2], but they have become increasingly popular in the last 30 years, from just seven reported in 1990, to over 120 in 2008 [3, 4]. Figure 1 provides an illustration of this increase in popularity by displaying the number of search results by year for "cluster randomised controlled trials" with restriction to publication title. Alongside such a rapid increase in the use of the CRCT design, there have been some attempts to develop new Bayesian methodology for the design and analysis of such trials. This ranges from utilising well-established Bayesian hierarchical modelling approaches to account for the clustered nature of the data [5], through to more novel approaches to study design and sample size calculation such as that developed by Turner et al [6, 7]. The Bayesian approach to analysis in particular may offer a number of advantages over the frequentist approach. In a random effects setting, as is often applicable in the analysis of a CRCT, the hierarchical Bayesian framework provides a flexible, intuitive approach to statistical inference. Furthermore, Bayesian analysis facilitates a more natural, probabilistic interpretation of results and moves away from frequentist hypothesis testing and *p*-values, an approach which has been criticised in recent years [8]. Whilst often criticised, the incorporation of prior information into a statistical analysis can facilitate more informative conclusions, which reflect all the available evidence as opposed to simply the evidence offered from the single dataset at hand. In many cases, the rationale for the inclusion of informative priors is sound, for example results from previous research or even existing data (such as pilot or feasibility studies). However, whilst the advantages of the Bayesian approach to both the analysis of clinical trials [9] and hierarchical data [10] are clear and have been documented, it is unclear whether such methods are being regularly utilised within the context of CRCTs.

With the increased use of CRCTs, the need for consistent, high-quality reporting is crucial. In response to this recognised need, the CONSORT extension to cluster randomised trials was first published in 2004 [11] and updated in 2012 [12]. The CONSORT statement provides recommendations for reporting of randomised trials, and whilst there is no extension for Bayesian trials, it was not written exclusively for frequentist methods. A recent review of the methodological quality of sample size calculations in a sample of 300 CRCTs published between 2000 and 2008, found that only 55.3% (166) presented a sample size calculation, of which only 61.4% (102) accounted appropriately for clustering [13]. A separate recently published review of the same sample of CRCTs examined the impact of the 2004 CONSORT extension on more general methodological quality and concluded that adherence to published reporting guidelines and quality remains low [14]. Similar reviews of CRCT reporting quality have been conducted and produced comparable conclusions [15, 16]. However, to our knowledge, none have focussed specifically on CRCTs

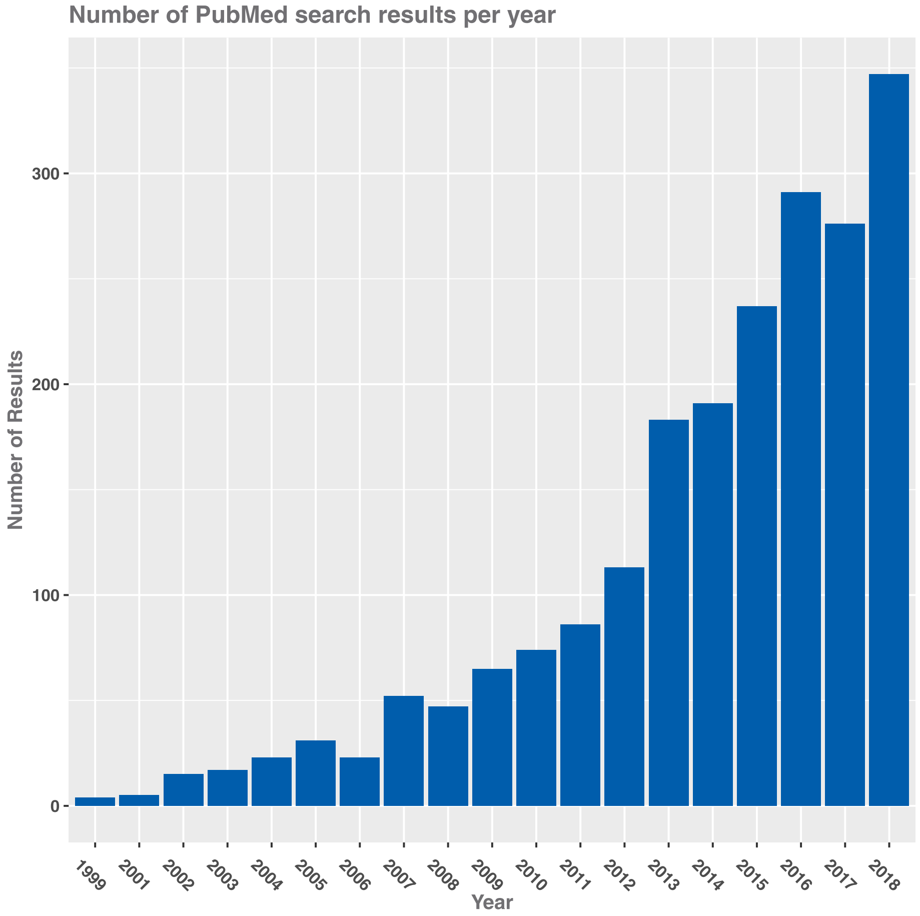


Fig. 1 Number of PubMed search results per year. Search term: "cluster randomized controlled trial"[Title] OR "cluster randomised controlled trial"[Title] NOT "stepped"[Title]. The search was conducted in February 2019 and partial data for that year was removed

which incorporated Bayesian methods, and so both the quantity and quality of these are unknown.

This review aims to:

(i) Quantify and explore the use of Bayesian methodology in the design and/or analysis of CRCTs;

(ii) Appraise the quality of reporting of CRCTs conducted in a Bayesian framework against the current relevant CONSORT guidelines and identify whether the reporting quality differs from previous reviews assessing reporting quality in CRCTs more generally (most of which likely, but not necessarily, pertain to frequentist trials).

The impact of the introduction of the CONSORT guidelines for CRCTs in 2004 and 2012 on reporting quality will also be appraised.

Methods

The protocol for this methodological systematic review was developed prospectively and made publically available online [17] before commencing the literature searches. The review was conducted and reported in accordance with the PRISMA guidelines [18].

Inclusion and exclusion criteria

We sought to identify all published parallel group CRCTs in which Bayesian methodology was used in either the study design (including sample size calculation) or statistical analysis. We also opted to include any papers in which Bayesian methodology was discussed or considered, even if such methods were not implemented in the study, whilst recognising that such a scenario would be unlikely. We did not restrict our search or inclusion on the basis of publication date, location, intervention type or population in any way, provided the relevant paper was published in the English language, due to resource limitations.

In order to be included in this review, it had to be evident that randomisation in the study occurred at a group level, in which multiple participants were randomised together, as per the definition of a CRCT.

We did not exclude references on the basis of type (category) of published paper. Specifically, we included not only primary reports of efficacy or effectiveness but also protocol papers, papers reporting secondary analyses and publications reporting results of pilot/feasibility studies. We also included studies reporting Bayesian methodological developments in the area of CRCTs. At the data extraction stage, we sought to identify supplementary literature related to the same study, if indicated, to obtain the required information, but only included such examples as a single entry. It was anticipated, for example, that this might include obtaining additional detail from a published protocol or monograph that had been omitted in the corresponding primary results paper.

