The University of Sheffield

A comparison of statistical methods for analysing cluster randomised controlled trials: *classical vs emerging methods*

By:

Bright C. Offorha

A thesis submitted to the University of Sheffield in partial fulfilment of the requirements for the degree of Doctor of Philosophy

University of Sheffield
Faculty of Health
Division of Population Health
School of Medicine and Population Health

Supervised by: Professor Stephen J. Walters and Dr Richard M. Jacques

August 2023

Acknowledgments

I would like to express deep appreciation to my supervisors, Professor Stephen J. Walters, and Dr Richard M. Jacques, for their support and guidance throughout my Ph.D. programme. I would also like to thank Professor Steven A. Julious and Dr Laura J. Sutton for chairing my first-year confirmation review, their suggestions guided me subsequently for the rest of my Ph.D. programme. I wish to thank the Chief Investigator of Bridging the Age Gap in breast cancer trial, Professor Lynda Wyld, Department of Oncology and Metabolism, University of Sheffield Medical School, Sheffield, UK, for permitting me to use the Bridging the Age Gap trial data.

I am grateful to the Nigeria Federal Government which sponsored my Ph.D. study through her agency – the Tertiary Education Trust Fund (TETFUND). Thanks to my employer, Abia State University, Uturu for releasing me to pursue my academic goals. I would also like to extend my gratitude to Dr Ruth Wong, an exceptional Information Specialist, who guided me in developing the search terms for the methodological scoping review conducted in this thesis.

I am grateful to my parents for their prayers and moral support that has kept me going. Massive thanks to my wife and son who have been my major support system from the beginning to the end of my studies. Thanks for keeping the family together, listening to my worries, and celebrating my little wins.

My deepest gratitude goes to my friends and fellow Ph.D. colleagues Dr. Ayodeji Oyedeji, Dr. Chinyere, Dr. Chinyereugo M. Umemneku, Dr. Yirui Qian, and Dr. Nahid Sultana for their academic support and brainstorming. Thanks to my football mates for providing the needed social distractions for a balanced lifestyle. Lastly, I am grateful to the pastor and members of the Ushering unit of Winners' Chapel International Church, Sheffield for their prayers and encouragement.

Declarations

I declare that this research or part of it has not been submitted for a different degree at any other University or the University of Sheffield. I declare that this research consists of my original works toward my Ph.D. degree.

Abstract

Introduction

Cluster randomised controlled trials (cRCTs) entails randomising groups of individuals, such as schools, care homes, hospital wards, and general practices to the treatment arms. The outcomes within a cluster are likely to be correlated. The chosen analytical approach must consider this correlation to obtain valid results. Ignoring the correlated outcomes by using standard statistical methods that treat the outcomes as being independent, may lead to underestimating the standard errors (SEs) of the parameter estimates and consequently obtaining narrower confidence intervals (CIs), false small P-values, and incorrectly overstating the effect of the intervention. The following research question were conceptualised to explore the statistical methods used to analyse outcome data from cRCTs 1) What are the appropriate, and available methods in the literature for analysing outcome data from cRCTs? 2) What statistical methods are used in practice for analysing outcome data from cRCTs? 3) What criteria should be used in deciding the appropriateness of the identified methods? 4) How well do the selected methods perform, when compared?

Methods

I conducted a methodological scoping review involving a systematic search of the online bibliography databases of MEDLINE, EMBASE, PsycINFO, CINAHL, and SCOPUS and a practice review involving a chronological search of the online table of contents of the National Institute for Health and Care Research (NIHR) Journal Library to identify gaps in knowledge, and four analytical approaches (GzLMM, GEE1, GEE2, and QIF) were identified. The methods were applied to four cRCT datasets with continuous and binary outcomes. Furthermore, three of the four methods (GzLMM, GEE1, QIF) were applied to simulated continuous outcome datasets.

Results

The methodological review identified 27 unique analytical methods. In the practice review, from the 79 included cRCT reports with 86 independent trials and 100 primary outcomes analysed the observed median intracluster correlation coefficient (ICC) was 0.02, of which 4 in 10 trials did not report the observed ICC. This act goes against the recommendations of the Consolidated

Standards of Reporting Trials (CONSORT) reporting guidelines. The analysis of the four example datasets with clusters ranging from 10 to 100, and individual participants ranging from 748 to 9,207 showed that the estimates of the treatment effect (and associated standard error, confidence interval, and P-value) from the methods were equivalent in most cases. However, in a few analyses, the QIF produced different results compared to the other three methods, especially in trials with small to moderate numbers of clusters. The estimates from GEE1 and GEE2 were the same, except in their estimates of the ICC, hence, GEE2 was dropped from further investigations. A simulation study involving continuous outcome data shows that GzLMM, GEE1, and QIF performed equivalently based on bias, empirical standard error, and mean square error. The number of clusters N, cluster sizes n_i, ICC ρ and effect sizes θ had no impact on these results. With regards to coverage, Type I error rate, and power, GzLMM (with identity link function and parameters estimated by MLE) performed better than GEE1 and QIF when the ICC is low. For moderate ICC, appropriate small sample correction should be applied in conjunction with the chosen method when the clusters are fewer than fifty.

Conclusions

The planning of cRCTs should consider the hierarchical nature of cRCT design in the sample size calculation. Adherence to the reporting guidelines of CONSORT with extension to cRCTs is suboptimal based on the reporting of the observed ICC. Researchers, peer reviewers, and editors should make efforts to improve on this. In most cases investigated, the GzLMM performed better than GEE1 and QIF, however, other factors should be considered in choosing the appropriate analytical method, such as the estimand and scientific question of interest. QIF have no advantage over GEE1, hence, the current practice should be maintained.

Table of Contents

Abbreviations

A(Q)IC	Akaike (Quasi-likelihood) Information Criterion
3EE	Three Estimating Equations
ABC-C	Aberrant Behaviour Checklist – Community
AGHQ	Adaptive Gauss-Hermite Quadrature
ALR	Alternating Logistic Regression
ANOVA	Analysis of Variance
AUGEE-IPW	Augmented Generalized Estimating Equations – Inverse Probability Weighting
BIC	Bayesian Information Criterion
CI	Confidence Interval
CLA	Cluster Level Analysis
CONSORT	Consolidated Standard for Reporting Trials
CPr	Coverage Probability
cRCT	Cluster Randomised Controlled Trial
CSM	Cluster-Specific Model
CV	Coefficient of Variation
DE	Design Effect
DESI	Decision support intervention
DGM	Data Generating Mechanism
DoF	Degree of Freedom
EME	Efficacy and Mechanism Evaluation
EORTC	European Organisation for the Research and Treatment of Cancer
EPDS	Edinburgh Postnatal Depression Scale
ESE	Empirical Standard Error
FG	Fay and Graubard
GEE1	First-Order Generalized Estimating Equations
GEE2	Second order Generalized Estimating Equations
GLS	Generalized Least Squares
GMM	Generalized Method of Moment
GP	General Practitioner
GzLM	Generalized Linear Model
GzLMM	Generalized Linear Mixed Model
HBV	Hepatitis B Vaccination
HL	Hierarchical Likelihood
HLA	Hierarchical Likelihood – Laplace
HPC	High Performance Computing
HSDR	Health and Social Care Delivery Research
HTA	Health Technology Assessment
IC	Informed Choice
ICC	Intracluster Correlation Coefficient

ICTMC	International Clinical Trials Methodology Conference
ILA	Individual Level Analysis
IQR	Interquartile Range
IRLS	Iteratively Reweighted Least Squares
ISRCTN	International Standardised Randomised Controlled Trial Number
LM	Linear Model
LMM	Linear Mixed Model
LRT	Likelihood Ratio Test
MAR	Missing At Random
MCMC	Markov Chain Monte Carlo
MCMC	Markov Chain Monte Carlo
MCR	Missing Completely at Random
MCSE	Monte Carlo Standard Error
MGF	Moment Generating Function
mGLM	Marginal Generalized Linear Model
MIDIRS	Midwives' Information and Resource Service
MLE	Maximum Likelihood Estimation
MNAR	Missing Not At Random
MSB	Between clusters Mean Square error
MSE	Mean Square Error
MSW	Within clusters Mean Square error
MVPA	Moderate-to-Vigorous Physical Activity
NIHR	National Institute of Health and Research
NOSH	Nourishing Start for Health
NR	Not Reported
NTCT	National Clinical Trial
OLS	Ordinary Least Squares
OR	Odds Ratio
OSF	Open Science Framework
PAM	Population average Model
PAM	Population Average Models
PGfAR	Programme Grants from Applied Research
PHR	Public Health Research
PICOTS	Population Intervention, Comparator, Outcome, Timing, Setting
PL	Pseudo Likelihood
PLML	Pseudo Maximum Likelihood
PLRS	Pseudo-Likelihood – Risk Set
PND	Postnatal Depression
PQL	Penalized Quasi Likelihood
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRO	Patient Reported Outcome
PROC	Procedure
QIF	Quadratic Inference Function

QoL	Quality of Life
RCT	Randomised Controlled Trial
REML	Restricted Maximum Likelihood Estimation
RMRSE	Regression Model with Robust Standard Error
SAT	Satterthwaite
ScR	Scoping Review
SD	Standard Deviation
SE	Standard Error
TMLE	Targeted Maximum Likelihood
WHO	World Health Organisation

Notations

Subscripts

i	Represent a cluster; $1,2,\dots,N$
j	Represent an individual; $1,2,\dots,n$
h	Represents basis matrix; $0,\dots,m$
p	Represents vector of covariates; $0,1,\dots,p$

Hypothesis testing

α	Type I error rate
β	Type II error rate
$pr(\alpha)$	Probability of committing a Type I error
$pr(\beta)$	Probability of committing a Type II error
H_0, H_1	Null and alternative hypotheses
$\hat{\theta}_{ll}, \hat{\theta}_{ul}$	Estimates of lower and upper limits of the estimated confidence interval

Statistics, parameters, and outcomes

σ^2	Total variance
σ_b^2	Between cluster variance
σ_w^2	Within cluster variance
A, B	Groups A and B
ρ	True intracluster correlation coefficient
$R(a)$	Working correlation matrix
$l(.)$	Full likelihood function
$\bar{n}$	Average sample size
$\bar{y}$	Sample mean

Clusters

n_i	ith cluster size
n_A, n_B	Sample sizes of groups A and B
N	Total number of clusters
n	Total sample size
CV	Coefficient of variation for the cluster sizes
n_{sim}	Number of simulations
p_{ij}	Probability of the outcome for a jth individual in the ith cluster

z_q	The $q\%$ critical value of a standard Normal distribution

Regression model

β	Regression model coefficient
y	Single outcome
x	Single covariate
$\hat{V}$	Variance estimator
ϕ	Scale parameter
θ	Treatment effect
τ	Cluster level error term
e	Individual level error term
$\eta(.)$	Link function
p	Probability/prevalence of the outcome of interest
π	True population prevalence of the outcome of interest
δ	True effect size/minimal clinically important difference
$V(.)$	A variance function

Vectors and Matrices

$\boldsymbol{\varphi}$	A specification of the vector of scale parameters
$\boldsymbol{\alpha}$	A specification of the vectors of correlation parameters
$\boldsymbol{z}$	A specification of the vector of scale parameters
$\boldsymbol{s}$	A specification of the vector of pairwise correlations
$\boldsymbol{Y}$	Vector of outcomes
$\boldsymbol{X}$	Matrix of covariates
$\boldsymbol{P}$	Matrix of true correlation parameter
$\boldsymbol{R}_h$	Sum of product of basis matrices and unknown constants
$\boldsymbol{M}_h$	Basis matrix
$\boldsymbol{K}_m$	Vector of unknown constants
$\boldsymbol{\theta}$	Vector of natural parameters
$\boldsymbol{\phi}$	Vector of nuisance parameters
$\boldsymbol{V}$	Working covariance matrix

List of figures

List of tables

Chapter 1

INTRODUCTION

1.1 Overview

In clinical trials, participants are randomly allocated to treatment arms to achieve balance in known and unknown prognostic factors, eliminate selection bias, and improve the internal validity of the study (Hayes and Moulton, 2009; Campbell and Walters, 2014a). If successful, randomisation should minimise the effect of the prognostic factors so that researchers can controllably study the effect of the intervention(s) of interest (Samsa and Neely, 2018). In individually randomised controlled trials (RCTs), individual participants are directly randomised, while in cluster randomised controlled trials (cRCTs) groups/clusters of individuals are randomised, such as schools, care homes, hospital wards, or general practices (Campbell and Walters, 2014a).

There are two levels of participants in cRCTs; the distinctive cluster level and the individual level which is nested within the cluster – with correlated outcomes (Hayes and Moulton, 2009; Walters, Morrell and Slade, 2011). An appropriate statistical method for analysing cRCTs will be any method that considers this hierarchical nature of the cRCT design (Bland, 2004; Christie, O'Halloran and Stevenson, 2009). Ignoring the correlated outcomes at the cluster level and using standard statistical methods that treat the outcomes as being independent, might lead to underestimating the standard errors (SEs) of the parameter estimates and consequently obtaining narrower confidence intervals (CIs), false small P-values, and incorrectly overstating the effect of the intervention (Hayes and Moulton, 2009; Campbell, 2014).

Furthermore, some of the common issues in the design and analysis of cRCT are (a) Ignoring clustering (Offorha, Walters and Jacques, 2022) (b) inadequate handling of missing data (Twardella, Bruckner and Blettner, 2005), (c) and poor reporting quality (Ivers et al., 2011; Offorha, Walters and Jacques, 2022). Newer analytical methods for handling clustering have been proposed in the literature of other study designs with clustered data, such as longitudinal study design. Notable ones are second-order generalized estimating equations (GEE2) (Prentice and Zhao, 1991; Yan and Fine, 2004), quadratic inference function (QIF) (Qu, Lindsay and Bing, 2000), and alternating logistic regression (ALR) (Carey, Zeger and Diggle, 1993).

Some of these newer methods are considered as alternatives to some already existing methods, for example, QIF is considered a promising alternative to GEE1 in the context of longitudinal studies (Qu, Lindsay and Bing, 2000; Odueyungbo *et al.*, 2008; Song *et al.*, 2009). It is worth noting, that these alternatives have not been comprehensively evaluated against already existing/established methods used in cRCTs to warrant their routine application (Westgate, 2012; Turner, 2017). For example, to the best of my knowledge, QIF have not been comprehensively compared against first-order generalized estimating equations (GEE1) and/or generalized linear mixed model (GzLMM) in the context of cluster randomised controlled trials. This could possibly one of the reasons for their slow uptake in cRCTs (Turner, 2017; Offorha, Walters and Jacques, 2022).

This thesis aims to contribute to the literature of the design and analysis of cRCTs with an in-depth investigation on the comparative performance of the statistical methods for analysing outcome data from cRCTs.

1.2 Chapter aim

This chapter aims to present a general introduction to this research. After conceptualising the potential research hypothesis, the research questions, aim, and objectives were deduced. This chapter also provides an overview of the structure of the thesis.

1.3 Research questions

The research questions were created to explore the primary aim of evaluating the statistical methods for analysing outcome data from cRCTs. The research questions are:

- What are the appropriate, and available methods, in the literature for analysing outcome data from cRCTs?
- What statistical methods are used in practice for analysing outcome data from cRCTs?
- What criteria should be used in deciding the appropriateness of the identified methods?
- How well do the selected methods perform, when compared?

1.4 Research aim

The primary aim of this research was to identify, describe, and compare the selected statistical methods for analysing outcome data from cRCTs and make recommendations about the most appropriate method of analysis in a specific cRCT scenario. This overall aim is further broken down into the following specific objectives.

1.5 Research objectives

The research objectives are:

- To identify what statistical methods are described in the literature for analysing cRCTs.
- To identify common statistical methods used in practice to analyse cRCTs, by reviewing cRCT reports published in the NIHR Journals Library.
- To identify the key criteria for declaring a statistical method to be appropriate for analysing cRCTs.
- To apply the criteria to the statistical methods found in the NIHR Journals Library and methodological reviews to select the methods to evaluate.
- To describe the selected statistical methods for analysing outcome data from cRCTs.
- To apply the selected statistical methods to various real-world cRCT datasets and evaluate their performance.
- To use a simulation study to compare the selected statistical methods for analysing cRCT based on the evaluation of their statistical properties.
- To make recommendations for the most appropriate method for analysing outcome data from cRCT design.

1.6 Structure of the Thesis

The remainder of this thesis is structured as follows: Chapter 2 presents the background information on major designs, explains key concepts in cRCTs, describes the classical analytical approaches for analysing outcome data from cRCTs, and states the consequences of not using appropriate analytical methods that account for clustering to analyse outcome data from cRCTs. Chapters 3 and 4 present the results of methodological and practice reviews, respectively. In

Chapter 3, the primary aim was to identify the available and appropriate analytical methods for analysing outcome data from cRCTs in the literature. The electronic bibliographical databases of MEDLINE, EMBASE, PsycINFO, CINALH, and SCOPUS were systematically searched with a developed search strategy. The practice review focussed on the methods used in cRCTs reported in the NIHR Journals Library – Efficacy and Mechanism Evaluation (EME), Health and Social Care Delivery Research (HSDR), Health Technology Assessment (HTA), Programme Grants from Applied Research (PGfAR), and Public Health Research (PHR). The findings of the practice review are presented in Chapter 4, with one publication in BMC Trials stemming from Chapter 4 (Offorha, Walters and Jacques, 2022).

Results from the two reviews were compared, and gaps in knowledge were identified and conceptualised into research questions, aim, and objectives presented in Chapter 5. Chapter 6 provides further technical descriptions of the two classical (Chapter 2) and two selected methods identified in the methodological review of Chapter 3. These four analytical methods were then applied to real-world outcome data from four cRCTs, and the results are presented in Chapter 7. Chapters 8 and 9 presented the protocol for conducting a simulation study and the results from the study, respectively. Chapter 10 summarised and discussed the findings of this research. the strengths and contributions of this research to the field of cRCTs with several recommendations made is also presented in Chapter 10. This research had some limitations, and future research areas to explore also in Chapter 10.

Chapter 2

Background of cRCTs

2.1 Introduction

In Chapter 1 the overview and purpose of this research were discussed, from which the research questions, aim, and objectives were deduced. The structure of the entire thesis was provided at the end of the chapter. This current chapter provides the background on cRCTs by explaining the reasons for choosing a cRCT design over RCT, the different types of cRCT designs, the computational formulae for the ICC, the power of a study, discusses the different approaches and classical methods for analysing outcome data from a cRCT and the consequences for ignoring the correlation among outcomes in a cluster. The chapter ends with a brief explanation of the CONSORT reporting guidelines extension for cluster randomised trials. The cRCT is a special type of clinical trial design that is increasingly being used in medical research. Other names for cRCTs are group or community randomised trials (Murray *et al.*, 2004). In this thesis, "cluster" is used to represent groups of individuals which is the unit of randomisation and not just individual participants. In cRCTs, choosing what criteria to use in grouping the individual participants is majorly driven by geographical, demographical, or medical considerations (Moberg and Kramer, 2015).