We excluded papers reporting only cost-effectiveness. We also excluded studies implementing a stepped-wedge or other longitudinal cluster randomised design, as the methodological considerations are different and the reporting quality metrics presented in the CONSORT extension to CRCTs [12] are not valid for such longitudinal designs. Since commencement of this systematic review, however, separate guidelines for stepped-wedge designs have been published [19]. Conference proceedings and masters and PhD dissertations were not included.

Data sources and search methods

We searched both Medline and Embase using Ovid, as well as the Cochrane Central Register of Controlled Trials (CENTRAL), for relevant publications on 24 July 2018, without restriction on date of publication. The full electronic search strategy was an extension of that presented by Taljaard et al. [20] to identify CRCTs, adapted to identify only studies including the word “Bayes” in the title, abstract or text. The full electronic search strategy used to search Medline and Embase is shown in Table 1, with minor syntactic adaptations required in order to run the search in CENTRAL. The searches were undertaken by BJ. Additional literature was included where appropriate through hand searching of the authors’ own collection of references.

Reference sifting and quality control

After conducting electronic searches, all references were downloaded and imported to Mendeley [21] for electronic deduplication. Following this, remaining references were exported and uploaded to Rayyan [22]. BJ and AS independently reviewed each reference and made a decision to include or exclude on the basis of the information available from the title and the abstract assessed against the pre-specified inclusion/exclusion criteria outlined in the protocol [17]. Rayyan includes a blinding feature, which was switched on during the independent sifts and then disabled. Any disagreements were resolved through discussion and, where required, SC made a final decision.

Table 1 Search strategy used to search Medline and Embase within Ovid on 24 July 2018

#	Search
Existing published strategy for randomised controlled trials	
1	(article OR randomized controlled trials).pt.
2	Animals/
3	Humans/
4	#2 NOT (2 AND 3)
5	#1 NOT #4
Cluster design–related terms	
6	(cluster$ adj2 randomi$).tw.
7	((communit$ adj2 intervention$) or (communit$ adj2 randomi$)).tw.
8	group$ randomi$.tw.
9	#6 OR #7 OR #8
10	intervention?.tw.
11	Cluster Analysis/
12	Health Promotion/
13	Program Evaluation/
14	Health Education/
15	#10 OR #11 OR #12 OR #13 OR #14
16	#9 OR #15
Bayesian search terms	
17	bayes$.af.
18	#16 AND #17
Final search	
19	#18 AND #5
20	limit #19 to (randomized controlled trial)

pt. represents publication type; / represents MeSH search; $ allows for truncation of words; adj allows for adjacency between search words; tw represents text words in abstract and/or title; af represents all fields; ? is a wildcard which retrieves one or 0 characters

After the initial sift, full-text articles were obtained for all remaining references. BJ examined the full texts and again made inclusion/exclusion decisions using Rayyan. SC or AB re-examined approximately half each of all full texts and independently made inclusion or exclusion decisions. Any disagreements were once again resolved through further discussion.

Data extraction

For the primary and secondary published reports of trial results, we collected a range of data including demographic data, technical detail regarding design and analysis methodology with relation to Bayesian techniques, and information regarding statistician involvement with the study and their respective affiliations. For papers reporting primary results, we also collected a selection of reporting quality metrics taken from the 2012 CONSORT extension to CRCTs [12]. In addition, we recorded whether or not *p*-values were reported for comparison of baseline demographics, as has been collected in previous systematic reviews of CRCTs [15, 23], Clinical Trial Unit (CTU) involvement in the study, and journal endorsement of the CONSORT guidelines.

We considered the paper as having statistician involvement, via a previously used criterion [15, 24, 25], if there was a clearly designated statistician, or if at least one of the co-authors belonged to a department of epidemiology or biostatistics. If it was not possible to obtain this information from the authorship list on the paper, online searching was undertaken to attempt to determine this from the qualification or affiliation of the authors. In any cases where it was not possible to obtain the required information, statistician involvement was recorded as “no”. We also recorded the statistician’s affiliation to a CTU, an academic statistical department, a commercial pharmaceutical company, a clinical research organisation (CRO) or “other”. CTU involvement in the study was determined if at least one author had a listed affiliation to a CTU. If author affiliations were not available in the paper or online, this was recorded as “no”.

We classified journal endorsement of the CONSORT statement using previously defined criteria [15]: a journal’s strength of endorsement was classified as high if the words “required”, ”must”, ”should” or “strongly recommended” were used in their author instructions, a medium endorser if words “encouraged”, ”recommended”, “advised” or “please” were used, and a low endorser if “may wish to consider” or ”see CONSORT” was used. We included a fourth category, “none”, if the journal included no mention of the CONSORT statement in its guidelines to authors.

Separate data extraction forms were developed for primary and secondary results papers to ensure that all the required information was obtained independently, consistently and without bias. The forms were piloted by BJ prior to data extraction. Formal data extraction was not undertaken for the methodological papers, but rather these papers were examined for the purpose of qualitative reporting and descriptive summaries of the methods developed in order to gain an understanding of the extent of methodological developments in this area.

BJ conducted data extraction on all primary and secondary results papers. SC, AB and AS independently conducted approximately one-third each of the data extraction on all papers, and final data was agreed by the whole study team. BJ and SC also each independently classified the results papers as primary or secondary. Any disagreements were resolved through discussion. Separately, BJ examined the methodological papers for qualitative reporting, but no second data extraction was undertaken. BJ double-entered all data from the data extraction forms into separate excel spreadsheets for primary and secondary papers.

Analysis

We present descriptive statistics of frequencies and percentages or means and standard deviations, as appropriate, for demographic qualities relating to each of the results publications, including trial location, number of participants recruited and type of primary outcome, by category of published results (primary or secondary). For the reporting quality measures, we present the number of primary results papers satisfying each criterion overall, by year (before or after the publication of the 2012 extension to the CONSORT guidelines for CRCTs [12]), by journal endorsement of the CONSORT guidelines (high or medium versus low or none) and by statistician involvement in the trial. We also summarise the use or consideration of Bayesian methods in the design and/or sample size calculation and/or analysis, as well as the level of information incorporated into the prior distributions specified. We also outline for which parameters the prior distributions were specified, if this information was available. Finally, a qualitative synthesis of the methodological papers was undertaken to summarise the areas of focus in the development of new methods.

Results

We identified 325 records from our electronic searches, of which 48 were identified as duplicates and removed. The remaining 277 records were screened on the basis of the detail available within the title and abstract, of which 219 were excluded (51 were the wrong study design (such as *N*-of-1 trials or meta analyses), 160 were individually randomised trials, and eight were papers reporting cost-effectiveness only). Full texts were obtained for the remaining 58 papers. At this final stage, following independent review of the full texts, a further 37 were removed (25 were individually randomised, five did not include any mention of Bayesian methodology, six were the wrong study design and one paper reported only cost-effectiveness results), leaving 21 papers from the electronic search. A further six papers, all of which were methodological, were added through additional hand searches, resulting in a total of 27 papers included (Fig. 2). The full list of references for the included papers is detailed in Table 2. Eleven (40.7%) were reports of CRCT results, of which seven (63.6%, R1–R7) were primary results papers and four (36.4%, R8–R11) reported secondary analyses. Thirteen papers (48.1%, M1–M13)

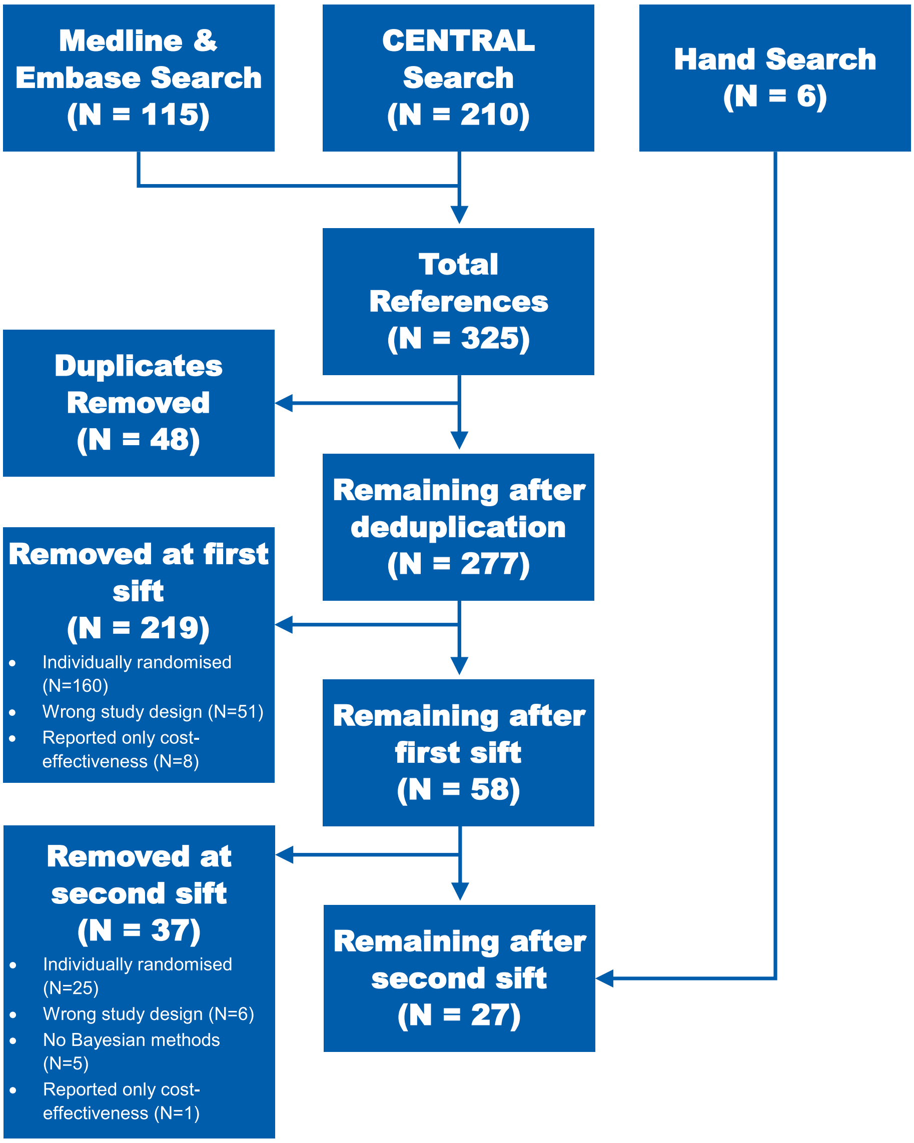

Medline & Embase Search (N = 115)
CENTRAL Search (N = 210)
Hand Search (N = 6)
Total References (N = 325)
Duplicates Removed (N = 48)
Remaining after deduplication (N = 277)
Removed at first sift (N = 219)
Individually randomised (N=160)
Wrong study design (N=51)
Reported only cost-effectiveness (N=8)
Remaining after first sift (N = 58)
Removed at second sift (N = 37)
Individually randomised (N=25)
Wrong study design (N=6)
No Bayesian methods (N=5)
Reported only cost-effectiveness (N=1)
Remaining after second sift (N = 27)

Fig. 2 Flow diagram of the identification process for the 27 publications included in this review

reported methodological developments and the remaining three (11.1%, C1–C3) reported comparisons of methods, assessing the performance of various existing methodology.

Demographics

Descriptions of demographics are displayed in Table 3. Target sample sizes and numbers of clusters were only collected for primary results papers. We deemed it necessary to distinguish "numbers approached" from target sample sizes, as the numbers approached seemed likely driven by logistical rather than statistical considerations, and so were not included in the summary statistics of the target sample sizes. Clear statistician association with a CTU was identified in one (12.5%) study. We were unable to identify more general CTU involvement with trial or data management in any instance.

Reporting quality

Reporting quality of the seven primary results papers was mixed (Table 4). Four (57.1%) included a description of the sample size calculation, but none of these clearly accounted for clustering, provided the intra-class correlation coefficient (ICC) used in the sample size calculation or took into consideration potential variability in cluster size or accounted for this in the sample size calculation. Similarly, none of the papers reported estimated ICCs for any of the primary or secondary outcomes, despite the potential value of