Chapter 3 goes beyond the classical methods and identified more methods for analysing outcome data from cRCTs. It presents the results of a methodological scoping review conducted to investigate the available and appropriate methods that can be used or have been used to analyse outcome data from cRCTs. A well-developed search strategy was employed to search the electronic databases of MEDLINE, EMBASE, PsycINFO, CINAHL, and SCOPUS. Whereas, Chapter 4, investigated the frequency of the practical application of the methods described in Chapter 2 and 3. The online table of contents of the five NIHR Journals Library chronologically. The results revealed the commonly used methods in practice.

2.2 Chapter aim

The aims of this chapter are:

- To provide background information on the major designs, relevant concepts, and analytical methods for cRCTs used throughout in this report.
- To provide examples illustrating the use of the described designs, concepts, and analytical methods in practice.
- To provide some plausible reasons why a researcher would choose a cRCT design.
- To provide brief descriptions of the classical statistical methods for analysing outcome data from cRCTs.

2.3 Why choose a cRCT design?

As explained in Chapter 1, there are two fundamental ways of randomly allocating participants to the treatment arms in randomised controlled trials (RCTs). First, by randomly allocating the individual subjects, and second by randomly allocating clusters of subjects. Randomising individual subjects gives rise to RCTs while that of clusters of subjects gives rise to cRCTs (see, **Figure 2.1**).

In medical research, RCTs are more common than cRCTs (Mollison et al., 2000). One of the major reasons is that generally, RCT is more efficient than cRCT because it produces a parameter estimate of the intervention effect with smaller SE (Hemming *et al.*, 2021). For example, if a RCT and a cRCT are powered on the same sample size, the latter is more likely to produce a less accurate estimate of the intervention effect with larger SE and wider CI, this is because for a cRCT there are two sources of variations in the outcome variable: the variation in the outcomes between the distinct clusters termed "between-cluster variation (σ_b^2)", and the variation, in the outcomes of the individual subjects in each cluster termed "within-cluster variation (σ_w^2)". Unlike in RCTs where only the within-cluster variation σ_w^2 is accounted for. Having two sources of variation is one of the major drawbacks of the cRCT design. For example, one of the consequences is that the power (as determined by the sample size) needed to produce more precise parameter estimates would be reduced (Julious, 2023).

This inherent disadvantage of cRCTs should be accounted for, both at the design and analysis stages. At the design stage, the sample size calculated on the premise of an RCT should be inflated by an inflation factor to adequately power the cRCT study (Julious, 2023). While at the analysis stage, an appropriate analytical approach that accounts for the effect of clustering should be used as a follow-up. Note that doing so will conform with the popular mantra *"analyse as you randomised"* (Campbell and Walters, 2014a). Other reasons for the popularity of RCT compared to cRCT is that it is easier to set up, conduct and analyse (Hayes and Moulton, 2009; Campbell and Walters, 2014a). However, conducting an RCT to make informed decisions in support of the principle of evidence-based medical practice might not be feasible or ethical, or statistically sound sometimes (Hayes and Moulton, 2009; Moberg and Kramer, 2015). Nonetheless, before choosing a cRCT over an RCT design, clear and convincing reason(s) should be provided (Hayes and Moulton, 2009). For example, a recent study acknowledged the difficulty of implementing the intended interventions at the individual level without risking protocol violation (Wyld *et al.*, 2021).

Box 2.1 provides some possible questions to consider when deciding if a cRCT design should be used instead of an RCT.

Box 2.1 Questions to help determine if a cRCT design is appropriate instead of an RCT (Pladevall *et al.*, 2014).

1. Does the technology/intervention been studied occur naturally at a "group" or "cluster" level (e.g., a medication prescribing guideline that is implemented throughout a group practice), with effects measurable at an individual patient level?
2. If individual participants (e.g., patients) were randomised, would it be difficult for those administering the intervention (e.g., physicians) to change behaviour according to the arm of the study to which an individual had been assigned?
3. Is it likely that individuals in the study (patients, clinicians, other staff) would have occasion to talk among themselves about the study and possibly "crossover" to the other arm (e.g., adopt a diet or exercise plan to which others had been assigned) based on those conversations?
4. Is it possible that individuals in the study could intentionally bias results of the study because of their knowledge of the characteristics of different study arms (e.g., patients assigned to usual care versus an exercise intervention might start exercising more)?
5. Would it be substantially easier or more efficient to apply the experimental intervention to clusters or units of individual participants rather than to individuals one at a time?
6. Will the participants subvert the randomisation (e.g., equipment in an ambulance or ED department where individual randomisation will be ignored as clinicians like the shiny new thing) or will clustering occur due to the care being given (e.g. surgery)?

Answering **"Yes"** to one or more of the questions in **Box 2.1** suggests that a cRCT design may be appropriate.

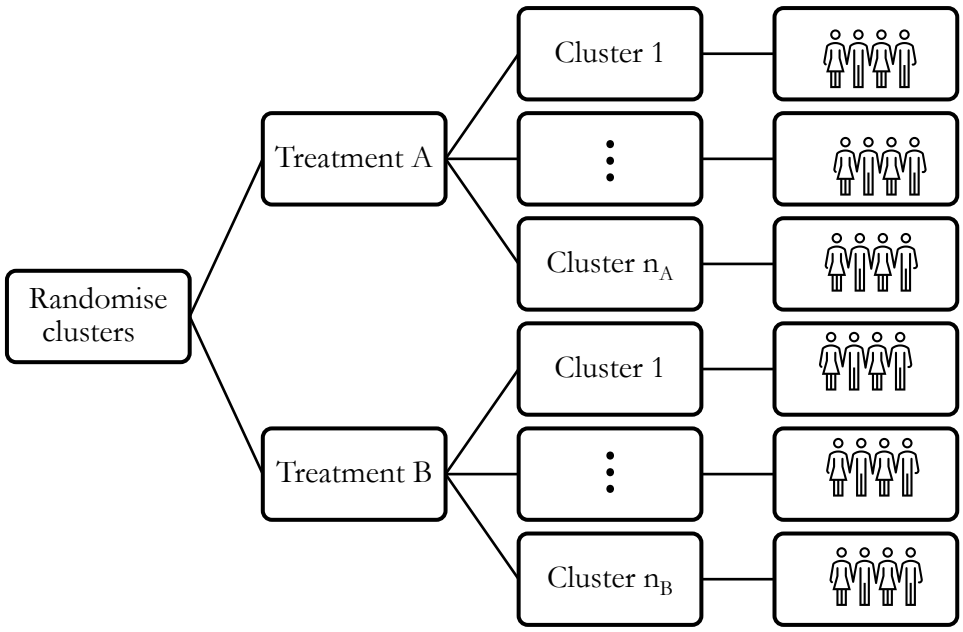


Figure 2.1 A simple cluster randomised controlled trial schematic

2.4 Types of cRCT design

These are the most common types of cRCT design (Offorha, Walters and Jacques, 2022), but this is not an exhaustive list, there are more in the cRCT literature for which the scope of this thesis does not cover.

2.4.1 Parallel group cRCT

This is the most straightforward and commonly used cRCT design (Offorha, Walters and Jacques, 2022). Here, identified clusters are randomly allocated to the distinct treatment arms separately and these clusters are expected to maintain their membership till the end of the trial (see, **Figure 2.2**). If the clusters are assigned at random to the different treatment arms without pairing, it is known as an *unmatched or simple* parallel group cRCT. An example of an unmatched parallel group cRCT is the Age-Gap trial (Wyld *et al.*, 2021). Bridging the Age Gap cluster RCT compared the decision support intervention (DESI) against usual care in the treatment decision-making of older women (aged 70+ years) with operable breast cancer.

In Bridging the Age Gap trial, there were two distinct groups - the intervention and the control. Of the Forty-six breast cancer units (the clusters), 21 clusters were randomly and separately allocated to the decision support (i.e., intervention arm) and the remaining 25 clusters to the usual care (i.e., control arm). At the end of the 6 months follow-up period, the primary outcome of interest; the global health status/quality of life was obtained from each subject. The outcomes from the individual subjects belonging to the two separate treatment arms were then compared analytically to determine if the intervention was effective.

In the simplest form, when two clusters are paired based on some identified important prognostic factors that do reasonably influence the outcome of interest before one member of the paired clusters is randomly assigned to the intervention arm and the other to the control arm, this process is known as the *matched paired* parallel group cRCT. A good example is the COMMIT study (Gail *et al.*, 1992). It was designed to investigate smoking cessation at the community level. The researchers identified that there would be some natural heterogeneity among the communities before allocation, on this basis the 22 communities involved were first matched in pairs based on population size, geographical proximity, age and sex composition, degree of urbanisation, and socioeconomic factors. Then one member of each pair of communities was randomly assigned to the intervention arm and the other to receive the control. While the unmatched design is the simplest of the two it has major setbacks. One of the two common setbacks is the unrealistic dependence on randomisation alone to achieve baseline covariates balance between the treatment arms (Campbell and Walters, 2014a). The second is the loss in efficiency, as determined by the sample size required to achieve the desired power (Gail *et al.*, 1992) .

similarly, the m*atched pair* cRCT design has its pros and cons (Chondros *et al.*, 2021). As an alternative to unmatched cRCT, it provides more balance between the treatment arms in some important baseline covariates. Also, this design enjoys some gain in efficiency in terms of the sample size needed to power the study, but not in all cases as would be explained further. Of recent, a study was conducted to comprehensively investigate specific cases where the unmatched pair or matched pair, or stratified randomisation gain efficiency when compared to one another. When the number of clusters per arm was 10 or above and the *matching correlation* was high, say $^1/_3$ or $^1/_2$, the match pair design was more efficient than the unmatched. The matching correlation is the within-stratum correlation of a matched pair concerning the outcome of interest (Campbell and Walters, 2014a; Chondros et al., 2021). For cRCTs with less than 10 clusters per arm and small

($^{1}/_{20}$ or $^{1}/_{10}$) matching correlation the opposite was the case i.e., the unmatched design was more efficient, that is it produced more precise estimates (Chondros *et al.*, 2021). These findings conformed to that of a previous study (The COMMIT Research Group, 1995).

The major reason behind these results was attributed to the loss of the degrees of freedom (DoF) of the intervention effect estimate SE available during analysis. For a matched pair cRCT of 11 pairs, the available DoF will be 10 clusters while for its unmatched counterpart, it would be 20 clusters. Diehr et al. (1995) showed that breaking the pairs during analysis would help regain the lost degree of freedom, which will provide more power and precision. If the researchers intend to use this strategy, at least, it should be stated in the protocol. Because this goes against the previously stated mantra "*analyse as you randomised*", which is counterintuitive.

Another example of a matched pair study where the matching was broken during analysis is the "Informed Choice trial". O'Cathain et al. (2002) aimed to investigate the effect of using leaflets to assist women using maternity services to make informed choices. The intervention could only be effectively delivered at the maternity unit level due to the risk of contamination, that is the possibility of the women being randomised to the different treatment arms (control and intervention) but assessing the same maternity unit, from sharing their experiences or study materials. Hence, the maternity units were randomly allocated instead of the individual subjects. Some maternity units are bigger than others both in size and caseloads. The 10 maternity units were paired based on their annual number of deliveries before one member of the pair was randomised to the control and the other to the intervention arm. Nonetheless, unmatched analyses were performed.

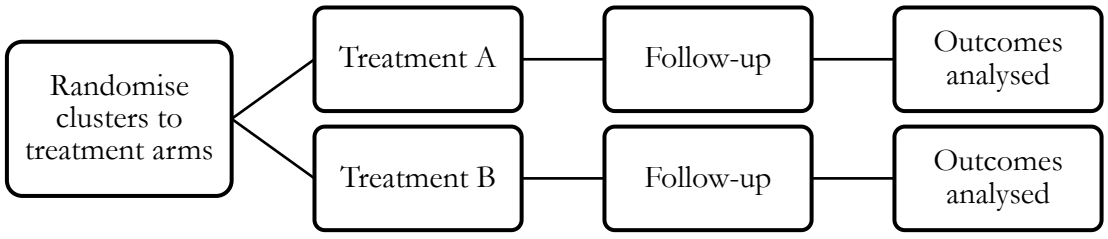


Figure 2.2 A schematic of a parallel treatment group cRCT design.

2.4.2 Crossover cRCT

This cRCT design is a good alternative to the parallel group design. It addresses the cons of a parallel group cRCT design. And it is efficient when the number of clusters in a study is small and the variability in the baseline prognostic covariates is presumed to be large (Hayes and Moulton, 2009; Chatterjee and Bandyopadhyay, 2017). Each of the clusters randomly receives all the treatments in different sequences indexed by time, with a washout period between the time points (if required). A washout period is a period between the end of treatment to the start of another by a treatment sequence (Campbell and Walters, 2014a). This is done to reduce the residual/spillover effect of the previous treatment, especially when the treatment is not the control (Hayes and Moulton, 2009).

To illustrate this, the simplest standard case is presented in **Figure 2.3**. This involves 2 treatment arms resulting in two-sequence, two-period, normally denoted as an AB/BA crossover cRCT design. The AB arm receives treatment A followed by B after possibly allowing for a defined washout period as shown in **Figure 2.3.** This design is not the most efficient because it could be difficult sometimes to eliminate the carryover effect (especially when one treatment is a control) in the second period, which could mask the potential effect of a treatment by period interaction (Campbell and Walters, 2014a). The AAB/BBA design is an improved version that would allow for the estimation of the carryover effects (Campbell and Walters, 2014a). This design is among the eight possible treatment sequences of a two-treatment three-period crossover design (i.e., 3×2) (Ebbutt, 1984; Chatterjee and Bandyopadhyay, 2017, 2019). The AAB/BBA design involves repeating the same treatment given in the first period, in the next (second), then a different treatment is given at the third (i.e., last period) (Senn, 2002; Machin and Campbell, 2005). This allows for the carryover effects to be estimated in the second period (when the other treatment is absence) and in the third period (when the other treatment is present). However, the use of this design has criticised by Ebbutt (1984), and Chatterjee and Bandyopadhyay (2017) (2019). First, they argue that the first two-period does not represent an ideal crossover design, and this goes against the major reason of choosing a crossover design (Ebbutt, 1984). For example, in a case where there are enormous dropouts in the third period, the data obtained from the first two periods cannot be analysed as a crossover design (Chatterjee and Bandyopadhyay, 2017). Second, that they the carryover effects from A to B cannot be evaluated in each subjects (Ebbutt, 1984). This second argument is true only for the first two periods (i.e., AA(B)\BB(A)).

In general, the crossover design gains efficiency over the parallel design because each cluster serves as its control reducing the internal variability across the clusters. This reduces the magnitude of total random errors which in turn improves the power of the study and the precision of the estimates (Hayes and Moulton, 2009; Crespi, 2016). One demerit of a crossover design is that it takes a longer time to complete which could cost more compared to a parallel group design. This is because each of the clusters will be exposed to each of the treatments one after the other (Hayes and Moulton, 2009).

A good example of a crossover cRCT is the Bridge-It study (Cameron *et al.*, 2020). The study investigated the effectiveness of using a complex intervention of a temporary oral contraception pill plus the offer to have expert meetings to discuss all available contraception methods compared to the control treatment of using emergency contraception alone. The primary aim was to promote the use of effective contraception methods among women to minimise the risks of unplanned pregnancies. Their justification for choosing a cluster trial was to meet recruitment targets. While the choice of a crossover cluster design was to achieve efficiency since each pharmacy (the cluster) will serve as its control and to maximise the power of the study for using a small number of clusters (29 pharmacies). One of the two groups received the control treatment followed by the intervention after a minimum of 2 weeks washout period while the other had the intervention followed by the control. The study lasted for about a year and six months.

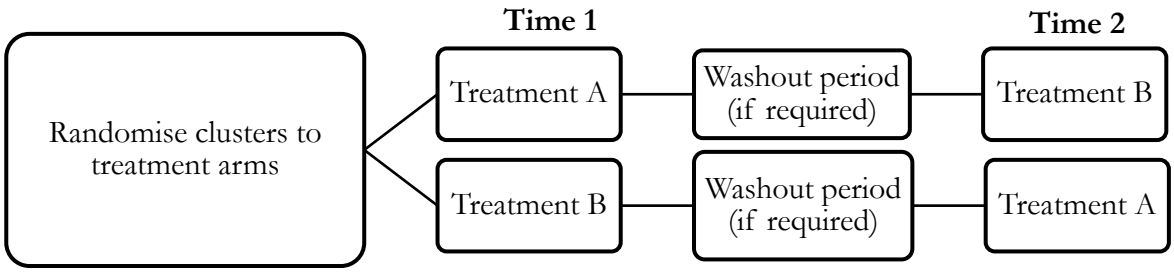


Figure 2.3 A schematic of a 2 × 2 crossover cRCT design.

2.4.3 Stepped wedge cRCT

This is a good alternative to both the parallel and crossover cRCT designs when it is not feasible to roll out the intervention treatment to all the intervention clusters simultaneously. This could be due to logistic challenges, high costs, or unavailability of the intended intervention (Group, 1987; Hemming *et al.*, 2015). Here, at the baseline of the trial, all the clusters start with the control treatment, then at the first "step", 1 cluster (or more) switches to begin the intervention treatment while the others remain on the control treatment. This process is repeated in each new step, in a regular time interval, until all clusters are exposed to the intervention in a random manner. This is depicted in **Figure 2.4** for better understanding.

The stepped wedge design shares some similarities with the crossover design explained in Section 2.4.2. Firstly, each cluster serves as its control. Secondly, it is efficient when a small number of clusters are available (Hussey and Hughes, 2007). This design presents some ethical advantages as well, it provides the opportunity for all participants to be exposed to the promising intervention treatment which could be beneficial, and at the same time better chances to stop earlier with minimal damages, if the intervention is too harmful. Hence, stepped wedge trials are most times implementation studies – implementing an intervention that is known to work (Group, 1987). An obvious disadvantage is that it would take a longer time to complete which comes with possible increased cost.

The Gambia Hepatitis Intervention Study is a good example, and it is considered to be one of the earliest clinical trials that used the term "stepped wedge" (Group, 1987). The study was initiated by the Gambia government in conjunction with WHO to investigate the long-term effects of a nationwide hepatitis B vaccination (HBV) programme for all children. Their interest was to assess the efficacy of the vaccine in preventing chronic liver disease and liver cancer in the future. The intervention could only be administered at the cluster level and in stages. So, 17 immunisation teams were created, and all were administered the standard vaccine at the start of the trial. At each new stage (usually about every 3 months) a new team is randomly activated to start immunising all the new-borns in a well-defined area with the HBV + standard vaccines. It took over 4 years for all the 17 teams to be activated and the study ended after that.

The major reasons for choosing the stepped-wedged design in the Gambia study were:

- The cost and scarcity of the HBV vaccines. This meant that it was difficult to roll the vaccines to all participants in the intervention arm simultaneously.

- A need for comparators at the same time.
- Logistics challenges if the intervention has been delivered at the individual participants' level.
- The hope to solve the scarcity problem of the vaccine and make it widely available as the trial progresses.
- The ethical justification of all the clusters getting the promising intervention at the end of the trial.

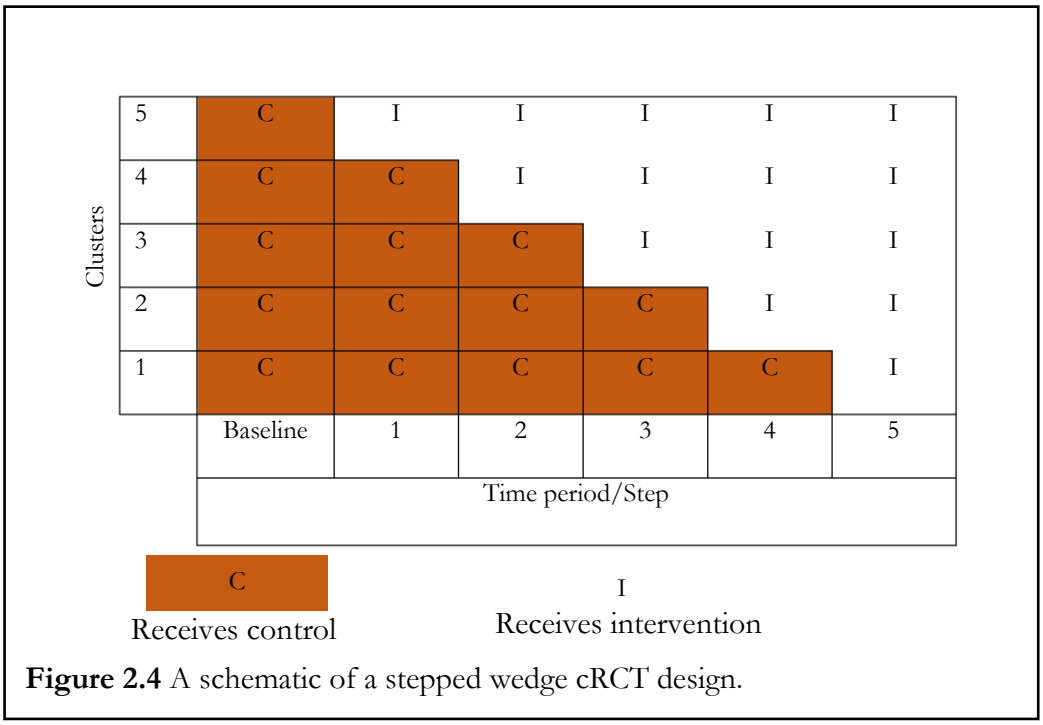


Figure 2.4 A schematic of a stepped wedge cRCT design.

2.4.4 2^k factorial cRCTs

This design gives the option of assessing the effect of two or more interventions without necessarily increasing the sample size. Aside from assessing the independent effects of the interventions, their combined effects can also be assessed simultaneously in the same trial. The generic representation is 2^k, where k is the number of intervention treatments to be assessed with each having 2 levels (Mdege *et al.*, 2014).

To illustrate the simplest and most common factorial design – the 2^2 factorial design, 2 interventions with each having 2 levels (Offorha, Walters and Jacques, 2022). Assuming that the two intervention treatments of interest to a researcher are treatments A and B, and one control treatment C. At the first level of randomisation, clusters are allocated to treatment A and treatment C, at random. At the second level, under each of the treatments of the first level, treatments B and

C are introduced. This will give four possible combinations of the treatment arms AB, A, B, and C. Clusters of participants are then randomly allocated to the four treatment arms. The arm AB is the interaction arm for assessing the combined effects of treatments A and B (**Figure 2.5**).

Although, the parallel group cRCT with independent treatment arms can also serve this purpose by having more arms (e.g., three treatment arms): two intervention arms with one control, thereby lowering the required sample size, since there would be no treatment arm created to investigate the interaction of the two treatments (i.e. AB group) (Mdege *et al.*, 2014). However, this three-armed parallel groups design is adequate compared to a 2^2 factorial design, only when there is no substantial correlation between the two treatment arms. If there a substantial correlation, then a 2^2 factorial design is a better option, though a larger sample size would be required to power it compared to a three-armed parallel groups cRCT design. Increasing the sample size to power the factorial study means eliminating one of the major merits of a factorial design (Mdege *et al.*, 2014).

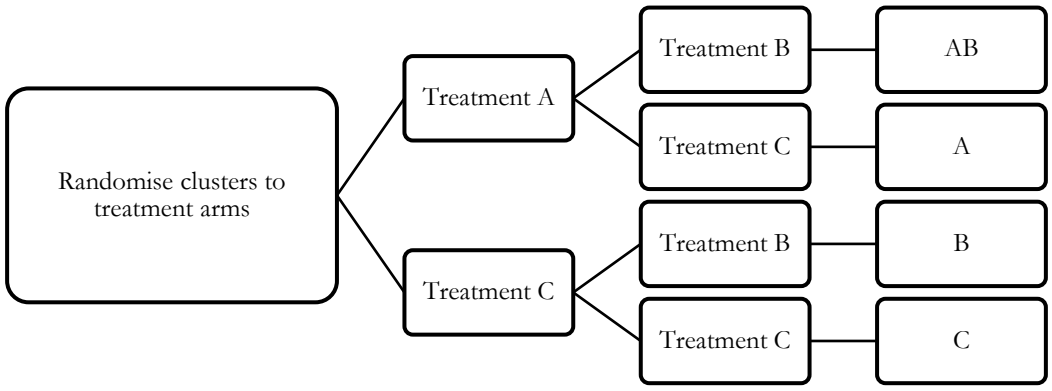


Figure 2.5 A schematic of a 2^2 factorial cRCT design

For example, Hartinger et al. (2020) employed a 2^k factorial design to evaluate the impact of two interventions – kitchen sink & hygienic education (IHIP), early child development programme (ECD), and the control treatment; on improving the health of children under 36 months living in a rural area. The four treatment arms created in the study are the control, a combination of IHIP and ECD represented as IHIP+; IHIP alone; and ECD alone. Ten clusters were randomised to each of the four treatment arms. Their reason for choosing this design was that it allows a comprehensive comparative evaluation of the two interventions using a single trial without necessarily increasing the sample size. Only baseline findings were reported in the paper, because

the main purpose of the study was to evaluate the feasibility of the trial design, for instance, to ascertain if the randomisation scheme would be successful in achieving considerable balance in baseline covariates.

2.5 Simple and stratified randomisation

Randomisation is the cornerstone of experimental clinical trials. If it is done well it could minimise, if not eliminate, the influence of confounding factors and selection bias. A good randomisation scheme should be reproducible and unpredictable. Hence, tossing a coin and picking numbers from an urn are precluded (Campbell and Walters, 2014a). Randomisation allows the investigator to controllably assess the effectiveness of the intervention(s) of interest. It is an important concept that qualifies experimental clinical trials as the gold standard of medical research. Compared to RCTs where individuals are randomised, in cRCTs, clusters (c in cRCTs) of individuals are randomised to the treatment arms. The following are the reasons why randomisation is important:

- Balance: For a reasonable sample size, randomisation would ensure a balance in quantity and quality of both known and unknown prognostic factors among participants in the treatment arms (Campbell and Walters, 2014a). This will isolate the true effect of the intervention from all other confounding factors.

- Objectivity: Randomisation minimises the conscious or unconscious introduction of the researcher's opinion or feelings into the recruitment and allocation processes, which may result in selection bias. In most recent RCTs, the person(s) in charge of the randomisation is excluded from all other aspects of the trial. This act does provide good assurance to all stakeholders on the findings of the trial.

- Replicability: Replicability is one of the hallmarks of medical research. Random allocation of clusters (with individuals nested) ensures that the process of a trial is impartial, and transparent, and could undoubtedly be repeated by other researchers.

- Blinding: The scheme of randomly allocating clusters will help support the purpose behind blinding. The participants will find it somewhat difficult to guess which treatment arm they have been assigned to. This is in the good interest of minimising things that could go wrong and undermine the findings of the trial.

The most common and straightforward randomisation scheme is the random allocation of clusters on a 1:1 ratio to the treatment arms; termed "*simple*" randomisation. Suppose there are two treatment arms A and B, and the first cluster is defined and randomly assigned to either A or B. The next second cluster will be assigned to B assuming the first was assigned to A. This randomisation scheme may fail to achieve balance in the treatment arms if the trial fails to recruit prespecified targets. An alternative is the **block** randomisation, here allocation to the treatment arms is done in batches. Suppose a trial wants to recruit 24 clusters to treatment A and B, this would be done in batches of 6. Once the first batch is realised, 3 clusters are allocated to A and the other three to B. Matched pair is the simplest form of block randomisation with a block size of two (Campbell and Walters, 2014a).

On the other hand, **stratified** randomisation is used to achieve balance on relevant and known prognostic factors, especially if the factor is known to influence the outcome of interest. This involves creating strata (levels) of the factor and randomly allocating clusters within each stratum to the treatment arms. The essence is for each treatment arm to have a mix of the strata to ensure that the target population is well represented. A good example where this scheme was used was a cluster trial to assess the effectiveness of decision support interventions (DESIs) in improving quality of life and survival among older women with operable breast cancer. The clusters (breast units) were stratified by the levels of current primary endocrine therapy (low or high) and chemotherapy rates and then randomly allocated to the DESIs intervention arm or control arm (Wyld *et al.*, 2021). The recommendation is that all covariates used for stratification should be adjusted during analysis, regardless if they are statistically significant or not (Campbell and Walters, 2014a).

2.6 Intracluster correlation coefficient (ICC), ρ

Participants in a cluster share the same study resources such as trial site, exposure to the same intervention, investigator, outcome assessor, prognostic factors, etc. Hence, patient outcomes from each cluster are likely to be similar. This is one of the crucial aspects of cRCTs, and the degree of this similarity is quantified by the ICC ρ. The ICC is an important requirement in the planning stage of a cRCT for sample size calculation and ideally should be accounted for in the analysis stage (Eldridge, Ukoumunne and Carlin, 2009). It is defined as the proportion of the total variation σ^2 in the outcome that is attributable to the variation between the clusters σ_b^2 (Campbell and Walters, 2014a). The total variation σ^2 in the participant outcomes, is made up of the sum of

two components, the between-cluster σ_b^2 and the within-cluster σ_w^2 variances, thus the ICC for a continuous outcome is given as:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \tag{2.1}$$

Generally, equation

(2.1) is the basic representation of the ICC with the assumption that the correlation between outcomes of any two participants j and j' in a cluster, is the same. This assumption that the ICC is homogenous for all participants across clusters is common in medical research and it is most times called exchangeable or compound symmetry ICC (Zeger and Liang, 1986). However, in practice, it is not uncommon for the ICC to be heterogeneous across clusters, which could be because of differences in cluster characteristics such as the number of clusters and cluster sizes (Eldridge, Ukoumunne and Carlin, 2009). This heterogeneity should be accounted for during analysis, especially in cases where σ_b^2 is high and σ_w^2 is low, this implies that ρ would be large (Crespi, 2016). An example of such cases is a family clinical trial investigating the relationship between gene and alcohol dependence of family members (Yan and Fine, 2004). To account for heterogenous ICC across the clusters

(*2.1*) could be rewritten as:

$$\rho_i = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \tag{2.2}$$

where i is the cluster index indicating a particular cluster, hence ρ takes different values for each i. For binary or dichotomous continuous outcomes, it's common knowledge that the variance of the individual level residuals is not homogenous because it's a function of the mean, hence σ_w^2 does not exist (Eldridge, Ukoumunne and Carlin, 2009). The strategy is to come up with another way of expressing the total variance σ^2 of the patients' outcomes. Suppose the prevalence of the outcome of interest y_{ij}, for a jth participant in the ith cluster is p_i and follows a distribution with mean π and variance $var(p_i) = \sigma_b^2$ (Eldridge, Ukoumunne and Carlin, 2009; Campbell and Walters,

2014a). Thus $E(y_{ij}) = E(p_i) = \pi$ and $var(y_{ij}) = \pi(1-\pi)$, hence, the ICC ρ could be calculated as

$$\hat{\rho} = \frac{\sigma_b^2}{\pi(1-\pi)}$$

(2.3)

where π is the true population prevalence of the outcome of interest to be estimated from the sampled data. Another version of (2.3) is obtained by expressing p_i on a log-odds scale $\log(p_i)$. So, in (2.3) the numerator which is the variance of p_i, is on the log-odds scale while the denominator is on the proportional scale, the two scales are not relatable. In this case, equation (2.3) can be expressed as (Eldridge, Ukoumunne and Carlin, 2009).

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \left(\pi^2/3\right)}$$

(2.4)

These two parameters, σ_b^2 and σ_w^2 , can be replaced with the extracts from the output of a one-way analysis of variance (ANOVA). According to Donner (1986), the following equations hold true

$$\sigma_b^2 = (MSB - MSW)/\bar{n}$$

$$\sigma_w^2 = MSW$$

(2.5)

where MSB is the between-cluster mean squared error, MSW is the within-cluster mean square error, both MSB, and MSW are the extracts from ANOVA, $\bar{n}$ is the average cluster size calculated with this formula:

$$\bar{n} = \frac{1}{N-1}\left(n - \frac{\sum_{i=1}^{N} n_i^2}{n}\right)$$

(2.6)

where N is the total number of clusters, n is the total sample size, and n_i is the i^{th} cluster size. If equation (2.5) is substituted into

(*2.1*) the ICC estimator will become

$$\hat{\rho} = \frac{MSB - MSW}{MSB + (\bar{n} - 1)MSW} \tag{2.7}$$

Obtaining either a positive or negative ICC depends on which estimator is used, while the ICC estimator of equation (2.1) is positive definite because its components are variances, the one of (2.7) can produce a negative ICC estimate because of the subtraction on its numerator, and this occurs when $MSB < MSW$. For the rest of the chapters, equations (*2.1*) and (2.4) are referred to when discussing the estimator of the ICC for continuous and binary outcomes respectively, except otherwise stated. There are several other ways of computing the value of ρ as reviewed by Ridout, Demétrio and Firth (1999) the simplest way is to perform an ANOVA on a balanced data set to obtain the two necessary components, σ_w^2 and σ_b^2. Another way is directly from the output of multilevel regression models such as the LMM, GzLMM, GEE1, GEE2, and QIF.

2.7 Power of a cRCT study

The power of a clinical trial is relatable to the sample size used (Julious, 2023). To power a study adequately to ensure that it picks up the least possible signal of the effectiveness of the intervention treatment, an optimal sample size must be used in the study. Usually the power of a study, $1 - pr(\beta)$, improves as the sample size $n \to \infty$, where $pr(\beta)$ is the probability of Type II error (see, **Table 2.1**). However, because clinical trials involve humans as the experimental units, it will be unethical to overpower a study – exposing more participants than is necessary to an intervention that has not shown any benefit and thus could be harmful. Additionally, it will be wasteful to underpower a study – using inadequate small numbers of participants. This could pose a challenge in picking up the signal of the effectiveness of the intervention, when it truly exists (Button *et al.*, 2013).

Given the importance of powering a study sufficiently, sample size determination for cRCTs is well-studied in the literature, and more recent studies have been published on this aspect of cRCTs

(Eldridge, Ukoumunne and Carlin, 2009; Gao *et al.*, 2015; Crespi, 2016; Ribeiro, Milosavljevic and Abbott, 2018; Li and Jung, 2020). For any cRCT, it is recommended that cogent reasons be provided regarding the number of participants included in the study (Campbell *et al.*, 2012).

Table 2.1 The power of a study explained with respect to the null hypothesis

	Correct status of the null hypothesis	
Decision	**The null hypothesis is true**	**The null hypothesis is not true**
Do not accept	Type I error with probability α	Correct decision with probability (Power: $1-\beta$)
Accept	Correct decision with probability $1-\alpha$	Type II error with probability β

As explained earlier in Section (2.3), because there are two levels of participants in a cRCT – the cluster of individuals and the individuals themselves, the total response variance σ^2 is made up of the cluster-level variance σ_b^2 and the individual-level variance σ_w^2. The presence of σ_b^2 is an additional noise, and it does have a negative impact on the sufficient sample size needed to power a cRCT. The impact of σ_b^2 is determined by the ICC ρ.

For a superiority trial with a continuous primary outcome, the common strategy for calculating the required sample size is to calculate it based on an RCT, and then inflate it using an inflation factor to account for the effect of clustering in cRCTs. To apply the above strategy to a parallel group cRCT design, a sample size calculation for the RCT is given as (Julious, 2023):

$$\hat{n}_{per\ arm} = \frac{2\sigma_{Plan}^2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\delta_{Plan}^2} \tag{2.8}$$

where $\hat{n}_{per\ arm}$ is the required sample size per arm assuming an equal number per arm, $z(.)$ are the critical values from a Normal distribution corresponding to areas or probabilities under the standard Normal distribution curve of 1-$\alpha/2$ and 1-β respectively, σ_{Plan}^2 is the population standard deviation of the outcome which is assumed to be the same in all the treatment arms and

estimated from the sampled data, δ_{Plan} is the target or planned effect size (minimal expected difference) between the treatment arms. To ensure that adequate sample size is used, (2.8) should be inflated using a factor known as design effect (DE) and given below:

$$DE = 1 + (\bar{n} - 1)\rho$$

(2.9)

where $\bar{n}$ is the average cluster size, and ρ is the ICC. It is not uncommon in real-world trials to have varying cluster sizes instead of fixed participants per cluster as assumed in (2.9). If this is the case, to account for the varying cluster size, a variant of (2.9) is given as (Eldridge and Kerry, 2012).

$$DE = 1 + [(CV^2 + 1)(\bar{n} - 1)]\rho$$

(2.10)

where CV is the coefficient of variation for the cluster sizes (i.e., the ratio of the standard deviation of the cluster sizes to the mean cluster sizes). For other strategies for calculating sample size depending on the type of cRCT design used, if covariate(s) is included, and if clustering occurs only in one arm see Julious (2023). For example, a study by Hassiotis et al. (2018) aimed at examining whether challenging behaviours among adults with intellectual disability could be reduced if staff members were trained in Point Behaviour Support skills compared to those that continued treatment as usual (the control). The cluster unit was the community intellectual disability service, and they were randomised on a 1:1 basis between the intervention and control arms. The primary outcome, Aberrant Behaviour Checklist – Community (ABC-C), is a continuous variable.