Table 2 References included in the review

R1	Carabin H, Millogo A, Ngowi HA, et al. Effectiveness of a community-based educational programme in reducing the cumulative incidence and prevalence of human Taenia solium cysticercosis in Burkina Faso in 2011–14 (EFECAB): a cluster-randomised controlled trial. *Lancet Glob Heal*. 2018;6(4):e411-e425. doi:10.1016/S2214-109X(18)30027-5
R2	Foxcroft DR, Callen H, Davies EL, Okulicz-Kozaryn K. Effectiveness of the strengthening families programme 10-14 in Poland: Cluster randomized controlled trial. *Eur J Public Health*. 2017;27(3):494-500. doi:10.1093/eurpub/ckw195
R3	Levy BT, Hartz A, Woodworth G, Xu Y, Sinift S. Interventions to Improving Osteoporosis Screening: An Iowa Research Network (IRENE) Study. *J Am Board Fam Med*. 2009;22(4):360-367. doi:10.3122/jabfm.2009.04.080071
R4	Ngowi HA, Carabin H, Kassuku AA, Mlozi MRS, Mlangwa JED, Willingham AL. A health-education intervention trial to reduce porcine cysticercosis in Mbulu District, Tanzania. *Prev Vet Med*. 2008;85(1-2):52-67. doi:10.1016/j.prevetmed.2007.12.014
R5	Rahme E, Choquette D, Beaulieu M, et al. Impact of a general practitioner educational intervention on osteoarthritis treatment in an elderly population. *Am J Med*. 2005;118(11):1262-1270. doi:10.1016/j.amjmed.2005.03.026
R6	Swanson KM, Chen H-T, Graham JC, Wojnar DM, Petras A. Resolution of Depression and Grief during the First Year after Miscarriage: A Randomized Controlled Clinical Trial of Couples-Focused Interventions. *J Women's Heal*. 2009;18(8):1245-1257. doi:10.1089/jwh.2008.1202
R7	Van Deurssen E, Meijster T, Oude Hengel KM, et al. Effectiveness of a Multidimensional Randomized Control Intervention to Reduce Quartz Exposure among Construction Workers. *Ann Occup Hyg*. 2015;59(8):959-971. doi:10.1093/annhyg/mev037
R8	Amza A, Kadri B, Nassirou B, et al. Community risk factors for ocular chlamydia infection in Niger: Pre-treatment results from a cluster-randomized trachoma trial. *PLoS Negl Trop Dis*. 2012;6(4). doi:10.1371/journal.pntd.0001586
R9	Hovi T, Ollgren J, Savolainen-Kopra C, T. H, J. O. Intensified hand-hygiene campaign including soap-and-water wash may prevent acute infections in office workers, as shown by a recognized-exposure -adjusted analysis of a randomized trial. *BMC Infect Dis*. 2017;17(1):47. doi:https://doi.org/10.1186/s12879-016-2157-z
R10	Barlis P, Regar E, Serruys PW, et al. An optical coherence tomography study of a biodegradable vs. durable polymer-coated limus-eluting stent: A LEADERS trial sub-study. *Eur Heart J*. 2010;31(2):165-176. doi:10.1093/eurheartj/ehp480
R11	See CW, O'Brien KS, Keenan JD, et al. The effect of mass azithromycin distribution on childhood mortality: Beliefs and estimates of efficacy. *Am J Trop Med Hyg*. 2015;93(5):1106-1109. doi:10.1111/sjos.12316
M1	Alexander N, Emerson P. Analysis of incidence rates in cluster-randomized trials of interventions against recurrent infections, with an application to trachoma. *Stat Med*. 2005;24(17):2637-2647. doi:10.1002/sim.2138
M2	Clark AB, Bachmann MO. Bayesian methods of analysis for cluster randomized trials with count outcome data. *Stat Med*. 2010;29(2):199-209. doi:10.1002/sim.3747
M3	Nixon RM, Duffy SW, Fender GR. Imputation of a true endpoint from a surrogate: Application to a cluster randomized controlled trial with partial information on the true endpoint. *BMC Med Res Methodol*. 2003;3:1-11. doi:10.1186/1471-2288-3-17
M4	Olsen MK, DeLong ER, Oddone EZ, Bosworth HB. Strategies for analyzing multilevel cluster-randomized studies with binary outcomes collected at varying intervals of time. *Stat Med*. 2008;27(29):6055-6071. doi:10.1002/sim.3446
M5	Thompson SG, Warn DE, Turner RM. Bayesian methods for analysis of binary outcome data in cluster randomized trials on the absolute risk scale. *Stat Med*. 2004; 23(3):389-410. doi:10.1002/sim.1567
M6	Turner RM, Prevost AT, Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med*. 2004; 23(8):1195-1214. doi:10.1002/sim.1721
M7	Turner RM, Omar RZ, Thompson SG. Modelling multivariate outcomes in hierarchical data, with application to cluster randomised trials. *Biometrical J*. 2006;48(3):333-345. doi:10.1002/bimj.200310147
M8	Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous responses. *Stat Med*. 2001;20(3):435-452. doi:10.1002/1097-0258(20010215)20:3<435::AID-SIM804>3.0.CO;2-E
M9	Kikuchi T, Gittins J. A behavioural Bayes approach for sample size determination in cluster randomized clinical trials. *J R Stat Soc Ser C Appl Stat*. 2010;59(5):875-888. doi:10.1111/j.1467-9876.2010.00732.x
M10	Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clin Trials*. 2005;2(2):108-118. doi:10.1191/1740774505cn072oa
M11	Turner RM, Omar RZ, Thompson SG. Constructing intervals for the intracluster correlation coefficient using Bayesian modelling, and application in cluster randomized trials. *Stat Med*. 2006;25(9):1443-1456. doi:10.1002/sim.2304
M12	Uhlmann L, Jensen K, Kieser M. Bayesian network meta-analysis for cluster randomized trials with binary outcomes. *Res Synth Methods*. 2016;8(October 2015):236-250. doi:10.1002/jrsm.1210
M13	Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med*. 2001;20(3):453-472. doi:10.1002/1097-0258(20010215)20:3<453::AID-SIM803>3.0.CO;2-L
C1	Peters TJ, Richards SH, Bankhead CR, Ades AE, Sterne JAC. Comparison of methods for analysing cluster randomized trials: An example involving a factorial design. *Int J Epidemiol*. 2003;32(5):840-846. doi:10.1093/ije/dyg228
C2	Pacheco GD, Hattendorf J, Colford JM, Mäusezahl D, Smith T. Performance of analytical methods for overdispersed counts in cluster randomized trials: Sample size, degree of clustering and imbalance. *Stat Med*. 2009;28(24):2989-3011. doi:10.1002/sim.3681
C3	Ma J, Thabane L, Kaczorowski J, et al. Comparison of Bayesian and classical methods in the analysis of cluster randomized controlled trials with a binary outcome: The community hypertension assessment trial (CHAT). *BMC Med Res Methodol*. 2009;9(1). doi:10.1186/1471-2288-9-37

Prefix "R" refers to results papers, "M" to methodological papers and "C" to comparison of methods papers

such estimates in informing the design of future studies. However, it was clear in six (85.7%) of the primary results papers how clustering was accounted for in the statistical analysis.

Reporting quality metrics have also been summarised by the following: (i) publication date before or after the publication of the CONSORT extension to CRCTs in 2012 [12]; (ii) journal endorsement of the CONSORT guidelines [12]; and (iii) involvement of a statistician in the study (Table 4). Due to the small number of available papers, we dichotomised journal endorsement of the CONSORT guidelines into “High” or “Medium” versus “Low” or “None”. We intended to summarise these results by three time periods (pre-2005, 2005–2012 and 2012–2018) to assess any effect of the publication of the CONSORT extensions for CRCTs in 2004 and 2012 on reporting quality. However, we were unable to identify any CRCTs using Bayesian methodology published before 2005. Pre-specified quality metrics are detailed in Table 4. However, due to the small number of primary results papers identified (seven in total), no meaningful comparisons can be made.

One of the papers retrieved was a pre-specified substudy and so was classified as a secondary results paper (Table 2, R10). We noted that reporting quality, despite not being a primary results paper and therefore not obligated to follow CONSORT guidelines, was high: a sample size calculation was presented and appropriately accounted for clustering, including specification of the assumed ICC; the flow of clusters and individuals through the study was well documented; and all levels of clustering were accounted for within a hierarchical modelling framework.

Use of Bayesian methodology

We were unable to identify any results papers in which a Bayesian approach was taken, or even discussed, for study design or sample size calculation. One secondary paper did, however, specify that the design factor used to inflate the sample size calculation was derived from the results of a Bayesian hierarchical model.