From an internal pilot study, they obtained a mean baseline ABC-C score of 45.4 (SD = 26.4), the minimal clinically important difference was an SD reduction of 0.45 on the ABC-C score in favour of the intervention arm. Applying the ANCOVA method to the ABC-C measurements from an internal pilot study to obtain a correlation of 0.48 between the baseline and post-intervention ABC-C scores. They planned a 90% power for the study and a two-sided 5% significance level. Their strategy was to calculate the required sample size for an RCT using equation (2.8) where $\sigma^2_{Plan} = SD = 26.4; \alpha = 0.05; \beta = 0.1; \delta = 0.45$. Hence, 80 participants are required per arm. They went ahead to inflate this estimate to account for clustering using equation (2.10), $\rho = ICC = 0.062$

obtained from an internal pilot study; $\bar{y} = 12(SD, 3)$, CV = 3/12 = 0.25, which gave a total of 276 participants required.

2.8 Approaches for analysing cRCTs

In this section, classical methods that are appropriate for analysing outcome data from cRCTs are described. As explained in Section (2.3), the allocation of experimental units and administration of interventions are done at the cluster level. So, intact clusters of participants such as schools, general practices, care homes, communities, etc., are randomly allocated to the treatment arms. The two main analytical approaches are:

2.8.1 General notation

A boldface letter mainly denotes a vector or matrix or as otherwise specified. The general notation is established as; let y_{ij} denote an outcome variable for the j^{th} subject in the i^{th} cluster ($i = 1, \dots, N;\ j = 1, \dots, n_i$); N is the number of independent clusters in the study and n_i denotes the different numbers of subjects in each cluster (i.e., the i^{th} cluster size), y_{ij} has a corresponding set of p-dimensional vector covariates $\boldsymbol{X}_{pij}^{T} = \left(x_{1i}, \cdots, x_{pij}\right)$ where x_{1i} denotes an indicator variable for the treatment arm to which a cluster belongs (x_{1i} = 0 indicates the control arm and x_{1i} = 1 the intervention arm) for a trial with two treatment arms, and $\boldsymbol{Y}_i = \left(y_{i1}, \cdots, y_{in_i}\right)^T$ *is* a $n_i \times 1$ vector of the collection of the individual level outcomes for the i^{th} cluster. Also, $\boldsymbol{\beta}_p = \left(\beta_0, \beta_1, \cdots, \beta_p\right)$ is an unknown p-dimensional vector of regression parameters and $\boldsymbol{\mu}_i = \left(\mu_{i1}, \cdots, \mu_{in_i}\right)^T$ is an $n_i \times 1$ vector of true means with $\mu_{ij} = E\left(y_{ij} \left| \boldsymbol{X}_{pij}^{T}\right.\right)$ being the conditional expectation for the j^{th} subject in the i^{th} cluster with covariates $\boldsymbol{X}_{pij}^{T}$.

2.8.2 Analysis at the cluster level

This analytical approach involves collapsing the individual participants' outcomes into an aggregate summary measure for each cluster. The implication is that standard statistical methods can be applied, and they would be optimal. The common standard parametric methods that are often applied are *two-sample t-test, and regression models* (Campbell and Walters, 2014b; Offorha, Walters and Jacques, 2022). Sometimes the underlying assumptions of these methods might not be met, to

avoid producing unreliable results non-parametric alternatives should be employed, such as *Wilcoxon's rank sum test*, *Mann-Whitney U test,* or a *permutation test* (Hayes and Moulton, 2009; Campbell and Walters, 2014a). How to implement this approach is demonstrated below using a standard linear regression model for a continuous outcome, more examples of other methods mentioned above are available in the literature (Hayes and Moulton, 2009; Walters, Morrell and Slade, 2011; Campbell and Walters, 2014b). Suppose $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ is a summary measure for the i^{th} cluster, the cluster-level mean regression model is specified as

$$\bar{y}_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \tag{2.11}$$

where β_0 is the mean outcome for the control arm conditioned on the fixed covariates; β_1 is the difference in the mean of the outcomes of the intervention and control arms known as the intervention effect conditioned on the fixed covariates, x_{1i} is the treatment arm indicator variable (x_{1i}: control = 0 and intervention = 1), x_{pi} is the pth cluster-level covariate with coefficients β_p and ε_i is the independent Normally distributed cluster-level residuals. The regression parameters of (2.11) can be estimated using ordinary least squares (OLS).

This approach mainly allows cluster-level covariates to be adjusted for, complication arises when there is a need to adjust for individual subject baseline covariates, and this is a major setback of this analytical approach. For examples of implementing this analytical approach see Walters, Morrell and Slade (2011) and (Campbell and Walters, 2014a).

2.8.3 Analysis at the individual-participant level

This analytical approach maintains the nested/hierarchical nature inherent in cRCT data. The response values are measurements from each participant within a cluster. So, we have a group of individual outcomes nested within a cluster. The methods discussed under this approach are regression methods, and they account for individual-level participant prognostic factors by including them as covariates. Cluster-level covariates can also be included. These regression methods are further classified into four modelling strategies based on how clustering is adjusted (Walters, Morrell and Slade, 2011).

2.8.3.1 Regression models with robust standard errors (RMRSE)

This modelling approach applies a post hoc correction to the SEs of the parameter estimates from an OLS model. The SEs of the OLS model equation (2.11) are smaller because σ_b^2 is not accounted for in the model. This might result in falsely increased precision (narrower CI) and inflated Type – I error rate. Let's illustrate how to correct this false positive result by using an individual-level regression model with the continuous outcome, given as

$$y_{ij} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pij} + \varepsilon_{ij}, \quad i = 1, \ldots, N;\ j = 1, \ldots, n_i$$
$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

(2.12)

where β_1 is the intervention effect conditioned on the fixed covariates, $x_{1i}\ and\ x_{pij}$ are the indicator and p^{th} variables respectively for the j^{th} individual in the i^{th} cluster, ε_{ij} is the residual for everyone.

The parameter estimates from an RMRSE model are the same as that of an OLS model. The difference lies in their estimate of the SEs of the parameter estimates, consequently, their confidence intervals and P-values are also affected. The RMRSE applies a post hoc correction to the SE estimates from an OLS model. Hence, the statistical inferences based on the estimates from RMRSE are more reliable than those of the OLS model. The basic strategy of RMRSE is to adjust the SEs produced by a model (called model SEs) to be more *robust* to departures from the basic assumptions of a regression model such as constant variance assumption, consequently resulting in valid statistical inferences (Campbell and Walters, 2014b). The robust variance estimator for an OLS model is calculated using the formula (Walters, Morrell and Slade, 2011)

$$\hat{V}(\beta)_{robust} = \hat{V}_{OLS}\left(\sum_{j=1}^{n} \boldsymbol{u}_j^T \boldsymbol{u}_j\right)\hat{V}_{OLS}$$

(2.13)

where $\hat{V}_{OLS}$ is the standard variance estimator given as $\sum_j^n (e_j^2)\sum_j^n (y_j - \bar{y})/n - 2$, e_j is the residual of the j^{th} participant, $\boldsymbol{u}_j$ is a vector of the j^{th} contribution to the $\partial logL/\partial \boldsymbol{\beta}$ and $\boldsymbol{u}_j^T$ is its transpose. Equation (2.13) is valid when we assume that $corr\left(e_j, e_{j'}\right) = 0$, which hold true

for analytical models for analysing outcome data from RCTs. But we already know that for a cRCT outcome data $corr\left(e_j, e_{j'}\right) \neq 0$, so if $C^1, C^2, \ldots, C^N$ represent the independent clusters, in light of this information we can rewrite (2.13) as

$$\hat{V}(\beta)_{cluster-robust} = \hat{V}_{OLS}\left(\sum_{i=1}^{N} \boldsymbol{u}_i^{C^T}\boldsymbol{u}_i^{C}\right)\hat{V}_{OLS} \tag{2.14}$$

where is $\boldsymbol{u}_i^{C}$ and $\boldsymbol{u}_i^{C^T}$ are the cluster version of $\boldsymbol{u}_j$ and $\boldsymbol{u}_j^{T}$.

To illustrate the application of RMRSE the GoActive study is a good example (Corder *et al.*, 2021), the study had 16 schools (clusters) randomised to two treatment arms with a total of 2,638 adolescents (individual participants). It was conducted to investigate the effectiveness of intervention (Get Others Active) in promoting increased physical activity among adolescents (Year 9 students). The intervention effect was the adjusted mean difference in change from baseline between the intervention and control groups. The primary outcome was the average daily minutes of moderate-to-vigorous physical activity (MVPA) measured at 10 months post-intervention and was analysed using a standard linear regression model, of which the standard errors produced were corrected to allow for clustering of the students within schools.

2.8.3.2 Cluster-specific regression models (CSM)

This analytical approach utilises an ad hoc process to adjust for clustering. The adjustment is made whilst simultaneously estimating the coefficients of the regression models, hence, it has a direct impact on the parameter estimates of the model. Therefore, the estimated coefficients from a CSM could be different from those of an OLS and RMRSE models. The estimate of the intervention effect from this analytical approach is interpreted as what will happen to individuals in a cluster if they receive the intervention treatment compared to them receiving the control treatment. The linear mixed model (LMM) is a common example of this approach for analysing continuous outcome data.

2.8.3.2.1 Generalized linear mixed model (GzLMM) with coefficients estimated by MLE

The GzLMM is also called a random (or mixed) effects model and is the most used cluster-specific model for analysing outcome data from cRCTs (Twardella, Bruckner and Blettner, 2005; Offorha,

Walters and Jacques, 2022). The LMM is a special case of a GzLMM. A GzLMM uses a single equation to assess the impact of the fixed effects of some covariates of interest and the random effects of the randomly selected clusters on the outcome(s) of interest in the study. The MLE is often used to simultaneously estimate the fixed and random effects parameters of a GzLMM. However, technically, the MLE algorithm first estimates the fixed effects component (ignoring the random effects component), then plugs the estimates into the algorithm to estimate the random effects component. This process is repeated until optimal estimates are obtained. However, ignoring the random effects component in the first step causes the MLE to produce negatively biased variance components, because it means ignoring the variations present in the estimates of the fixed effects, which could be substantial when the sample size is small (McNeish and Stapleton, 2016; Leyrat et al., 2018; Thompson et al., 2022). Also, the MLE does not adjust for the DoF lost in estimating the parameters of the fixed effects component (McNeish and Stapleton, 2016). Hence, the MLE is likely to produce estimates of SEs that are too small, resulting in smaller P-values, and inflated Type I error rates, especially when there are few clusters.

An alternative likelihood-based estimation method is the restricted maximum likelihood estimation (REML) which can be utilised to circumvent these problems. For large sample sizes, these problems are not noticeable, and the estimates from MLE and REML are approximately the same. But for cRCTs with small samples, the problems are more pronounced (Leyrat *et al.*, 2018; Thompson *et al.*, 2022). The REML first transforms the outcome data to remove the fixed effects, before estimating the random effects component. Then, it applies a generalized least squares estimator to obtain the estimates of the fixed effects component within its algorithm. Put differently, REML obtains the estimates of the fixed effects and random effects components separately, starting with the random effects component (McNeish and Stapleton, 2016). To appropriately adjust for the loss in the DoF we applied the Satterthwaite correction on the DoF, consequently obtaining the correct P-values and CIs (Leyrat *et al.*, 2018).

For example, let y_{ij} denote a continuous outcome from a j^{th} individual in an i^{th} cluster. A specific example of the LMM called the random intercept LMM (because it adjusts for the random cluster effects using a random intercept term in the mixed model) is given as

$$y_{ij} = \overbrace{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pij}}^{Fixed\ effects} + \overbrace{\tau_i + \varepsilon_{ij}}^{Random\ effects}, \quad i = 1, \ldots, N;\ j = 1, \ldots, n_i$$
$$\tau_i \sim N(0, \sigma_b^2);\ \varepsilon_{ij} \sim N(0, \sigma_w^2)$$

(2.15)

Model equation (2.15) is the random intercept LMM. The GzLMM models other common types of outcome data, and it is specified as

$$\eta\left(E\left(y_{ij}\right)\right) = \eta\left(\mu_{ij}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pij} + \tau_i$$

(2.16)

where y_{ij} could be a binary, count, rate, or proportion outcome variable, $\eta(.)$ is a link function that linearly relates the mean response values to both the fixed effects and random effects components of the model, β_p is the p^{th} regression coefficient to be estimated. For example, if $y_{ij} \sim Bi\left(n, \Pr\left(y_{ij} = 1\right)\right)$ then model equation (2.16) is specified using a logit link function as

$$logit\left(\Pr\left(y_{ij} = 1\right)\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pij} + \tau_i$$

(2.17)

where $\Pr\left(y_{ij} = 1\right)$ is the probability of $y_{ij} = 1$ and $logit\left(\Pr\left(y_{ij} = 1\right)\right) = \frac{\Pr\left(y_{ij}=1\right)}{\left(1 - \Pr\left(y_{ij}=1\right)\right)}$, the other parameters are the same as defined previously. The regression coefficients of the fixed effects component $\beta's$, and the constant variances for the error terms, σ_b^2(for τ_i) and σ_w^2 (for ε_{ij}) of equations (2.15) and (2.16) can be estimated using the MLE (or REML). A good example of the application of the CSM approach would be the original analysis used in the Informed Choice trial (O'Cathain *et al.*, 2002). The study was conducted to evaluate the effect of using leaflets to help women in making informed choices when assessing maternity services. The maternity units (made up of individual women) are the clusters that were randomised to the treatment arms. The primary outcome was the change in the proportion of women who agreed to have made informed choices post-intervention. The intervention effect was estimated using multilevel modelling similar to equation (2.16) with the individual-level outcomes from the women as the response values. The following individual-level covariates were adjusted for: the women's age, the age at which they left full-time education, parity, and preference for involvement in decision-making.

2.8.3.3 Population average models (PAM)

The regression models under this class are appropriate for assessing the mean intervention effect across the populations. Inferences are made across the subpopulations of the treatment arms

rather than on the individual subjects. They are formulated based on the marginal likelihoods of the correlated response values for the i^{th} cluster, $\boldsymbol{Y}_i$, hence are considered semi-parametric models. In other words, the outcomes $\boldsymbol{Y}_i$ of the i^{th} cluster are not conditionally related to the random term but only to the fixed term of the model. A PAM is like the GzLMM of model equation (2.16) but with the random effects component τ_i modelled separately, that is, the correlation among any pair of outcomes within a cluster is accounted for using a separate working covariance matrix $\boldsymbol{V}_i$. In general, a PAM is given as

$$\eta\big(E(\boldsymbol{Y}_i)\big) = \eta(\boldsymbol{\mu}_i) = \boldsymbol{X}_{pij}^{T}\,\boldsymbol{\beta}_p \tag{2.18}$$

The marginal variance of a univariate response value y_{ij} is specified as $\phi v\big(\mu_{ij}\big)$, where $v(.)$ is a known variance function and ϕ is a scale parameter and equals 1 for a binary outcome, and σ^2 (needs to be estimated) for a continuous outcome. In practice, GEE1 is the most commonly used estimator of the PAM (Twardella, Bruckner and Blettner, 2005; Offorha, Walters and Jacques, 2022).

2.8.3.3.1 First-order generalized estimating equations (GEE1)

GEE1 allows for the correlation between observations in a cluster without explicitly explaining the origin of the correlations, so there is no explicit likelihood. It is suitable when the cluster random effects and their variances are not of inherent interest, as it describes the correlation among outcomes in a cluster without explaining its source. Here, the focus is on estimating the average response to the treatments administered across the population ("population average" effects) rather than the regression parameters that would enable the prediction of the effect of changing one or more components of $\boldsymbol{X}_{ij}$ on a given individual.

GEE1 is usually used in conjunction with Huber–White SE estimates, also known as "robust SE" or "sandwich variance" estimates. Parameter estimates from the GEE1 are consistent even when the variance-covariance structure is misspecified, but the loss in efficiency could be significant, especially when the correlation is substantial (Liang and Zeger, 1986; Qu, Lindsay and Bing, 2000; Leyrat *et al.*, 2018). GEE1 is a semiparametric analytical method because it relies on the specification of the first two moments only. GEE1 as a semiparametric method is a popular alternative to the likelihood based LMM and GzLMM, which are more sensitive to covariance

structure specification. In general, fitting a GzLMM is more computationally complex and intensive than fitting GEE1. GEE1 is suitable for obtaining the parameter estimates of the model equation (2.18). It treats the correlations within a cluster as a nuisance, such that, it does not explicitly model their effect. However, GEE1 accounts for the paired correlations using a "working" covariance matrix, and the ICC characterises the working covariance matrix.

Let the univariate response value y_{ij} be the same as explained in (2.16), if its marginal probability density function (or probability mass function for discrete distribution) can be expressed as belonging to the linear exponential family distribution, then the first and second moments of y_{ij} could be solved by taking the partial derivative of the log of the moment generating function (MGF) parameterized in the mean. It is worth noting that the nuisance parameter is also contained in the MGF but without itself being estimated. The GEE1 draws its strength from the linear exponential family distribution to be further discussed in Chapter 6 (Ziegler, 2011). Liang and Zeger (1986) proposed a unique class of GEE1, that uses a working covariance matrix to solve (2.16) and it is given as

$$U_i(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{V}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) = 0 \tag{2.19}$$

where $\boldsymbol{V}_i$ is the $n_i \times n_i$ working covariance matrix for $\boldsymbol{Y}_i$ (i.e., $\boldsymbol{V}_i = Cov(Y_i)$) characterised by the working correlation matrix $\boldsymbol{R}(\alpha)$ and defined as

$$\boldsymbol{V}_i = \phi \boldsymbol{G}_i^{\frac{1}{2}} \boldsymbol{R}_i(\alpha) \boldsymbol{G}_i^{\frac{1}{2}} \tag{2.20}$$

where $\boldsymbol{G}_i = diag\{v(\mu_{i1}), \cdots, v(\mu_{in_i})\}$ is a diagonal matrix with the diagonal elements $v(\mu_{ij})$, which is the variance function for each response y_{ij}, and $\boldsymbol{R}_i(\alpha)$ is an $n_i \times n_i$ working correlation matrix specified by the ICC, α. The GEE1 estimator computes asymptotically consistent estimates $\widehat{\boldsymbol{\beta}}$, regardless of the choice of $\boldsymbol{R}_i(\alpha)$ but provided that the mean structure is correct. However, it may suffer some loss in efficiency if the choice of $\boldsymbol{R}_i(\alpha)$ is wrong (Qu, Lindsay and Bing, 2000). The parameter estimates $\widehat{\boldsymbol{\beta}}$ are iteratively obtained by alternating between a modified Fisher scoring algorithm for $\boldsymbol{\beta}$ and the moment estimation of α and ϕ, and its residual $N^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is a multivariate Normally distributed residual with mean zero and a robust sandwich variance-covariance matrix $\boldsymbol{\xi}_i$.

A real-world example of using this approach is the PoNDER study (Morrell *et al.*, 2009), carried out to assess the effectiveness of training health visitors to recognise and treat depression among postnatal women. A total of 100 general practices (the clusters) were randomised between the intervention and control arms. The primary outcome was the proportion of women who had an initial six-week EPDS score ≥ 12, and a six-month EPDS score ≥ 12. The estimated intervention effect was the average mean difference of the primary outcome in all the clusters belonging to the intervention against those of the control. This estimate was obtained using the GzLM with the model coefficients estimated with GEE1 (in conjunction with robust SE and an exchangeable correlation structure).

2.8.3.4 Relationship between the parameter estimates of CSM and PAM

Recall that model equation (2.15) represents a CSM while (2.18) represents a PAM, it is obvious that both models are targeting different inferences and estimates of the intervention effect. In general, the interpretation of the estimates of the two models' coefficients is not the same or equivalent, except in some cases (Hubbard *et al.*, 2010). When y_{ij} is modelled using linear (or log-linear) regression model with only the treatment arm indicator covariate x_{1i} included, LMM of CSM and GEE1 of PAM are targeting the same estimate/inference of the intervention effect. Hence, the estimate of β_1 has the same interpretation for both CSM and PAM as the average change in the population means between the intervention vs control groups. However, if the CSM model has any individual-level covariate included, the interpretations of the parameter estimate of the intervention effect between CSM and PAM are not equivalent (Hubbard *et al.*, 2010).

The differences in the interpretation of the treatment effect β_1 between CSM and PAM is more profound for non-linear regression models – Logistic CSM vs. Logistic PAM. In these two cases, the estimate $\hat{\beta}_1^{\dagger}$, is interpreted as the log-odds ratio and not the mean changes, where † indicates that the interpretation of this later estimate is different from the former $\hat{\beta}_1$. Furthermore, $\hat{\beta}_1^{\dagger}$ from the intercept term logistic CSM is interpreted as what would be the average cluster chance (i.e., log OR) of observing the outcome of interest if a participant receives the intervention treatment compared to the participant receiving the control. For a logistic PAM with an exchangeable correlation structure, it is the average chance of observing the outcome across all clusters in the intervention arm vs those in the control arm. In Thompson et al. (2022) an equation that numerically approximated the interpretation of CSM and PAM in all cases is provided as

$$\beta_{PAM} \approx \beta_{CRM}\left(\left[{16\sqrt{3}}/{15\pi}\right]^2 \sigma_b^2 + 1\right)^{-1/2} \tag{2.21}$$

In general, the interpretation of the result of the estimate of the treatment effect follows the chosen estimand framework stipulated at the beginning of the trial. Decisions about the analytical approach can unintentionally result in answering different questions about the interventions (i.e., target estimands). The 'participant-average treatment effect' answers the question 'How effective is the intervention for the average participant?' whereas the 'cluster-average treatment effect' answers the question 'How effective is the intervention for the average cluster?' (Offorha, Walters and Jacques, 2022; Kahan et al., 2023).

2.9 Ignoring clustering

In previous sections, appropriate and recognised analytical approaches for analysing outcome data from cRCTs were presented. A review of publicly funded cRCTs showed that researchers strive to use these analytical methods efficiently, however, a few still ignore clustering and use standard methods (Offorha, Walters and Jacques, 2022). The review found that about 5 in 100 primary outcomes were analysed with analytical methods that ignore clustering in the outcome data. Previously, it was found to be more severe as about 21% of trials did not account for clustering in their primary analyses (Twardella, Bruckner and Blettner, 2005).

Ignoring clustering in simple terms means not accounting for the extra variance, σ_b^2, introduced by randomising groups of participants (**Figure 2.1**). This implies that the total variance would be restricted to just $\sigma^2 = \sigma_w^2$, instead of explicitly partitioning it as $\sigma^2 = \sigma_w^2 + \sigma_b^2$. That is to say, the total variance of the outcomes would be smaller than it should be. The consequences of ignoring clustering may appear to be little but they are impactful, especially when the presence of σ_b^2 is substantial. Ignoring clustering would result in inaccurate SEs, and consequently an increased inferential test statistic, smaller P-value, and highly precise CIs. All these could increase the chances of obtaining a false positive result, even when the Null hypothesis of "no difference" between treatment arms is true (Bland, 2004; Christie, O'Halloran and Stevenson, 2009) .

For example, one of the work packages in the report of Perez et al. (2016) was a cRCT. The study was aimed at assessing the effect of two interventions (theory-based educational vs. a postal information campaign), and a "*treatment as usual*" control on early detection of individuals at high

risk of developing psychosis following effective care after first-episode psychosis. About 104 general practices serving as the cluster units were randomised among the three treatment arms. The key outcome was a count data type - the number of referrals per practice. This outcome datasets were analysed using standard Poisson regression models with practice size and the number of GPs in each practice included as covariates.

2.10 Consolidated Standards for Reporting Trials

The Consolidated Standards of Reporting Trials (CONSORT) statement was first published in 1996 to guide the reporting of RCTs (Begg *et al.*, 1996). The extension of the CONSORT statement to cover cluster trials was first suggested in 2001 (Elbourne, Campbell and D.R., 2001), and was then extended in 2004 (Campbell, Elbourne and Altman, 2004) which was based on the revision of the CONSORT statement in 2001. There were still inadequacies in the reporting of RCTs; hence, in 2010, the previous version of 2001 was updated (Schulz *et al.*, 2010). The 2012 extension to cover cluster trials was based on this updated CONSORT 2010 statement (Campbell *et al.*, 2012).

These guidelines are meant to aid researchers in the planning, conducting, analysing, and reporting of cluster trials to reduce the problems occurring from the poor reporting of cRCTs. Most of the information extracted from each trial reviewed in this study is based on CONSORT 2010 statement: extension to cluster randomised trials. Adherence to the CONSORT reporting guidelines for cluster trials and its impact on the quality of reporting cluster trials has attracted the interest of researchers since it was published (Ivers *et al.*, 2011; Rutterford *et al.*, 2015; Agbla and DiazOrdaz, 2018).

The adherence to different aspects of the CONSORT reporting guidelines for cluster trials has been of interest to researchers, for example, a review found that though some aspects of treatment compliance by the participants in the studies are reported, in general, comprehensive reporting of treatment compliance by participants is poor and inadequate (Agbla and DiazOrdaz, 2018). Another review concluded that despite the availability of the CONSORT reporting guidelines for cluster trials, the reporting of all aspects of sample size calculation is inadequate (Rutterford *et al.*, 2015). Ivers et al. (2011) went a step further and investigated adherence to all the new items included in the CONSORT extension for cluster randomised trials; they found that improvement was only evident in a few aspects, while in general, the adherence to the CONSORT statement

extension for cluster trials was inadequate. The success of any guideline can be measured by the rate of its implementation in practice (Gogtay, 2019).

One of the objectives of the practice review of Chapter 4 is to contribute to the debate in the literature on the adherence to the CONSORT reporting guidelines extension for cluster trials; I focussed only on the aspect of the reporting quality of the observed ICC in the papers that were reviewed, and this is because the final research questions of my thesis have not yet been determined at the time conducting the practice review in Chapter 4. It is justifiable to investigate how well the extended CONSORT reporting guidelines for cluster trials is been implemented in practice, and consequently recommend how to improve the quality of reporting cluster randomised trials (if necessary).

2.11 Summary

Having introduced the general overview, research questions, aim, and objectives of this research in Chapter 1. This chapter discussed the rationale for choosing a cRCT design over an RCT design, types of randomisations, description of the ICC, power of a cRCT, the classical methods for analysing outcome data from cRCTs, and the CONSORT standard reporting guidelines extension for cRCTs. Chapter 3 describes a methodological scoping review involving a systematic search of 5 bibliographic databases to identify other appropriate and available statistical methods for analysing outcome data from cRCTs. This is followed by Chapter 4 which reviewed and summarised the statistical methods that have been used in practice to analyse outcome data from cRCTs. In Chapter 4, the online table of content of the five NIHR Journals Library were searched chronologically to review the analytical methods used.

Chapter 3

A methodological scoping review of methods for analysing outcome data from cRCTs

3.1 Introduction

In Chapter 2, the classical methods for analysing outcome data from cRCTs were described and examples of their usage were provided. This chapter concerns a methodological scoping review to identify statistical methods available in the literature, in addition to the "classical" methods described previously in Chapter 2, that can be used to analyse outcome data from cRCTs. This involved systematic searches, auditing, and synthesising of the literature. The scope of this review covered three types of studies; they are:

- Studies that proposed a statistical method for analysing cRCTs.
- Studies that refined an already existing statistical method.
- Studies that compared the analytical methods for analysing outcome data from cRCTs.

Several statistical methods have been proposed that are appropriate for the analysis of cRCTs, most of these methods are multilevel to account for the nesting of individuals within the clusters (CORNFIELD, 1978; Donner, 1985, 1998; Donner and Klar, 1994; Allan, Donner ; Neil, 2001). Due to the rapid increase in the use of the cRCT design to evaluate the effectiveness of public health interventions, social interventions, educational policies, new health technologies, new drugs, etc., there has been an accompanied rapid methodological development, refining of methods, and also comparisons of these methods to find the most efficient and optimal method(s) for analysing outcome data from cRCTs for different data types and scenarios (Bellamy *et al.*, 2000; Omar *et al.*, 2000; Austin, 2007, 2010; Ukoumunne, Carlin and Gulliford, 2007; Hubbard *et al.*, 2010; Walters, Morrell and Slade, 2011; Leyrat *et al.*, 2018).

Several methodological reviews on the different aspects of cRCTs exist in the literature (Donner, Brown and Brasher, 1990; Chuang, Hripcsak and Jenders, 2000; Murray *et al.*, 2008; Fiero *et al.*, 2016), but to the best of my knowledge, this is the first methodological scoping review on papers proposing new methods, refining existing ones, or comparing existing methods for analysing outcome data from cRCTs.

3.2 Chapter aim

This chapter presents the results of a methodological scoping review of the literature conducted to identify the available and appropriate methods for analysing outcome data from cRCTs.

3.3 Aims of the review

The primary aim, of this methodological scoping review, is to identify appropriate statistical methods, that have been used or can be used to analyse outcome data from cRCTs.

The secondary aims are:

1. To briefly describe the identified methods.
2. To identify the frequency at which each method was studied.
3. To identify and summarise the gaps in knowledge, regarding appropriate statistical methods for analysing outcome data from cRCTs.

3.4 Methods

This scoping review provides a comprehensive overview by mapping the evidence in the published literature on the development, and refinement of methods for analysing cRCTs, and comparisons that have been made to evaluate the statistical properties of the methods. Arksey and O'Malley (2005) were used as a guide for the methodological framework of this review, for a more detailed guide in preparing and describing all the sections and Sections in this review, the Joanna Briggs Institute Scoping Review Methods Group published guidelines were used (Peters *et al.*, 2020).

The rationale for choosing a scoping review is to comprehensively review both the published and unpublished literature on the broad topic of "statistical methods for analysing cRCTs". A scoping review is often used to conduct an initial investigation to clarify a concept, identify gaps in

knowledge, or identify the scope of the body of evidence in the literature (Grant and Booth, 2009). This methodological review focuses on identifying what has been done, the scope, and the size of the research regarding new methods, or refinement or comparison of the methods for analysing outcome data from cRCTs. For scientific transparency and integrity, the protocol for this review was registered with Open Science Framework (OSF) via the website https://osf.io/8erfk (last assessed on the 8th of December 2022).

3.4.1 Research questions

The following are the research questions that this current review is focussed on answering:

1. What are the available statistical methods for analysing outcome data from cRCTs?
2. How well do these identified methods perform when their statistical properties are compared?
3. What are the methodological gaps that exist in the literature on the comparison of the methods?
4. What are potential areas to explore further in future research?

This review employed the PRISMA-ScR (Preferred Reporting Items for Systematic Review and Meta-Analysis to Scoping Review) extended guidelines for conducting a scoping review, as a guide to ensure high-quality and comprehensive reporting (Tricco et al., 2018).

3.4.2 Sources and search strategy

To develop an efficient search strategy an experienced information specialist was consulted throughout this review. The search terms were used systematically to search the electronic bibliographic databases of MEDLINE, EMBASE, PsycINFO, CINAHL, and SCOPUS.

To identify unpublished (grey or difficult-to-locate) literature the databases of Web-of-Science, Scopus conference proceedings, and OpenGrey were searched electronically. Pearl growing was also employed to help identify all (or most) of the relevant literature. In pearl growing, the reference list of included key papers were hand-searched to identify more relevant studies, in the same manner, the reference lists of the relevant papers that were identified from the included articles were investigated. The same search strategy was used to search each of the databases, minor

adjustment was made for the Scopus database search. The adjustment is the use of double quotation " " marks when searching a phrase whereas parentheses ( ) were used in other databases.

The search was from 1st January 2003 to 19th December 2020. Firstly, the primary justification for choosing 2003 as the starting year, is because this was a year prior to the extension of the CONSORT reporting guidelines to cover cRCTs (Campbell, Elbourne and Altman, 2004), hence choosing the time period would make it easier to identify most (if not all) of the published reports of cRCTs. Secondly, a review on cRCTs similarly to this currect review was published in March 2004 by Murray (Murray *et al.*, 2004), so it is ideal that I do not duplicate the findings in that publication by setting my search period far beyond 2004. Recently, a review was conducted as an update to that of Murray's 2004 review, however, the review focussed on all aspects of recent methodological developments and less on the statistical methods for analysing cluster randomised trials (Turner, 2017). Lastly, limiting the searching period to January 2003 – December 2020 would make the number of articles identified to be more manageable, and big enough to adequately power this current review.

Although the increased use of cRCT design dates to 1980, before that there were few publications on cRCTs. A further slight increase in the use of cRCT design was recorded after a methodological review by Donner, Brown and Brasher (1990), and in 2008 there were over 120 cRCTs recorded in the literature (Moberg and Kramer, 2015). It is worth noting that only published and unpublished (grey) studies in English were considered due to limited resources.

3.4.3 Search strategy

Box 3.1 presents the search strategy used to search the five bibliographical databases – MEDLINE, PsycINFO, EMBASE, SCOPUS, and CINAHL.

Box 3.1 Search terms used to systematically search the five bibliographic databases

MEDLINE/PsycINFO/EMBASE via Ovid
1. (statistic* model* or statistic* method*).mp.
2. (group randomi* or community randomi*).ti.
3. cluster.tw.
4. 2 or 3
5. randomi* controlled trial.ti.
6. randomi*.mp.
7. placebo.mp.
8. 5 or 6 or 7
9. 1 and 4 and 8

SCOPUS and CINAHL (via Ebsco)
1. "statistic* model*" or "statistic* method*"
2. "group randomi*" or "community randomi*"
3. cluster
4. 2 or 3
5. randomi* controlled trial
6. randomi*
7. placebo
8. 5 or 6 or 7
9. 1 and 4 and 8

Limiters:
1. Language: English only
2. Time frame: from 1st January 2003 to 19th December 2020.

3.4.4 Eligibility criteria

For a study to be considered for inclusion, it must satisfy the following PICOTS (Population, Intervention, Comparator, Outcome, Timing, and Study design/Settings) criteria explained below:

Population: Studies that reported the proposal of a statistical method(s) or refined or compared already existing methods of estimating the intervention effect. The comparison should be based on their frequentist statistical properties to evaluate the performance of the methods. The frequentist statistical properties of the methods were of interest because they focus on the long-run performance of the methods.

Intervention: Statistical methods used or can be used to estimate the intervention effect in cRCTs.

Comparator: Not applicable, this is a methodological review, and the interest is in method papers.

Outcome: The primary outcome of interest is the method/estimator that was used to estimate the intervention effect in a cRCT.

Timing: The time the article was published online will be considered, only articles published from 1st January 2003 to 19th December 2020 will be considered. The reasons for choosing this time frame are stated in Section 3.4.2 above.

Settings/Design: Only studies where clusters/groups of individuals were randomised to the treatment arms are considered.

3.4.4.1 Inclusion criteria

Studies to be included would have satisfied the PICOTS specified in Section 3.4.4 above. Specifically, the inclusion criteria are:

1. Only articles whose primary target of inference is the intervention effect with clustering considered, are included.
2. Only articles written in English are to be included due to limited resources.
3. Articles that reported enough information that is of interest in this review as indicated in the pre-piloted standardized data collection form.
4. Only methods papers (that proposed or refined or compared statistical methods for analysing cRCTs) are included.

3.4.4.2 Exclusion criteria

1. Articles that only applied (or described) statistical methods to analyse outcome data from cRCTs were excluded. That is studies that only report the results of a cRCT were excluded unless it reports the results from comparing different statistical methods for analysing cRCT.
2. Articles that reported other study designs such as quasi-experimental design, observational studies, editorials, reviews, tutorials, surveys, and RCT were excluded.
3. Articles where the full-text report is not available via open source, subscriptions, or contact with the author for a copy.

4. Other reports such as thesis, dissertation, books, protocols, feasibility, and pilot studies were excluded.

3.4.5 Study selection

Studies identified through the systematic searches were imported into Zotero (version 6.0.26) referencing manager for processing. Duplicates and ineligible studies were removed. Only studies that met all the inclusion criteria were processed for full-text retrieval. Any further disagreement after subjecting an article to the inclusion and exclusion criteria is addressed through extensive discussion with my supervisors. A PRISMA flow chart is presented showing the search and selection processes of the articles, and the number of studies excluded at each stage with reasons provided (**Figure 3.1**).

3.4.6 Data extraction

A standardised online spreadsheet that serves as the data extraction tool was developed and piloted on 20 randomly selected studies. The data collection tool was adjusted appropriately after the piloting phase. The relevant information extracted includes but is not limited to the study characteristics such as study title, an article full-text online link, lead author, journal, volume, issue, year of publication, and origin of publication.

Methodological characteristics such as; the purpose of the study (proposed a new method, or refined or compared already existing methods), the method for accounting for clustering, a brief description of the method, the level of clustering considered, the principle of analysis, type of cRCT design, method of parameter estimation, statistical properties evaluated (e.g. Type I error, Type II error, bias, power, coverage, etc), level of analysis, method of allocation (stratification, pair-matching, minimisation), data used to facilitate comparison (simulation or example data), type of model (linear vs non-linear), type of follow-up/data collection (cohort, cross-sectional, etc), the method used to address missing data, univariate vs multivariate outcome. Also, the results from the analytical method such as the intervention effect estimate, and its associated CI, SE, and P-value. Other relevant information such as software used for the analysis and the recommendations made were extracted.

3.4.7 Assessing risk of bias

This is a methodological scoping review so no formal assessment of the risk of bias was done on the individual studies that were included.

3.4.8 Data summary and result synthesise

The study and methodological characteristics extracted were summarised using frequencies and percentages, this was used to produce a narrative report. Since this review is mainly a methodological review, a brief description of the methods identified is provided. Numerical characteristics were summarised using the median (and interquartile range). A summary of the gaps in knowledge identified was presented at the end of this chapter, as this is one of the major rationales of the study.