Of the eleven results papers included in the review, all adopted some form of Bayesian approach to statistical analysis (Table 5). In nine (81.8%; R1–R7, R9, R10) of the 11 papers, hierarchical modelling techniques were employed to account for the clustered structure of the data. Another study employed Bayes Model Averaging (R8) in order to mitigate the risks of overfitting that can be associated with stepwise regression in model-fitting. One study conducted a literature search of Cochrane Reviews and extracted the key summary statistic (mortality) before converting each into a log-odds ratio. These statistics were combined into a single arithmetic mean in order to construct an empirical prior. This prior was then combined with the likelihood from the CRCT to obtain a Bayesian posterior distribution of the relative risk of mortality in the intervention group versus the control group (R11).

In these results papers, prior distributions were informative in two (18.2%; R3, R11) papers; in one, (R3) “collateral” information from a previous study was used to construct a prior distribution for the variation in practice effects (specifically, the standard deviation for practice-level rates); in the other (R11) an informative prior distribution for the treatment effect parameter within a negative binomial regression was constructed based on a meta-analysis of relevant reviews obtained from the Cochrane library, and used to inform the estimation of the outcome of interest (the relative risk of childhood mortality). No information was provided on the prior distributions placed on the variance components. Weakly informative prior distributions were used in one (9.1%; R2) study, by placing Student’s *t* priors centred at 0 on the treatment effect parameter and other fixed logistic regression coefficients, which the authors acknowledged would only affect inference if the data provide little information about the parameters. No detail was provided on the prior distributions specified for the variance components in this paper. Five (45.5%; R1, R3, R5, R9, R10) papers specified the use of non-informative prior distributions, although only one of these (R5) provided more specific detail, stating normal prior distributions for the treatment effect and each of the fixed logistic regression coefficients, and uniform prior distributions for the variance components. Four studies (36.4%; R4, R6, R7, R8) did not specify their choice of prior distribution. One paper fitted two Bayesian models (R3) - one model implementing a non-informative prior and the other utilising “collateral” information, so we recorded the use of both an informative and a non-informative prior.

Bayesian methodological developments

We categorised 13 (48.1%) of the 27 papers included as methodological papers, where the focus was on the development of Bayesian methods for use in the design or analysis of CRCTs, as opposed to applying existing methods to data from CRCTs. Of these 13 papers, we defined 11 (84.6%) as “pure” methods papers, in which Bayesian methodological developments are reported independently of an applied scenario (although study data may have been used to demonstrate the method). We categorised two (15.4%) papers as being methodological but with the developments being driven by a specific statistical problem encountered in a CRCT, in which the method is presented and subsequently used to analyse the data of interest. Finally, we categorised three (11.1%) of the 27 papers as comparison of methods papers, in which existing methodology (both Bayesian and frequentist) were applied to the same data for comparative purposes.

Of the 11 “pure” methodological papers, seven presented analysis methods (63.6%; M2, M4, M5, M7, M11, M12, M13), two presented methods for design/sample size calculation (18.2%; M6, M9) and two presented elements of both (18.2%; M8, M10). Both papers driven by

Table 3 Demographic data for the eleven results papers

N (%) unless otherwise stated	Total (*N* = 11)	Primary (*N* = 7)	Secondary (*N* = 4)
Year of publication			
Pre 2005	0 (0.0)	0 (0.0)	0 (0.0)
2005–2012	6 (54.5)	4 (57.1)	2 (50.0)
Post 2012	5 (45.5)	3 (42.9)	2 (50.0)
Location of first author[a]			
UK	2 (18.2)	1 (14.3)	1 (25.0)
US/Canada	5 (45.5)	4 (57.1)	1 (25.0)
Europe excl. UK	3 (27.3)	1 (14.3)	2 (50.0)
Australia/New Zealand	0 (0.0)	0 (0.0)	0 (0.0)
Africa	2 (18.2)	1 (14.3)	1 (25.0)
Asia	0 (0.0)	0 (0.0)	0 (0.0)
Other	0 (0.0)	0 (0.0)	0 (0.0)
Location of study[a]			
UK	1 (9.1)	0 (0.0)	1 (25.0)
US/Canada	3 (27.3)	3 (42.9)	0 (0.0)
Europe excl. UK	4 (36.4)	2 (28.6)	2 (50.0)
Australia/New Zealand	0 (0.0)	0 (0.0)	0 (0.0)
Africa	4 (36.4)	2 (28.6)	2 (50.0)
Asia	0 (0.0)	0 (0.0)	0 (0.0)
Other	0 (0.0)	0 (0.0)	0 (0.0)
Target sample size; mean (SD) [range]	N/A	*N* = 3[b] 1466.7 (1868.6) [120, 3600]	N/A
Target number of clusters; mean (SD) [range]	N/A	*N* = 2[c] 200.0 (198.0) [60, 340]	N/A
Recruited Sample Size; mean (SD) [range]	*N* = 11 10898.5 (19816.1) [116, 66204]	*N* = 7 2484.6 (3700.1) [116, 9928]	*N* = 4 25662.8 (28762.5) [683, 66204]
Recruited Number of Clusters; mean (SD) [range]	*N* = 11 58.8 (95.6) [5, 341]	*N* = 7 69.1 (121.6) [5, 341]	*N* = 4 40.8 (13.2) [21, 48]
Randomisation unit			
Medical facility	1 (9.1)	1 (14.3)	0 (0.0)
Village/community/district	6 (54.5)	4 (57.1)	2 (50.0)
Organisation	1 (9.1)	1 (14.3)	0 (0.0)
Couple	1 (9.1)	1 (14.3)	0 (0.0)
Individual	1 (9.1)	0 (0.0)	1 (25.0)
Working unit (office)	1 (9.1)	0 (0.0)	1 (25.0)
Primary outcome type			
Binary	9 (81.8)	5 (71.4)	4 (100.0)
Continuous	2 (18.2)	2 (28.6)	0 (0.0)
Statistician involvement	8 (72.7)	5 (71.4)	3 (75.0)
Statistician association			
Clinical trials unit	1 (12.5)	0 (0.0)	1 (33.3)
Academic statistical department	7 (87.5)	5 (100.0)	2 (66.6)
Commercial pharmaceutical company	0 (0.0)	0 (0.0)	0 (0.0)

Table 3 Demographic data for the eleven results papers *(Continued)*

N (%) unless otherwise stated	Total (*N* = 11)	Primary (*N* = 7)	Secondary (*N* = 4)
Clinical research organisation	0 (0.0)	0 (0.0)	0 (0.0)
Other	0 (0.0)	0 (0.0)	0 (0.0)
Journal endorsement of CONSORT guidelines			
High	N/A	3 (42.9)	N/A
Medium	N/A	1 (14.3)	N/A
Low	N/A	0 (0.0)	N/A
None	N/A	3 (42.9)	N/A

[a] One author was associated with an institution in both Europe and the UK, and the associated study was run across both locations. The denominator used for the calculations is based on the number of papers

[b] Two studies specified the number of participants approached but these were not explicitly stated/justified recruitment targets and so were excluded

[c] Four studies specified the number of clusters approached but these were not explicitly stated/justified recruitment targets and so were excluded

specific application presented analysis methods (M1, M3).