3.4.9 Ethics and Dissemination

This is a methodological review, so the data collated have no direct link or cause direct harm to the participants in the individual studies included, hence no ethical approval was required. This study will inform various stakeholders in cRCTs on what has been done so far and what could be done in the future. The findings of this study were publicised through journal publications, and oral, and poster conference presentations.

3.5 Results

The literature search identified 1573 articles, and with duplicates removed, 1073 articles were remaining. After the title and abstract of each of the identified articles were screened, 116 articles were shortlisted with 55 finally included in the list of relevant articles for data extraction (these include 12 articles identified through pearl growing). These articles are methodological and application papers and are referenced throughout. Seventy-three other articles were excluded for other reasons provided (**Figure 3.1**).

3.5.1 Study general characteristics

The search and selection process of the included articles is summarised in **Figure 3.1**. Among the 55 included papers; 34 (62%) compared already existing methods, 25% of the papers proposed new statistical estimators, and 13% refined already existing ones. There is no clear pattern in the development, advancement, or comparison of statistical methods for analysing cRCTs in the last two decades approximately (**Figure 3.2**). In this review, 15 articles studied cluster-level methods for analysing cRCTs (27%, 15/55), and all the articles were comparative studies. Two of the studies compared the unweighted cluster-level t-test and Wilcoxon rank-sum test to other methods (Pacheco *et al.*, 2009; Walters, Morrell and Slade, 2011). The number of times each method was studied in the 55 articles and their references are summarised in **Table 3.1**. The methodological characteristics of each of the articles are summarised in **Table 3.2**. This review identified 27 unique statistical methods for analysing outcome data from cRCTs which are briefly described in **Table 3.3**, and these methods were studied 112 times in total (**Table 3.1**). Regression models with parameters estimated by GEE1 were the most studied (23/112, 21%) followed by MLE (16%). Among the newer methods, QIF was the most studied (5%).

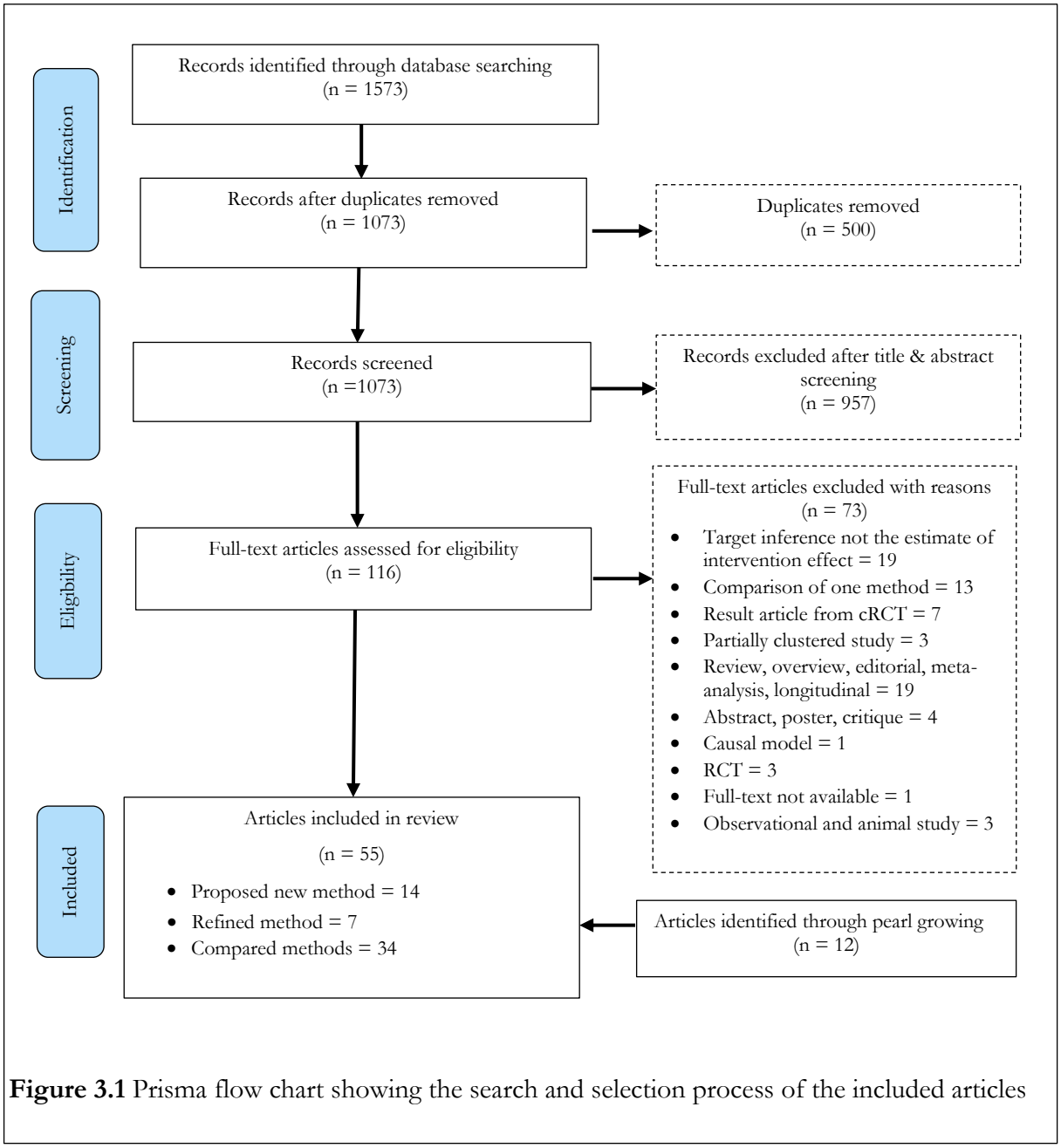


Figure 3.1 Prisma flow chart showing the search and selection process of the included articles

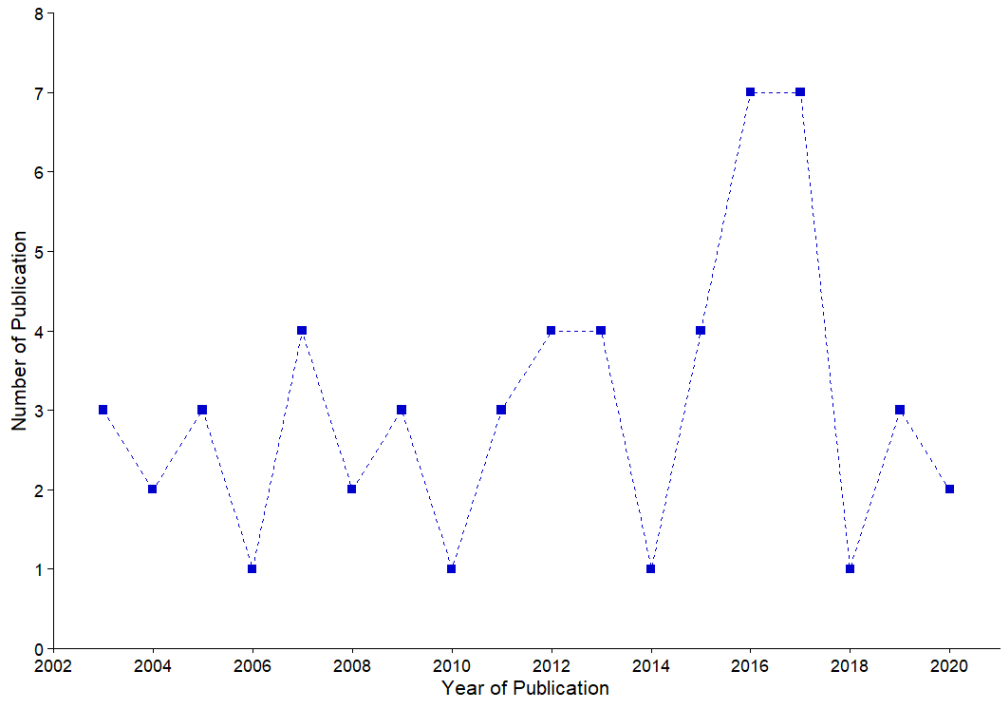


Figure 3.2 Trend of published papers on statistical methods for analysing outcome data from cRCTs, from January 2003 to December 2020.

Table 3.1 The frequency of study of each statistical method for analysing outcome data from cRCTs (N = 112)

Statistical method for parameter estimation	n	Reference
Cluster-level analysis		
Generalized least squares (GLS)	1	(Du and Lee, 2019)
Weighted Jack-knife	1	(Du and Lee, 2019)
t-test	7	(Austin, 2007, 2010; Ukoumunne *et al.*, 2008; Pacheco *et al.*, 2009; Walters, Morrell and Slade, 2011; Peek, Goud and De Keizer, 2013; Hossain *et al.*, 2017)
Wilcoxon rank sum test	2	(Austin, 2007; Leyrat *et al.*, 2018)
Kruskal Wallis	1	(Kim *et al.*, 2006)
Permutation test	3	(Murray *et al.*, 2006; Austin, 2007; Wang *et al.*, 2017)
Ordinary least squares (OLS)	4	(Ukoumunne, Carlin and Gulliford, 2007; Walters, Morrell and Slade, 2011; McNeish and Stapleton, 2016; Du and Lee, 2019)
Weighted least squares	2	(Ukoumunne, Carlin and Gulliford, 2007; Johnson *et al.*, 2015)
Restricted MLE (REML)	1	(Ukoumunne, Carlin and Gulliford, 2007)
Individual-level analysis		
Quadratic inference function (QIF)	5	(Westgate, 2012; Westgate and Braun, 2012, 2013; Yang and Liao, 2017; Yu, Li and Turner, 2020)
First-order generalized estimating equations (GEE1)	23	(Heo and Leon, 2005; Kim et al., 2006; Austin, 2007, 2010; Ukoumunne et al., 2008; Ma et al., 2009, 2013; Pacheco et al., 2009; Walters, Morrell and Slade, 2011; Yelland, Salter and Ryan, 2011; Westgate, 2012; Westgate and Braun, 2012, 2013; Peek, Goud and De Keizer, 2013; Forbes et al., 2015; Yelland et al., 2015; McNeish and Stapleton, 2016; Morgan et al., 2016; Barker et al., 2017; Hossain et al., 2017; Leyrat et al., 2018; Borhan et al., 2019; Yu, Li and Turner, 2020)
Corrected orthogonalized residual	1	(Perin and Preisser, 2016)
Quantile-GEE1	1	(Bossoli and Bottai, 2018)
AUGEE-IPW†	1	(Prague *et al.*, 2016)
Maximum likelihood estimation (MLE)	18	(Lam and Ip, 2003; Peters et al., 2003; Heo and Leon, 2005; Kang, Lee and Lee, 2005; Kim et al., 2006; Young et al., 2007; Olsen et al., 2008; Ma et al., 2009, 2013; Yelland, Salter and Ryan, 2011; Sauzet et al., 2013; Charvat et al., 2016; McNeish and Stapleton, 2016; Barker et al., 2017; Hossain et al., 2017; Pedroza and Truong, 2017; Chen and Wang, 2019; Tawiah et al., 2019)
REML	11	(Lam and Ip, 2003; Molas and Lesaffre, 2010; Walters, Morrell and Slade, 2011; Sauzet et al., 2013; Johnson et al., 2015; McNeish and Stapleton, 2016; Leyrat et al., 2018; Borhan et al., 2019, 2020; Du and Lee, 2019; Tawiah et al., 2019)

N = The total number of times the methods were studied, n = The number of times each method was studied. †AUGEE-IPW: Augmented generalized estimating equations – inverse probability weighted.

Table 3.1 The frequency of study of each statistical method for analysing cRCTs (N = 112) (cont'd)

	Individual-level analysis	
Statistical method for parameter estimation	**n**	**Reference**
Iteratively reweighted least squares (IRLS)	1	(Borhan *et al.*, 2020)
Multivariate penalized likelihood (MPL)	1	(Chen and Wang, 2019)
Hierarchical likelihood (HL)	2	(Kang, Lee and Lee, 2005; Christian, Ha and Jeong, 2016)
Hierarchical likelihood-Laplace (HLA)	1	(Kang, Lee and Lee, 2005)
Penalized quasi-likelihood (PQL)	3	(Heo and Leon, 2005; Olsen *et al.*, 2008; Yelland, Salter and Ryan, 2011)
Pseudo-likelihood (PL)	1	(Pacheco *et al.*, 2009)
Pseudo-likelihood – risk set (PLRS)	1	(Lu and Wang, 2005)
Adjusted Chi-square test	3	(Kim *et al.*, 2006; Austin, 2007; Peek, Goud and De Keizer, 2013)
Two-stage estimator	1	(Chen and Yu, 2012)
Quantile estimator	1	(Cai and Kim, 2003)
Bayesian methods (with MCMC)	13	(Peters et al., 2003; Thompson, Warn and Turner, 2004; Müller, Quintana and Rosner, 2007; Olsen et al., 2008; Ma et al., 2009; Pacheco et al., 2009; Clark et al., 2010; Brown, 2013; Ho et al., 2013; McNeish and Stapleton, 2016; Li, Xu and Shen, 2017; Pan, Cai and Wang, 2017; Pedroza and Truong, 2017)
Ordinary least squares (OLS)	1	(Borhan *et al.*, 2019)
Targeted maximum likelihood estimation (TMLE)	1	(Balzer, Petersen and J, 2016)

N = The total number of times the methods were studied, n = The number of times each method was studied

3.5.2 Study methodological characteristics

The common methodological characteristics of the included studies are summarised in **Table 3.2**. Statistical methods that are appropriate for binary outcomes were studied more than other types of outcomes (43%, 24/55), and few studies proposed estimators that can be used to assess more than one type of outcome simultaneously (**Table 3.2**). The impact of missing data on the performance of the methods was assessed in 13 studies, and most of the studies handled missing data using only available cases, otherwise known as complete case analysis (4/55, 7%), or with a combination of missing data analytical methods (4%). Although most of the studies did not clearly state how missing outcome data was handled (83%).

Multiple imputation which is an advanced and highly recommended method of handling missing was used twice (4%) or in conjunction with other missing data analytical methods (4%). One study went further to propose a method of parameter estimation that incorporates a method of handling missing data in its algorithm (Prague *et al.*, 2016). Most studies used simulated data primarily but still used real-world outcome data. The median number of simulations was 1000 (IQR: 500 – 1000 simulations).

Of the 55 studies, 9(16%) used only real-world data to facilitate comparison, and another 9(16%) used only simulated data. The majority of the studies used both simulated and example outcome data (68%). The performance of the methods was evaluated based on the estimates of the intervention effect, its SE, CI, and P-value for studies that used only real-world data. Studies that conducted simulations used several performance measures, such as power, Type I error rate, mean square error, empirical SE, and coverage to evaluate the methods.

Table 3.2 Summary of the methodological characteristics of the 55 articles included

Characteristic	n (%)
Type of article (N = 55)	
Articles proposing a new method	14 (25)
Articles that refined method	7 (13)
Articles comparing methods	34 (62)
Type of outcome (N = 55)	
Continuous	11 (20)
Binary	24 (43)
Count	6 (11)
Time-to-event	9 (16)
Methods applied to more than one outcome	
Continuous and binary	3 (6)
Continuous, binary, and count	1 (2)
Binary and time-to-event	1 (2)
Type of cRCT design (N = 55)	
Parallel	48 (86)
Crossover	2 (4)
Factorial	3 (6)
Stepped wedged	1 (2)
Parallel and stepped wedge	1 (2)
Source of data analysed (N = 55)	
Empirical	9 (16)
Simulation	9 (16)
Both	37 (68)
Handling of missing data (N = 55)	
Complete case analysis	4 (7)
Multiple imputations	2 (4)
Augmented inverse probability weighted GEE1	1 (2)
Not reported[1]	46 (83)
combined methods of handling missing data	
Complete case and multiple imputations	1 (2)
Complete case, standard, and within-cluster multiple imputations	1(2)

Note: The percentages of some of the factors add up to 101%, hence, 1% was deducted from the highest. cRCT = cluster randomised controlled trial; GEE1 = First order generalized estimating equations.

[1] These are studies that did not clearly state how missing outcome data was handled.

3.5.3 Statistical methods for analysing cRCTs

The two broad classifications of the analytical approaches for analysing outcome data from cRCTs have been discussed previously in Section 2.8. This classification is based on how the statistical methods account for clustering. The discussion in this section will be based on these classifications.

3.5.3.1 Cluster-level analysis (CLA)

The common methods used in CLA are the likelihood approaches (such as t-test and regression models), resampling (e.g., permutation test), and ranking (e.g., Wilcoxon rank-sum test). With these methods, it is complicated to adjust for individual-level covariates, and they do not produce point estimate of the intervention estimate. These methods are used for hypothesis testing, and it is possible to obtain permutation-based confidence intervals for these methods (Forbes *et al.*, 2015; Morgan *et al.*, 2016; Hossain *et al.*, 2017; Leyrat *et al.*, 2018; Borhan *et al.*, 2019). Furthermore, from the Wilcoxon rank-sum test it is possible to calculate the Hodges-Lehmann point estimate of the difference in location shift for two populations, and its CI.

3.5.3.2 Individual-level analysis (ILA)

Results under this section are based on the ILA approach discussed in Section 2.8.3.

3.5.3.2a Summarised results for RMRSE

Of the 40 studies under ILA, few adjusted for clustering by correcting the naïve model SEs to robust SEs to avoid an inflated Type I error rate that could cause increased false positive results (5/40, 13%). All the studies identified were comparative and intended to evaluate the relative performance of the methods. One out of the five studies used the same correction as that of equation (2.14) (see, Section 2.8.3.1). A study used the square root of the inflation factor given in (2.9) as a correction factor to inflate the abnormally small model SEs. From McNeish and Stapleton (2016) the correction factor is given as

$$Correction\ factor = \sqrt{DE} = \sqrt{1 + (\bar{n} - 1)\rho} \tag{3.1}$$

A study also used the correction factor of (2.9), but was incorporated into the formula of the standard chi-square test statistic, this new corrected test statistic is commonly known as the adjusted chi-square test (Austin, 2007). Additionally, another study used the popular Huber and White sandwich variance estimator to correct the model SEs from a standard logistic model (Peters et al., 2003). Lastly, Cai and Kim (2003) used bootstrap and kernel smoothing methods to obtain robust SEs that account for clustering.

3.5.3.2b Summarised results for PAM

PAMs have been described in Section (2.8.3.3). GEE1 was the most studied method for adjusting for clustering under PAM (**Table 3.1**). Forty-two percent (23/55) of the papers studied GEE1 as a method for analysing outcome data from cRCTs. Prague et al. (2016) refined GEE1 and called it "Augmented inverse probability weighting" (AUG-IPW). This method utilizes the GEE1 framework to handle missing data, imbalances in baseline covariates among treatment arms, and the interaction between outcomes and baseline covariates within its algorithm. These simultaneously occurring processes enable the AUG-IPW to obtain efficient and consistent parameter estimates. An acclaimed alternative to GEE1, called QIF was identified, and among the newer methods identified, it was the most studied method (**Table 3.1**). QIF has all the good qualities of GEE1 and it is acclaimed to be more efficient than GEE1 when the working correlation structure is misspecified (Qu, Lindsay and Bing, 2000).

Table 3.3 Brief descriptions of the unique statistical methods for analysing outcome data from cRCTs that were identified

Cluster-level analysis

Statistical method	Description	Type of outcomes analysed	Key references
Ordinary least squares (OLS)	This is the standard method for estimating the unknown parameters of a simple LM by minimising the sum of squares of the residuals.	Continuous	(Ukoumunne, Carlin and Gulliford, 2007; Walters, Morrell and Slade, 2011; McNeish and Stapleton, 2016; Du and Lee, 2019)
Generalized/Weighted least squares (G/W LS)	GLS is a generalization of the OLS method used for situations where some of the main assumptions of OLS like equal variances and independence among observations are violated. WLS is a special case of GLS where the residuals are uncorrelated with weights attached to each of them.	Binary and continuous	(Ukoumunne, Carlin and Gulliford, 2007; Johnson *et al.*, 2015; Du and Lee, 2019)
t-test	This is the standard two-sample t-test applied to the cluster level summaries to test the H_0: the mean cluster level summaries of the treatment arms are equal. This method has two variants the *unweighted* and *weighted versions*.	Binary, counts, continuous.	(Austin, 2007, 2010; Ukoumunne *et al.*, 2008; Pacheco *et al.*, 2009; Walters, Morrell and Slade, 2011; Peek, Goud and De Keizer, 2013; Hossain *et al.*, 2017)
Wilcoxon rank sum	This is a nonparametric alternative to the parametric t-test that involves ranking all the cluster level summaries of the treatment arms, then the sum of the ranks of the ith treatment arm is compared to its expected sum of ranks under the null, to obtain a test statistic. The comparison is often between the treatment arm with the smaller sum of ranks and the Normal approximation to the sampling distribution.	Continuous and binary.	(Austin, 2007; Leyrat *et al.*, 2018)
Kruskal Wallis	Is a non-parametric alternative to one-way ANCOVA and extends the Wilcoxon rank sum algorithm to enable the comparison of two independent treatment arms.	Binary	(Kim *et al.*, 2006)
Weighted Jack-Knife	A refinement to the standard jack-knife procedure, here a cluster is left out instead of an individual during the resampling process.	Continuous	(Du and Lee, 2019)
Permutation test	This algorithm can be incorporated into most test statistics. It involves calculating a P-value from the samples generated from the random permutation of cluster allotment between treatment arms. These permutations should not affect the observed effect measure if the clusters of the treatment arms are from the same underlying population.	Continuous and binary	(Murray *et al.*, 2006; Austin, 2007; Wang *et al.*, 2017)

Table 3.3 Brief descriptions of the unique statistical methods for analysing outcome data from cRCTs that were identified (Cont'd)

Statistical method	Description	Type of outcome analysed	Key references
	Individual-level analysis		
GEE1	GEE1 models the fixed effects of covariates and random effects of clusters separately using a working correlation structure. It focuses on modelling the mean parameters and considers the association parameters as a nuisance; however, it is consistent even when the working correlation structure is not correctly specified but may suffer a loss in efficiency.	Binary, continuous, and counts	(Ukoumunne, Carlin and Gulliford, 2007; Walters, Morrell and Slade, 2011; Yelland, Salter and Ryan, 2011; Westgate and Braun, 2013; Forbes *et al.*, 2015; Yelland *et al.*, 2015; Morgan *et al.*, 2017; Leyrat *et al.*, 2018; Yu, Li and Turner, 2020)
Quantile GEE1	This is a variant of GEE1 for modelling marginal quantiles with the working correlation structure specified by odds ratios instead of correlations.	Continuous	(Bossoli and Bottai, 2018)
QIF	QIF is an acclaimed alternative to GEE1 for obtaining the parameter estimates of PAM. It approximates the inverse of the working correlation structure of the GEE1 algorithm by summing the multiplication of known basis matrices and their unknown coefficients to form a linear combination. QIF is said to be efficient even when the correlation structure is misspecified since it avoids the direct estimation of the correlation.	Binary, continuous, and counts	(Westgate, 2012; Westgate and Braun, 2012, 2013; Yang and Liao, 2017; Yu, Li and Turner, 2020)
Corrected orthogonalized residual	This is an extension to the orthogonalized residuals (ORs) method to correct the finite small sample size biases in the sandwich covariance estimator of its algorithm. ORs is a generalisation of alternating logistic regression in which odds ratios are used to model the correlation within a cluster instead of correlation parameters.	Binary	(Perin and Preisser, 2016)
Quantile estimator	A non-parametric method of estimating the parameters of a quantile survival regression model with correlated failure times.	Time-to-event	(Cai and Kim, 2003)
AUGEE1-IPW	AUG-IPW is a doubly robust method for estimating the marginal intervention effect from a cRCT with missing at random continuous outcome data, imbalance in baseline covariate, and interactions between outcome variable and baseline covariates.	Continuous	(Prague *et al.*, 2016)
MLE	This involves obtaining parameter estimates of a statistical model that maximise the of realizing the observed data. Usually, when it involves a random-effects model the integral of the log-likelihood has no analytical solution, especially when the number of clusters is big. The consensus is to carry out numerical approximation or integration using techniques such as Newton-Raphson (NR), expectation-maximization (EM) algorithm, Gaussian, and adaptive Gaussian quadrature.	Continuous, binary, counts, time-to-event, and nominal.	(Lam and Ip, 2003; Peters et al., 2003; Heo and Leon, 2005; Kang, Lee and Lee, 2005; Kim et al., 2006; Young et al., 2007; Olsen et al., 2008; Ma et al., 2009, 2013; Yelland, Salter and Ryan, 2011; Sauzet et al., 2013; Charvat et al., 2016; McNeish and Stapleton, 2016; Barker et al., 2017; Hossain and Bartlett, 2017; Hossain et al., 2017; Pedroza and Truong, 2017; Chen and Wang, 2019; Tawiah et al., 2019)
REML	REML is a special case of MLE, and it is also called "reduced or residual maximum likelihood". While MLE maximises the information from the complete data set, REML uses the information from a uniquely reduced version of the data set to obtain estimates of the parameters of a model.	Continuous, binary, counts, time-to-event, and nominal.	(Lam and Ip, 2003; Molas and Lesaffre, 2010; Walters, Morrell and Slade, 2011; Sauzet *et al.*, 2013; Johnson *et al.*, 2015; McNeish and Stapleton, 2016; Leyrat *et al.*, 2018; Borhan *et al.*, 2019, 2020; Du and Lee, 2019; Tawiah *et al.*, 2019)

Table 3.3 Brief descriptions of the unique statistical methods for analysing outcome data from cRCTs that were identified (Cont'd)

Individual-level analysis

Statistical method	Description	Type of outcomes analysed	Key references
IRLS	This is an extension to the weighted least squares method which involves an iterative method of solving the weighted least squares in each step. It is commonly used to obtain the ML estimates of GzLM.	Counts	(Borhan *et al.*, 2020)
Multivariate penalized likelihood	This method was proposed to obtain the parameter estimates of a joint model with multiple types of outcomes – binary and time-to-event outcomes. It combines the theories of penalized partial likelihood (with Laplace approximation) and jack-knife resampling techniques to obtain consistent SE estimates.	Binary and time-to-event	(Chen and Wang, 2020)
Penalized quasi-likelihood	PQL is well suited for non-linear regression, it uses Taylor's series expansion to approximate the quasi-likelihood function of a distribution, to obtain estimates of the parameters of interest when the likelihood function has no analytical solution.	Continuous and binary	(Heo and Leon, 2005; Austin, 2007; Olsen *et al.*, 2008; Yelland, Salter and Ryan, 2011)
Hierarchical likelihood	This method maximizes a generalization of Henderson's joint likelihood (*h-likelihood)* from hierarchical HzGLMs using PQL on the *h-likelihood* to estimate the fixed effects component and on a generalization of the restricted maximum likelihood to estimate the dispersion components.	Binary	(Kang, Lee and Lee, 2005; Christian, Ha and Jeong, 2016)
Hierarchical likelihood Laplace	Is a modified version of the Hierarchical likelihood estimator that uses the Laplace algorithm instead of PQL.	Binary	(Kang, Lee and Lee, 2005)
Pseudo-likelihood	This is an alternative to full likelihood estimation to solve the problem of the intractability of the full likelihood estimators. It involves representing the joint full likelihood function with simplified approximates called pseudo likelihoods to obtain a tractable solution.	Binary and time-to-event	(Pacheco *et al.*, 2009)
Pseudo-likelihood – risk set	This is a modified version of the original Pseudo-likelihood in which a risk set sampling procedure is used to formulate the Pseudo-likelihood.	Time-to-event	(Lu and Wang, 2005)
Adjusted Chi-square test	This is an extension of the standard chi-square test for independent outcomes to allow for clustering. This is achieved by calculating a correlation correction factor for each treatment arm, which is incorporated into the standard chi-square formula.	Binary	(Kim *et al.*, 2006; Austin, 2007; Peek, Goud and De Keizer, 2013)
Two-stage estimator	In the first stage, this estimator estimates the marginal mean parameters under the independence working assumption, and in the second stage, the correlation parameter is obtained using MLE with the marginal mean parameters plugged in. This method is mostly applied to semiparametric transformation models.	Time-to-event	(Chen and Yu, 2012)
Bayesian methods (with MCMC)	This is a numerical integration method in the Bayesian multilevel regression model, MCMC iteratively generates a series of estimates using previous samples to randomly generate the next samples needed for estimating the posterior parameters.	Continuous, binary, counts and time-to-event	(Peters et al., 2003; Thompson, Warn and Turner, 2004; Müller, Quintana and Rosner, 2007; Olsen et al., 2008; Ma et al., 2009; Pacheco et al., 2009; Clark et al., 2010; Brown, 2013; Ho et al., 2013; McNeish and Stapleton, 2016; Li, Xu and Shen, 2017; Pan, Cai and Wang, 2017; Pedroza and Truong, 2017)
Targeted maximum likelihood estimation	Originally developed for causal inference modelling, TMLE is becoming an increasingly popular alternative to MLE. It's a doubly robust and efficient semiparametric substitution parameter estimating method.	Binary	(Balzer *et al.*, 2016)

3.5.3.2c Summarised result for CSM

CSMs are typified by conditional regression models with GzLMM and frailty survival models being common examples for this group. A CSM has been briefly explained in Section 2.8.3.2. Of the unique 27 statistical methods identified (**Table 3.3**), 20 (80%) occurred under the CSM approach (**Table 3.1**), and 2 methods occurred under CSMs and PAMs. MLE is the most studied estimator under the CSMs (18/112, 16%). Among the Twenty unique methods for obtaining parameter estimates of a CSM, the most common methods are MLE, REML, Bayesian methods adjusted chi-square, and PQL.

3.5.4 Comparison of the statistical methods

Thirty-four (62%) of the 55 articles reviewed were comparative studies, and their primary aim was to evaluate the relative performance of different methods for analysing cRCTs. Comparison papers were the highest among the three considered (**Figure 3.1**). This indicates the plausible growing interest among methodologists in evaluating the optimality of the statistical methods in different specific settings. This review identified 7 (7/34, 21%) comparative studies based on Bayesian methods (Peters et al., 2003; Thompson, Warn and Turner, 2004; Olsen et al., 2008; Ma et al., 2009; Pacheco et al., 2009; McNeish and Stapleton, 2016; Pedroza and Truong, 2017). Using empirical data, Ma et al. (2009) conducted a direct comparison between the Bayesian method and classical analytical methods with binary outcome data. The specific methods compared are OLS, REML, MLE, RMRSE (MLE with Huber sandwich estimator), and GEE1. The results showed that all the methods considered performed well, with the Bayesian logistic random effects model having the biggest SEs, hence the widest CIs.

Similarly, using an example data set, Olsen et al. (2008) compared PQL, MLE (with adaptive quadrature), and Bayesian method (with MCMC) with binary outcomes from a factorial study cRCT design involving multiple assessments. To capture the longitudinal nature of the trial the multilevel models had three levels. Their results showed that the credible interval for the Bayesian method was wider. Pacheco et al. (2009) study involving the comparison of t-test, GEE1 (with model and robust SE estimators), GzLMM, and Bayesian hierarchical model (Bayes-HM) with over-dispersed count data, supported the common knowledge that the number of clusters, and the degree of ICC do affect bias, power, and coverage. In the study, it was found that in general GzLMM and Bayes-HM performed better than the other methods in terms of good estimates of

between-cluster CV (and random effects), coverage, Type I error rate, and power (Pacheco *et al.*, 2009).

Peters et al. (2003a) conducted a comparative study of three cluster-level methods: weighted and unweighted logistic regression models, random-effects meta-regression of log odds, and five individual-level models: standard logistic regression, logistic regression with robust SEs, GEE1, random effects logistic model and Bayesian random-effects logistic model in a factorial cRCT design. The comparison was based on estimates of the intervention effect, its SE, and P-value using real-world binary outcome data. Stratification and baseline covariates were adjusted for the models. The results showed that standard logistic regression was highly anti-conservative and that the methods produced differing estimates, but with an adequate number of clusters, all the methods produced valid results except the standard individual-level logistic regression model.

Young et al. (2007) compared GzLMM (a CSM) with parameters estimated by MLE and GEE1 (a PAM) with counts outcome data from a non-randomised cRCT. The results from the trial indicated that the two methods are not robust to outliers, and both methods produced equivalent estimates of the intervention effect. Similarly, Ma et al. (2013) used simulated binary outcome data with some missingness. The interest was the performance of the two methods compared by Young et al. (2007) with a focus on accurate and efficient parameter estimations. With the assumption that the missingness is covariate dependent, missing outcomes were handled using complete case analysis, and standard multiple imputation. The results showed that GEE1 performed better in most scenarios except for a few, compared to GzLMM with MLE. The important message from these two studies is that the comparison of a PAM with parameters estimated by GEE1 and a CSM with parameters estimated by MLE is possible in certain circumstances where the parameter estimates from the two methods are equivalent as explained in Section 2.8.3.4.

Studies on a direct comparison between GEE1 and QIF were also identified in this review. Four studies (5/34, 15%) out of the total 34 comparative studies were identified (Westgate, 2012; Westgate and Braun, 2012, 2013). QIF uses the GEE1 framework but instead of inverting the working correlation matrix, which could be problematic for large matrices, QIF approximates the inverse of the working correlation matrix by summing the multiplication of known basis matrices and their unknown coefficients to form a linear combination (Song *et al.*, 2009). Further technical description of QIF is presented in Chapter 6. Westgate and Braun (2012) compared the impact of imbalances in the number of clusters, cluster sizes, and/or covariates on GEE1 and QIF with

exchangeable covariance structure for both. However, in this circumstance, GEE1 and QIF are not in the same class and hence incomparable. However, QIF was modified using correlation weights, and then modified QIF and GEE1 were then compared with binary outcome data. The comparison was based on the accuracy of the methods and was facilitated using mean square error (MSE) as the performance measure. The number of clusters (20 or 100 clusters), average cluster sizes (5-20, or 25-150), and ICC varied. Results showed that in most settings QIF was less precise than GEE1 due to the impact of the imbalance in the number of clusters and cluster covariates, rather than in the average cluster size. The study concluded that for trials with small to moderate numbers of clusters with imbalances in average cluster sizes and cluster covariates, QIF seems to produce a parameter estimate of the intervention effect with larger SE indicative of lesser precision compared to GEE1 even when the correlation structure is misspecified, as opposed to claims by.

PAMs are known to produce biased estimates for SE for small sample studies, which is also the case for QIF (Westgate and Braun, 2012). This motivated Westgate (2012) to investigate the performance of bias corrected QIF, standard QIF, and GEE1 for trials with small to moderate numbers of clusters. Two corrections were proposed for the standard QIF. The comparison was facilitated with real-world and simulated clustered outcome data mimicking both longitudinal and cRCT designs. The results showed that in general corrected QIF produced more varying estimates than expected theoretically. Also, as the number of clusters increases the performance of QIF relative to GEE1 improved, with the corrected QIF performing better. A most recent comparative study between GEE1 and QIF in the context of cRCT, a simulation study with continous outcome data was not comprehensive. Some of the parameters of the data generating mechanism of the simulation study had a single level – fixed number of clusters (100 clusters, i.e., 50 per arm) and an average cluster size of 25. The comparison was based on relative bias, empirical SE, and mean robust SE performance measures. The results showed that for a complex model with an individual-level covariate, QIF had better efficiency (less SE estimates) than GEE1 but suffered from an inflated Type I error rate (P-value > 0.05). In summary, the study found that QIF had better efficiency than GEE1 in general, but should be used with caution (Yu, Li and Turner, 2020).

3.6 Gaps in Knowledge

A summary of the major gaps in knowledge identified from the findings of this review is presented in this section. These gaps in knowledge relate to the statistical methods for analysing outcome data from cRCTs, that is, the analytical method for estimating the intervention effect. Specifically,

the focus is on gaps in knowledge regarding the comparison of the methods. Firstly, while GEE1 and QIF were developed in the context of longitudinal study design, GEE1 has enjoyed great attention from researchers in the literature of cRCT but this is not the case for QIF. QIF was developed about two decades ago and seems to have promising potential; in terms of producing an estimate of intervention effect that is more efficient (smaller SEs) especially when the correlation structure is misspecified and substantial, shown to be more robust to outliers and produces objective function like the twice negative likelihood compared to GEE1 (Qu, Lindsay and Bing, 2000; Odueyungbo *et al.*, 2008; Song *et al.*, 2009). To the best of my knowledge, only four studies have compared GEE1 and QIF in the context of cRCT (Westgate, 2012; Westgate and Braun, 2012, 2013; Yu, Li and Turner, 2020). In all the studies the comparison was based on the accuracy of the methods with MSE as the only performance measure (Westgate, 2012; Westgate and Braun, 2012, 2013), except for Westgate (2012) which included test size and coverage probability. The most recent of the four studies was not comprehensive – the number of clusters (100 clusters) and cluster size (25 participants each) were not varied (Yu, Li and Turner, 2020).

Secondly, studies have compared CSMs typified by GzLMM with parameters estimated by MLE/REML to PAMs typified by GEE1 in the context of cRCTs with binary outcome data (Ma *et al.*, 2013; Thompson *et al.*, 2022). In Ma et al. (2013), the main purpose was to compare the performance of the methods with different amounts of missing binary outcome data. The results of the study showed that in general GEE1 performed better than GzLMM on all four performance measures calculated – empirical SE, root MSE, standardised bias, and coverage probability. However, QIF has not enjoyed much evaluation against cluster-specific models, to the best of my knowledge, no study has evaluated QIF against any CSM in the context of cRCTs. It is worth knowing how well it would perform compared to popular CSMs like GzLMM.