The analysis methods papers predominantly presented hierarchical modelling methodology applied to dealing with a range of data types, such as incidence rates (M1), count data (M2) and binary data (M4, M5,M13), in a Bayesian setting, citing flexibility of modelling and the ability to incorporate prior information and account for the complex variance structures as key advantages. One paper reports Bayesian methods for modelling multivariate outcomes (M7), which allow for multiple outcomes without concern for multiplicity whilst accommodating complex correlation structures. Another paper presents Bayesian network meta-analysis methods for CRCTs (M12), allowing for comparison of multiple treatment arms whilst accounting for the complex correlation structure inherent in clustered data.

A number of methodological papers identified within our review focus on the ICC. One such paper centres on analysis only, presenting methods for constructing intervals for the ICC and suggesting prior distributions for use in modelling (M11). The two papers in which both design and analysis are discussed focus heavily on the ICC; one provides a range of options for choice of prior distribution alongside recommendations, before discussing briefly how the uncertainty in the ICC can be accounted for in sample size calculations (M8). The other paper presents methods for formulating prior distributions for use in sample size calculations and statistical analysis on the basis of multiple previous estimates, whilst incorporating the relevance of the studies from which they were obtained (M10). One of the papers presenting only study design methodology also focused on ICCs, and developed methods to formulate prior distributions from single and multiple previous ICC estimates for use in sample size calculations (M6).

The remaining study design paper presented a behavioural Bayes approach (M9), extending existing methodology [26–29] for sample size determination in individually randomised trials to CRCTs. The method incorporates estimated financial costs and benefits of the intervention to produce a net benefit, rather than being based on the more usual difference in primary outcome alone.

Discussion

To the best of our knowledge, this is the first methodological systematic review of the use, or consideration of, Bayesian methods in CRCTs.

As the number of included papers is small, drawing robust conclusions regarding overall reporting quality between subgroups (Table 4) is not possible. However, in 2013, Diaz-Ordaz presented a summary of reviews of CRCT quality, in which the percentage of studies accounting for clustering in the sample size calculation and statistical analysis ranged from 0% to 71% and 37% to 92%, respectively [15]. We have identified an additional review of reporting and methodological quality of CRCTs published in 2016 [16]. Including the data from the more recent review together with Diaz-Ordaz's summary, the mean (SD) percentage of studies accounting for clustering in the sample size calculation and analysis was 34.6 (23.7) and 64.2 (16.3), respectively. For comparison, our study identified no papers which clearly accounted for clustering in the sample size calculation, and six (85.7%) papers accounting for clustering in the analysis. Although our review included only a small number of papers, reporting quality according to these key metrics may differ somewhat between studies using Bayesian methodology and the wider pool of CRCTs, as none of the papers we identified clearly accounted for clustering in sample size calculation. Hence, there is a need to further improve the reporting of CRCTs utilising Bayesian methodology. Conversely, Bayesian CRCTs seem to more often account for clustering in analysis. This is likely due to the popularity of Bayesian hierarchical modelling within the set of included papers, which is a natural way to conduct mixed or random effects modelling and therefore inherently account for clustering.

Table 4 Reporting quality metrics for seven primary results papers

Reporting quality criteria *N* (%)	**Total (*N* = 7)**	**Year of publication**		**Journal endorsement of CONSORT guidelines**		**Statistician involvement**	
		2012 or earlier (*N* = 4)	**2013 onwards (*N* = 3)**	**High/medium (*N* = 4)**	**Low/none (*N* = 3)**	**Yes (*N* = 5)**	**No (*N* = 2)**
Description of sample size method	4 (57.1)	2 (50.0)	2 (66.7)	2 (50.0)	2 (66.7)	2 (40.0)	2 (100.0)
Was clustering clearly accounted for in sample size calculation	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Specification of the required number of clusters	2 (50.0)	1 (50.0)	1 (50.0)	1 (50.0)	1 (50.0)	1 (50.0)	1 (50.0)
Specification of the assumed cluster size	2 (50.0)	1 (50.0)	1 (50.0)	1 (50.0)	1 (50.0)	1 (50.0)	1 (50.0)
Specification of whether equal or unequal cluster sizes are assumed	1 (25.0)	1 (50.0)	0 (0.0)	0 (0.0)	1 (50.0)	0 (0.0)	1 (50.0)
Variability in cluster size accounted for	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Specification of the ICC used for the sample size	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Indication of the uncertainty of the ICC	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Accounted for the uncertainty in the ICC	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Other CONSORT metrics							
Details of how clustering was accounted for in the analysis	6 (85.7)	4 (100.0)	2 (66.7)	4 (100.0)	2 (66.7)	5 (100.0)	1 (50.0)
Specification of the number of clusters randomised	7 (100.0)	4 (100.0)	3 (100.0)	4 (100.0)	3 (100.0)	5 (100.0)	2 (100.0)
Specification of the number of clusters receiving intended treatment							
Explicit	5 (71.4)	3 (75.0)	2 (66.7)	4 (100.0)	1 (33.3)	4 (80.0)	1 (50.0)
Implied	2 (28.6)	1 (25.0)	1 (33.3)	0 (0.0)	2 (66.7)	1 (20.0)	1 (50.0)
Specification of the number of clusters analysed for the primary outcome at the primary endpoint							
Explicit	2 (28.6)	1 (25.0)	1 (33.3)	2 (50.0)	0 (0.0)	2 (40.0)	0 (0.0)
Implied	5 (71.4)	3 (75.0)	2 (66.7)	2 (50.0)	3 (100.0)	3 (60.0)	2 (100.0)
Details of cluster-level losses and exclusions							
Explicit	3 (42.9)	2 (50.0)	1 (33.3)	2 (50.0)	1 (33.3)	2 (40.0)	1 (50.0)
Implied	4 (57.1)	2 (50.0)	2 (66.7)	2 (50.0)	2 (66.7)	3 (60.0)	1 (50.0)
Details of individual-level losses and exclusions	4 (57.1)	2 (50.0)	2 (66.7)	2 (50.0)	2 (66.7)	2 (40.0)	2 (100.0)
Individual-level baseline characteristics presented	7 (100.0)	4 (100.0)	3 (100.0)	4 (100.0)	3 (100.0)	5 (100.0)	2 (100.0)
Cluster-level baseline characteristics presented	2 (28.6)	2 (50.0)	0 (0.0)	1 (25.0)	1 (33.3)	1 (20.0)	1 (50.0)
Coefficients of intracluster correlation provided for primary outcomes							
All	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Some	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Coefficients of intracluster correlation provided for secondary outcomes							
All	0 (0.0)[a]	0 (0.0)	0 (0.0)[a]	0 (0.0)	0 (0.0)[a]	0 (0.0)	0 (0.0)[a]
Some	0 (0.0)[a]	0 (0.0)	0 (0.0)[a]	0 (0.0)	0 (0.0)[a]	0 (0.0)	0 (0.0)[a]

Table 4 Reporting quality metrics for seven primary results papers *(Continued)*

Reporting quality criteria *N* (%)	Total (*N* = 7)	Year of publication		Journal endorsement of CONSORT guidelines		Statistician involvement	
		2012 or earlier (*N* = 4)	2013 onwards (*N* = 3)	High/medium (*N* = 4)	Low/none (*N* = 3)	Yes (*N* = 5)	No (*N* = 2)
***P*-values provided for baseline comparisons**	5 (71.4)	3 (75.0)	2 (66.7)	3 (75.0)	2 (66.7)	3 (60.0)	2 (100.0)
Clustering accounted for in the calculation of the *p*-values							
Yes	1 (20.0)	1 (33.3)	0 (0.0)	1 (33.3)	0 (0.0)	1 (33.3)	0 (0.0)
Unclear	1 (20.0)	1 (33.3)	0 (0.0)	1 (33.3)	0 (0.0)	1 (33.3)	0 (0.0)

[a] One study did not have any secondary outcomes

Evidently, the use of Bayesian methods in the design or analysis of CRCTs remains uncommon relative to the use of frequentist methods (Fig. 1), with only eleven primary or secondary results papers reporting doing so. This is despite the increasing use of CRCT designs, with over 120 reported in 2008 alone [4] and the number of PubMed search results rising almost year-on-year since 2006 (Fig. 1) reaching 347 in 2018. This methodological systematic review failed to identify a single reported CRCT which utilised a Bayesian approach to conduct the sample size calculation, despite some efforts to develop methodology in this area, as highlighted in the methodological aspect of our review. Whilst explaining the reason for this lack of uptake of Bayesian methodology in the design of CRCTs would be little more than speculation, possibilities include fundamental disagreements with the approach, still limited development of methodology, inaccessibility of software to implement the methods or lack of knowledge or understanding. Whilst we have shown that there has been some Bayesian methodological developments in both design and analysis of CRCTs, these have been limited in comparison to the development of classical methods which are now well-established in the literature. None of the thirteen published methodological papers appears to have developed publicly available software in order to aid implementation (although some papers reported that code is available from the authors on request), whereas classical analysis and sample size calculations for CRCTs can be conducted with relative ease in standard statistical software. As such, there is need to increase the availability and accessibility of these methods, which can offer advantages over the frequentist approach within the CRCT context.

A common criticism of the Bayesian approach in general, and in particular within the analysis of clinical trial data, is the subjective nature of the choice of prior distribution, although it is strongly recommended that sensitivity analyses be performed in order to assess the strength of the effect of the prior [30]. Interestingly, however, only two (18.2%) of the 11 results papers that were identified utilised an informative prior distribution, and one (9.1%) utilised a weakly informative prior. Five (45.5%) specified an uninformative prior (of which one employed two models). It is likely that the four (36.4%) papers that did not report their choice of prior used an uncontroversial, uninformative formulation, and in doing so, a likely total of nine (81.8%) studies

Table 5 Summary of Bayesian Methods used in primary and secondary results papers

N (%)	Total (*N* = 11)	Primary (*N* = 7)	Secondary (*N* = 4)
Sample Size (used)	0 (0.0)	0 (0.0)	0 (0.0)
Sample Size (discussed)	0 (0.0)	0 (0.0)	0 (0.0)
Analysis (used)	11 (100.0)	7 (100.0)	4 (100.0)
Priors used			
Informative	2 (18.2)[a]	1 (14.3)[a]	1 (25.0)
Weakly Informative	1 (9.1)	1 (14.3)	0 (0.0)
Non-informative	5 (45.5)[a]	3 (42.9)[a]	2 (50.0)
Unspecified	4 (36.4)	3 (42.9)	1 (25.0)
Analysis (discussed)	N/A	N/A	N/A

[a] One paper reported the use of two Bayesian models — the first model implementing a non-informative prior and the second model utilising "collateral" information

circumvented the perceived issues surrounding the choice of an informative prior. Despite this, the use of a well-justified, informative prior distribution has the potential to add value to a statistical analysis, and methodological development for informative yet rigorous prior specification for CRCTs may enhance the uptake of Bayesian methods in this area.

Strengths and limitations

A protocol for this methodological systematic review was published before commencement of the electronic search [17] and the review was conducted according to the PRISMA guidelines [18]. The electronic search strategy to identify Bayesian approaches in CRCTs was adapted from a previously published strategy, which was demonstrated to have high precision [20] in identifying CRCTs. In this study, each stage of the reference sifting and data extraction process was fully conducted twice, independently, to ensure accurate inclusion of references and high-quality data for examination. We developed data extraction forms for primary and secondary results papers in order to aid in the accurate and consistent collection of data. Furthermore, the final data extraction was agreed by all four members of the study team.

The reporting quality metrics collected are predominantly a subset of the CONSORT checklist for CRCTs, a well-accepted set of criteria. We added a small number of additional items such as whether cluster size variability had been accounted for in the presented sample size calculation [4] and whether *p*-values for baseline comparisons were provided, in order to facilitate a robust judgement of reporting quality.

Despite this, we acknowledge the possibility that we may have missed some publications in which Bayesian methodology was used or considered in the design or analysis of CRCTs. In particular, we opted for a search strategy in which specificity was maximised, rather than sensitivity, in order to make the sifting process more manageable with limited resource. We added six additional methodological papers through hand searching, but were unable to identify any additional trial results papers. This is not surprising given the search strategy was developed to identify the latter, but may suggest a greater risk that further methodological papers have been missed compared to trial results papers.

Furthermore, we present reporting quality metrics by journal endorsement of the CONSORT guidelines. However, we acknowledge that the guidelines may, in some cases, have changed since the date of the associated publications, and as a result, a journal's endorsement may have been intensified since the included papers were accepted for publication. To the best of our knowledge, this issue has not been raised in previous systematic reviews of trial reporting quality; archiving of journal guidelines would help researchers conducting quality assessment systematic reviews in the future. Similarly, we sought to identify author affiliations during data collection, but again acknowledge that these may have changed since publication of the research, particularly for papers published some time ago.

We intended to summarise the pre-specified reporting quality metrics by time periods (pre-2005, 2005–2012 and 2012–2018) according to publication date to assess the effect of the relevant CONSORT statements on reporting quality. We acknowledge that the time delay between completion of the study and submission of the final report for publication may have resulted in some studies being categorised as published after the publication of the CONSORT extension guidance, when in fact it was designed, conducted and possibly even analysed before.

Conclusion

The use of Bayesian methods in the statistical analysis of CRCTs is rare and was not found at all in the design of any of the reviewed studies or their sample size calculations. There have been some developments in Bayesian methodology for CRCTs but far less so than within the frequentist paradigm. Reporting quality may differ between CRCTs utilising Bayesian methodology compared with previous reviews of CRCT quality, although the number of papers identified in this review is small. There is a need for further Bayesian methodological developments in the design and analysis of CRCTs, including approaches for the specification of prior distributions, as well as statistical software development to allow easier implementation of methods, in order to increase the accessibility, availability and, ultimately, use of the approach.

Abbreviations

CENTRAL: Cochrane Central Register of Controlled Trials; CRCT: Cluster randomised controlled trial; CRO: Clinical research organisation; CTU: Clinical trials unit; ICC: Intra-class correlation coefficient; RCT: Randomised controlled trial

Acknowledgements

We thank Amanda Wanner, Information Specialist, for her advice during the conceptualisation of the project and the development of the search strategy.

Authors' contributions

BJ led on the conception and design of the study with contributions from the other authors. BJ conducted all reference sifting and extraction, and SC, AB and AS were each involved in part of the independent sift and data extraction, with RM providing independent input on Bayesian methodological developments and implementation in particular. BJ conducted the data analysis and drafted the manuscript. All authors have read and approved the final manuscript.

Funding

This project was completed as part of a PhD programme, for which funding was provided by The University of Plymouth.

Availability of data and materials

The datasets generated and/or analysed during the study are available on request.

Declarations

Ethical approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing Interests

The authors declare that they have no competing interests.

Author details

[1]Medical Statistics, Faculty of Health: Medicine, Dentistry and Human Sciences, University of Plymouth, Room N15, ITTC Building 1, Plymouth Science Park, Plymouth, Devon PL6 8BX, UK. [2]NIHR ARC South West Peninsula (PenARC), College of Medicine and Health, University of Exeter, Exeter, Devon, UK. [3]Klinische Epidemiologie, Institut für Epidemiologie und Sozialmedizin, Westfälische Wilhelms-Universität Münster, Münster, Germany. [4]School of Computing, Electronics and Mathematics, Faculty of Science and Engineering, University of Plymouth, Plymouth, Devon, UK. [5]Peninsula Clinical Trials Unit, Faculty of Health: Medicine, Dentistry and Human Sciences, University of Plymouth, Plymouth, Devon, UK. [6]Exeter Clinical Trials Unit, College of Medicine and Health, University of Exeter, Exeter, Devon, UK.

Received: 14 August 2019 Accepted: 11 March 2021
Published online: 31 March 2021

References

1. Eldridge SM, Kerry S. A Practical guide to cluster randomised trials in health services research: Wiley; 2012.
2. Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. BMC Med Res Methodol. 2004;4(1):21. https://doi.org/10.1186/1471-2288-4-21.
3. Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. Int J Epidemiol. 1990;19(4):795–800. https://doi.org/10.1093/ije/19.4.795.
4. Moberg J, Kramer M. A brief history of the cluster randomised trial design. J R Soc Med. 2015;108(5):192–8. https://doi.org/10.1177/0141076815582303.
5. Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous responses. Stat Med. 2001;20(3):435–52. https://doi.org/10.1002/1097-0258(20010215)20:3<435::AID-SIM804>3.0.CO;2-E.
6. Turner RM, Prevost AT, Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. Stat Med. 2004;23(8):1195–214. https://doi.org/10.1002/sim.1721.
7. Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. Clin Trials. 2005;2(2):108–18. https://doi.org/10.1191/1740774505cn072oa.
8. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond " p < 0.05." Am Stat. 2019;73(sup1):1-19. doi:https://doi.org/10.1080/00031305.2019.1583913
9. Lewis RJ, Wears RL. An introduction to the Bayesian analysis of clinical trials. Ann Emerg Med. 1993;22(8):1328–36. https://doi.org/10.1016/S0196-0644(05)80119-2.
10. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. Third.; 2014. https://www.crcpress.com/Bayesian-Data-Analysis/Gelman-Carlin-Stern-Dunson-Vehtari-Rubin/p/book/9781439840955
11. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. Bmj. 2004;328(7441):702–8. https://doi.org/10.1136/bmj.328.7441.702.
12. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345(7881). https://doi.org/10.1136/bmj.e5661.
13. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. J Clin Epidemiol. 2015;68(6):716–23. https://doi.org/10.1016/j.jclinepi.2014.10.006.
14. Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. BMJ. 2011;343(sep26 1):d5886. https://doi.org/10.1136/BMJ.D5886.
15. Diaz-Ordaz K, Froud R, Sheehan B, Eldridge S. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. BMC Med Res Methodol. 2013;13(1):127. https://doi.org/10.1186/1471-2288-13-127.
16. Tokolahi E, Hocking C, Kersten P, Vandal AC. Quality and reporting of cluster randomized controlled trials evaluating occupational therapy interventions: a systematic review. OTJR Occup Particip Heal. 2016;36(1):14–24. https://doi.org/10.1177/1539449215618625.
17. Jones B. The use of Bayesian statistics in the design and analysis of cluster randomised controlled trials and their methodological and reporting quality: a protocol for an international methodological review. Open Science Framework. Published 2018. https://osf.io/2azrc/
18. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7):e1000097. https://doi.org/10.1371/journal.pmed.1000097.
19. Hemming K, Taljaard M, McKenzie JE, et al. Reporting of stepped wedge cluster randomised trials: Extension of the CONSORT 2010 statement with explanation and elaboration. BMJ. 2018;363:1614. https://doi.org/10.1136/bmj.k1614.
20. Taljaard M, McGowan J, Grimshaw JM, Brehaut JC, McRae A, Eccles MP, et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: Low precision will improve with adherence to reporting standards. BMC Med Res Methodol. 2010;10(1). https://doi.org/10.1186/1471-2288-10-15.
21. Elsevier. Mendeley. https://www.mendeley.com/
22. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210. https://doi.org/10.1186/s13643-016-0384-4.
23. Froud R, Eldridge S, Diaz Ordaz K, Marinho VCC, Donner A. Quality of cluster randomized controlled trials in oral health: A systematic review of reports published between 2005 and 2009. Community Dent Oral Epidemiol. 2012;40(SUPPL. 1):3–14. https://doi.org/10.1111/j.1600-0528.2011.00660.x.
24. Delgado-Rodriguez M, Ruiz-Canela M, De Irala-Estevez J, Martinez-Gonzalez A, Llorca J. Participation of epidemiologists and/or biostatisticians and methodological quality of published controlled clinical trials. J Epidemiol Community Health. 2001;55(8):569–72. https://doi.org/10.1136/jech.55.8.569.
25. Dechartres A, Charles P, Hopewell S, Ravaud P, Altman DG. Reviews assessing the quality or the reporting of randomized controlled trials are increasing over time but raised questions about how quality is assessed. J Clin Epidemiol. 2011;64(2):136–44. https://doi.org/10.1016/j.jclinepi.2010.04.015.
26. Pezeshk H, Gittins J. A fully bayesian approach to calculating sample sizes for clinical trials with binary responses. Ther Innov Regul Sci. 2002;36(1):143–50. https://doi.org/10.1177/009286150203600118.
27. Gittins JC, Pezeshk H. A decision theoretic approach to sample size determination in clinical trials. J Biopharm Stat. 2002;12(4):535–51. https://doi.org/10.1081/BIP-120016234.
28. Gittins J, Pezeshk H. How large should a clinical trial be? J R Stat Soc Ser D Stat. 2000;49(2):177–87. https://doi.org/10.1111/1467-9884.00228.
29. Gittins J, Pezeshk H. A behavioral bayes method for determining the size of a clinical trial. Ther Innov Regul Sci. 2000;34(2):355–63. https://doi.org/10.1177/009286150003400204.
30. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. (Senn S, Barnett V, eds.). John Wiley & Sons, Ltd; 2004.

Publisher's Note