Lastly, direct comparisons between GEE1 and QIF in the context of cRCT were based mostly on efficiency, whereas a lot is yet to be known about the relative performance of these methods in terms of their capability to maintain a reasonable Type I error rate, power, coverage probability, and convergence rate. Also, these methods have not been compared with data from complex cRCT designs like crossover, factorial, repeated cross-sectional, and stepped wedge. Additionally, as far as I know, no study has compared the performance of GEE1 and QIF when the distribution of the cluster level random effects does not follow a Normal distribution. Westgate (2012) acknowledged these gaps and called for more studies to provide answers to these scientific questions. The next chapter, Chapter 4, as a complement to this methodological review, will focus

on reviewing what statistical methods are commonly used for analysing outcome data from cRCTs in practice. The review will be based on publicly funded cluster trials published by the UK NIHR Journals Library. One of the main purposes of the practice review in Chapter 4 is to identify more research gaps.

3.7 Summary

Chapter 3 describes a methodological scoping review conducted to identify the statistical methods that have been used or can be used for analysing outcome data from cRCTs. The review identified 55 articles, 14 (25%) of the articles proposed a method, 7(13%) refined already existing methods, and the remaining 34 (62%) compared the different statistical methods. There was no clear trend in the development of statistical methods for analysing cRCTs in the past two decades considered. In total, this review identified 27 unique statistical methods for analysing cRCTs which were classified into two broad approaches: cluster-level analysis, and individual-level analysis. The individual-level analysis was discussed further on the sub-classifications: PAMs and CSMs (adapted from Walters, Morrell and Slade (2011)).

The most common methods identified were GEE1, GzLMM (with MLE/REML), Bayesian methods (with MCMC), standard t-test, QIF, PQL, permutation test, Wilcoxon rank-sum test, and adjusted chi-square test (**Table 3.1**). These findings are like that of a recent review by Turner (2017) than a previous review by Murray et al. (2004). However, it is worth noting that both studies were focussed on other aspects of cRCTs (other than the analysis of outcome data) and were more of a narrative review. This current review employed a systematic searching approach and was focused primarily on the statistical methods for analysing cRCTs; hence this review was able to identify more methods than the other previous two. It is important to re-establish the main objectives of this review, which is to identify appropriate methods that have been used or can be used to analyse cRCTs.

Some of the methods identified were selected for evaluation, and they are technically described in Chapter 6. These methods are GzLMM (with MLE/REML), GEE1, QIF, and GEE2 were selected to be potentially studied in more depth. Two of the four methods, GzLMM (with MLE/REML) and GEE1, were selected because they are the most studied methods (possibly indicating high interest rate) for analysing cRCTs as indicated in **Table 3.1**. These two methods could be considered as the "classical methods", because they are appropriate for analysing outcome

data from cRCTs, they have been around for a longer time, and their properties have been extensively studied/established in the context of cRCTs. Briefly, GzLMM regresses the outcome variable of interest on both the fixed effects of covariates and random effects of clusters. In other words, it uses a single regression equation to model both the fixed effects of covariates and random effects of clusters simultaneously (Walters, Morrell and Slade, 2011; Campbell and Walters, 2014a). The MLE and REML are two common methods for estimating the model parameters of GzLMM (Lam and Ip, 2003; Zhang, 2015; Offorha, Walters and Jacques, 2023).

Similarly, GEE1 was identified as the most studied marginal/population average model (**Table 3.1**). GEE1 is a semi-parametric method that estimates the fixed effects of covariates and the random effects of clusters separately (Liang and Zeger, 1986; Zeger and Liang, 1986). GEE1 is based on Liang (Liang and Zeger, 1986) approach of using a separate working covariance matrix characterised by a working correlation matrix to model the correlation between any two random outcomes in a cluster. Similar to GzLMM, the statistical properties of GEE1 have been extensively established in the context of cRCTs, and its use span over decades (Liang and Zeger, 1986; Zeger and Liang, 1986).

The remaining two methods, QIF and GEE2, were selected as newer/emerging methods because they were recently proposed as promising alternatives to the classical methods and have not received much attention in the context of cRCTs, hence their properties are not well understood (Turner, 2017). GEE2 is based on the framework GEE1 with some little differences. GEE2 was proposed to handle situations where the correlation between pairs of outcomes in a cluster in of primary interest, and the correlation structure has a complex form (Ziegler *et al.*, 2000). GEE2 as opposed to GEE1, simultaneously models both the fixed effects of covariates and random effects of clusters using separate estimating equations.

The last method, QIF, is also a recent alternative based on the framework of GEE1. The results of the review conducted in this chapter indicated that QIF is the most studied newer method (**Table 3.1**). QIF was proposed as an upgrade to GEE1 in situation where the correlation structure could have been misspecified (Qu, Lindsay and Bing, 2000). The impact of misspecifying the correlation structure is more severe when the correlation among outcomes in a cluster is substantial (Prentice, 1988; Prentice and Zhao, 1991; Qu, Lindsay and Bing, 2000). Misspecifying the correlation structure when there is substantial correlation among outcomes could cause potential loss in efficiency (i.e., higher variability in the estimates) (Qu, Lindsay and Bing, 2000; Westgate,

2012). QIF avoids the use of the "working correlation matrix", instead it uses the sum of the product of basis matrices and unknown constants to represent the inverse of the working correlation matrix. This strategy enables QIF to eliminate the impact of misspecifying the working correlation structure (Qu, Lindsay and Bing, 2000). However, this advantages that QIF have over GEE1 has not been comprehensively investigated in the context of cRCTs.

The common analytical approach for handling missing outcome data identified in this current review were complete case analysis, and multiple imputations. The mechanism that generated the missing outcome data do have an impact on the adequacy of the analytical approach used to handle the missing outcome data. Rubin (Rubin, 1976) classified the mechanisms in which the missing outcome data are generated into three categories, 1) missing completely at random (MCAR) – is the likelihood that the missing outcome is not dependent on the observed or unobserved variables, 2) missing at random (MAR) – occurs when the probability of missing outcome data is dependent on the observed variables alone, 3) missing not at random (MNAR) – is the likelihood of the missing outcome data is dependent on the unobserved variables (Dziura *et al.*, 2013; Austin *et al.*, 2021; Heymans and Twisk, 2022).

In complete-case analysis (CA), subjects with at least one missing value on any variable of interest are excluded from further statistical analysis, therefore, only subjects with complete set of data are included in the statistical analysis (Heymans and Twisk, 2022). Complete-case analysis is easier to conduct, has gain popularity over the years, and is adequate (and appropriate) under MCAR, and MAR with missing outcome data (Austin *et al.*, 2021). However, aside that the reduction in sample size will affect the precision of the parameter estimates and power of the study, this approach has a major limitation (Austin *et al.*, 2021).

To resolve these pitfalls of CA, a common alternative is the "imputation" analytical approach. Recently, imputation has been increasing in popularity due to the sound technicality of its procedure. In imputation, the missing values of the variables are replaced with artificial/estimated value(s). Single imputation involves using just a single value, such as mean, median or last observed value to replace the missing value. The major limitation of this approach is that the uncertainty of using estimates to replace the missing values cannot be determined, and it reduces the standard deviation of the variables (i.e., variation in the data). To circumvent this problem, another prominent approach "multiple imputation (MI)" is often used (Austin *et al.*, 2021).

In multiple imputation each of the missing values is replaced with a set of predicted values from an imputation regression model, which results in several versions of complete datasets for each of the missing values. It is recommended that the outcome variable should be part of the regression model together with other covariates that are related and/or correlated to the outcome missing variable (Austin *et al.*, 2021). There are three major phases in MI approach; imputation of the missing outcomes, analysis of the imputed complete datasets, and pooling of the estimated values to get a single final estimate (Heymans and Twisk, 2022). Multiple imputation has some advantages compared to complete case analysis. First, when the mechanism of the missing data is MCAR, MI is more precise compared to CA (although CA can be used). Second, MI has been suggested to be adequate in cases where the missing data of a variable is even more than 50% of the observed data. However, in a situation where >50% of the data is missing, that says a lot about the "not so good quality" of the observed data. It would be more reasonable to drop a variable with such high amount of missing data from further analysis. Lastly, the uncertainty regarding using estimates to replace missing values is indicated by the standard errors of the pooled estimates (Dziura *et al.*, 2013; Austin *et al.*, 2021; Heymans and Twisk, 2022).

The next chapter presents the findings of a practice review. The practice review of Chapter 4 reveals the discrepancies between methods that are available in the cRCTs literature, and the common methods used in practice which is also one of the objectives of this thesis. The findings and research gaps identified from the methodological review of this Chapter, and the practice review of Chapter 4 led to the formulation of the research questions, aim, and objectives of this research which are presented in Chapter 5.

Chapter 4

A review of statistical methods used in practice for analysing cRCTs

4.1 Introduction

This chapter reports the process and results of a review conducted to investigate the statistical methods used in practice for analysing cRCTs, to complement the immediate previous chapter. Chapter 3 presented a methodological scoping review conducted to identify available and appropriate methods for analysing outcome data from cRCTs, in addition to the classical statistical methods described in Chapter 2. The main aim of conducting these two reviews is to evaluate their findings to identify relevant discrepancies between methods that are available in the literature and those that are used in practice. This review presents the results of publicly funded cRCTs published in the online National Institute for Health and Research (NIHR) Journals Library between January 1997 and July 2021. Part of Chapter 4 has been published as an article in Trials Journal.

Established in 2006, NIHR is now the largest funder of health and social care research in England. In 2018/2019 NIHR increased its funding capacity by £91 million compared to £226 million spent in 2017/2018 on research projects. Health Technology Assessment (HTA) received the highest amount of about £113.1 million (NIHR, 2019). HTA funds a lot of cRCTs that assess the safety, clinical, and cost-effectiveness of new technologies in England. The NIHR publishes its commissioned research of high quality, in different health areas through the following journals: Public *Health Research* (PHR), *Health Services and Delivery Research* (HSDR), *Efficacy and Mechanism Evaluation* (EME*), Programme Grants for Applied Research* (PGAR) and the *Health Technology Assessment* (HTA).

This review was carried out to evaluate the statistical methods that are used in practice for analysing publicly funded cRCTs. The findings of this chapter will enable the investigation of the discrepancies between the available methods described in the literature (summarised in Chapter 3) and those used in practice for analysing cRCTs.

4.2 Chapter aim

The primary aim of this chapter is to present the process and result of a review of the statistical methods that are used in practice for analysing cRCTs. Hence this chapter addresses one of the research objectives (Section 1.5).

4.3 Aims of the practice review

The aims of this review are:

- To determine the frequency of the usage of the identified methods.
- To ascertain the appropriate approaches used in practice to adjust for clustering.
- To note common methods that are available in the literature but not used in practice.
- To evaluate patients (and clusters) recruitment and retention in practical cRCTs.
- To evaluate adherence to the CONSORT reporting guidelines for cluster randomised trials.

4.4 Methods

The processes taken to conduct this review are explained in this section.

4.4.1 Search strategy

The online table of contents of each of the five NIHR journals was manually searched from 1st January 1997 to 15th July 2021 chronologically. The time frame was chosen for the following reasons. The starting year, January 1997, was the inception of NIHR Journal Library starting with HTA. So, it is ideal to start from January 1997, if not it will be challenging if not impossible to know when the first report on cRCT was published. Furthermore, the essence of conducting a scoping review is to comprehensively synthesise evidence on a topic. This lend to the reason of searching the NIHR Journal Library from its inception (January 1997), this would allow evidence synthesis on the broad topic of interest which is the bedrock of a scoping review (Arksey and O'Malley, 2005; Munn et al., 2018; Tricco et al., 2018) . The ending date (July 2021) corresponds to the present date that the search was conducted.

The title and abstract of each report were screened to identify if a cRCT was reported. If the title and abstract did not provide sufficient information to determine whether a cluster trial was reported, the introduction and methodology chapters of the report were screened to decide if the report should be included.

4.4.2 Trial identification

To identify reports to be included in this review, we followed the procedure described in Section 4.3.1. Apart from the HTA Journal which published its first volume in 1997, the other four journals are recent editions of the NIHR Journals Library. The HSDR, PGfAR, and PHR journals published their first volume in 2013 while EME published their first volume in 2014. A search through the HTA archive from 1st January 1997 to 15th July 2021 showed that the first report of a cRCT was published in 2000 (Turner *et al.*, 2000). Nonetheless, without starting the search from 1997 it would have been impossible to know when the first cRCT was reported before searching. Also, choosing 1997 enabled the assessment of the adherence to the CONSORT reporting guidelines before and after the publication of the CONSORT 2010 statement extension for cluster randomised trials. It is important to note that only trials in which groups of individuals were randomised to the treatment arms were considered in this review.

The cRCT reports were obtained from the NIHR Journals Library website (https://www.journalslibrary.nihr.ac.uk/#/ date last accessed 9 August 2021) along with any previously published trial paper, protocol paper, or trial protocol, where available. For trials with published International Standardised Randomised Controlled Trial Number (ISRCTN), it was used to check the ISRCTN register of clinical trials for any additional information, trial website, or any previously unobtainable trial reports (cf. http://www.isrctn.com/ date last accessed 9 August 2021). The trial reports published in the NIHR Journals Library were used as the main resource when there were discrepancies in reporting.

4.4.3 Data extraction

Articles reporting cRCTs were scrutinised and relevant information was extracted from them. A standardised and piloted data extraction form was used to extract all relevant information from each included article. In cases where the information of interest was not found, "Not Reported (NR)" was used to indicate this. Missing information could be because the author(s) did not

consider or make use of the method/item of interest or might have used the method/item of interest but did not report it. All extracted information was stored in an Excel spreadsheet for further analysis. Most of the information extracted was based on the review of Walters et al. (2017) and the CONSORT reporting guidelines for cluster randomised trials (Campbell *et al.*, 2012).

Specifically, the following information was extracted: details of the article, sample size calculation, recruitment, follow-up, clustering, allocation, design/type of trial, primary outcome, primary analysis, and results. See **Appendix 1** for the list and description (where necessary) of all the items extracted, while **Appendix 2** presents the list and URL of all the included reports. The extracted information was analysed and reported following the PRISMA guidelines where applicable (Page *et al.*, 2021).

4.4.4 Analysis

During the data extraction process, it became obvious that some of the cRCT reported the results of two or more separate cRCTs (Christian, Evans and Cade, 2014; Kitchener *et al.*, 2016; Raine *et al.*, 2017; Ballard *et al.*, 2020; Foy *et al.*, 2020), and some trials reported the results of more than one primary outcome (Macarthur *et al.*, 2003; Lawlor *et al.*, 2016; Wright *et al.*, 2016; Sumnall *et al.*, 2017; Connolly *et al.*, 2018; Ramsay *et al.*, 2018; Thompson *et al.*, 2018; Wykes *et al.*, 2018; Foy *et al.*, 2020; Gaughran *et al.*, 2020). The unit of analysis in this review was mainly the trials published in the NIHR Journals Library (86 trials); in some cases, the primary outcome was the unit of analysis (100 primary outcomes in total). Where a primary outcome has been analysed multiple times (repeated measurements) or the experimental arms were more than two, the maximum number of participants in the analysis and the maximum observed ICC is reported. Results were summarised with frequencies and percentages for categorical outcomes of interest, while mean, standard deviation, range, median and interquartile range were obtained for numerical outcomes. All analysis was done using an Excel spreadsheet (Microsoft® Excel for Mac, version 16.51) and R studio (Version 1.4.1717).

4.4.5 Eligibility

Only full-scale main studies that randomised whole groups or clusters of individuals to the treatment arms were eligible (basically cRCTs). The report must be published in any of the five online NIHR Journals library, from 1 January 1997 to 15 July 2021. Other study designs were

excluded. Pilot and/or feasibility cRCTs were excluded as these have separate specific design and analysis issues including outcomes, sample size, statistical analysis and reporting. Full texts of identified reports were retrieved for further assessment.

4.5 Results

The results presented in this section have been published in Trials journal, the article is provided in **Appendix 3** (Offorha, Walters and Jacques, 2022).

4.5.1 Trial characteristics

Relevant reports published from 1st January 1997 to 15th July 2021 were retrieved from the five online NIHR Journals Library. In total, 1,942 reports were screened for eligibility, and 118 cRCTs reports met the initial inclusion criteria, of which three were stepped wedge cRCTs (Snooks *et al.*, 2018; Wykes *et al.*, 2018; Peden *et al.*, 2019). Two reports were excluded because their trials were stopped due to poor recruitment, so they only reported their qualitative findings (Speed *et al.*, 2010; Simmons *et al.*, 2016). Thirty-seven pilot/feasibility cRCTs were excluded. Seventy-nine reports containing the results of 86 cluster trials were included. Five reports contained the results of multiple trials (4 reports of 2 cRCTs each and 1 report included 4 cRCTs) (Christian, Evans and Cade, 2014; Raine *et al.*, 2017; Ramsay *et al.*, 2018; Ballard *et al.*, 2020; Foy *et al.*, 2020). A total of 100 primary outcomes (11 trials in 10 reports had multiple primary outcomes) were assessed in this review. The search and selection processes are presented in **Figure 4.1**.

Table 4.1 summarises the characteristics of the 86 trials included in this review. Most of the trials reviewed were conducted in different regions but solely within the United Kingdom (UK) except for Simmons et al. (2016) which involved other European locations. The trial design used was mostly a parallel-group cluster trial that involved a direct comparison between the intervention and control treatment arms (85%, 73/86), and this was mostly done using two experimental arms for comparison (81%) (**Table 4.1**).

4.5.2 Statistical methods used in practice

Of the 100 primary outcomes reported in the 86 trials, the data type of most of the primary outcomes was continuous (65%, 65/100), followed by binary outcomes (28%) and then counts

(5%), then time-to-event (Davies *et al.*, 2017; Peden *et al.*, 2019) was the least (2%). In the description of the statistical analysis of the primary outcomes of the cRCTs, a variety of phrases were used to describe the multilevel regression methods used to account for clustering, such as *generalized linear mixed-model, two-level hierarchical model, mixed-effects, multilevel regression, two-level heteroscedastic linear regression model,* hence, we used a generic name "generalized linear mixed model (GzLMM)" to cover all the multilevel regression methods.

Of the 100 analysed primary outcomes in the trials, 80% (80/100) used a GzLMM to account for clustering, 7% used RMRSE and 6% used GEE1 to estimate the regression coefficients for the models. Most of these analyses were carried out using individual participant outcomes (**Table 4.2**). Only 2 trials used aggregated cluster-level outcomes as data points in their primary analyses (Wright *et al.*, 2016; Cameron *et al.*, 2021). The different statistical methods used to account for the clustering of outcomes at the analysis phase are presented in **Table 4.2**.

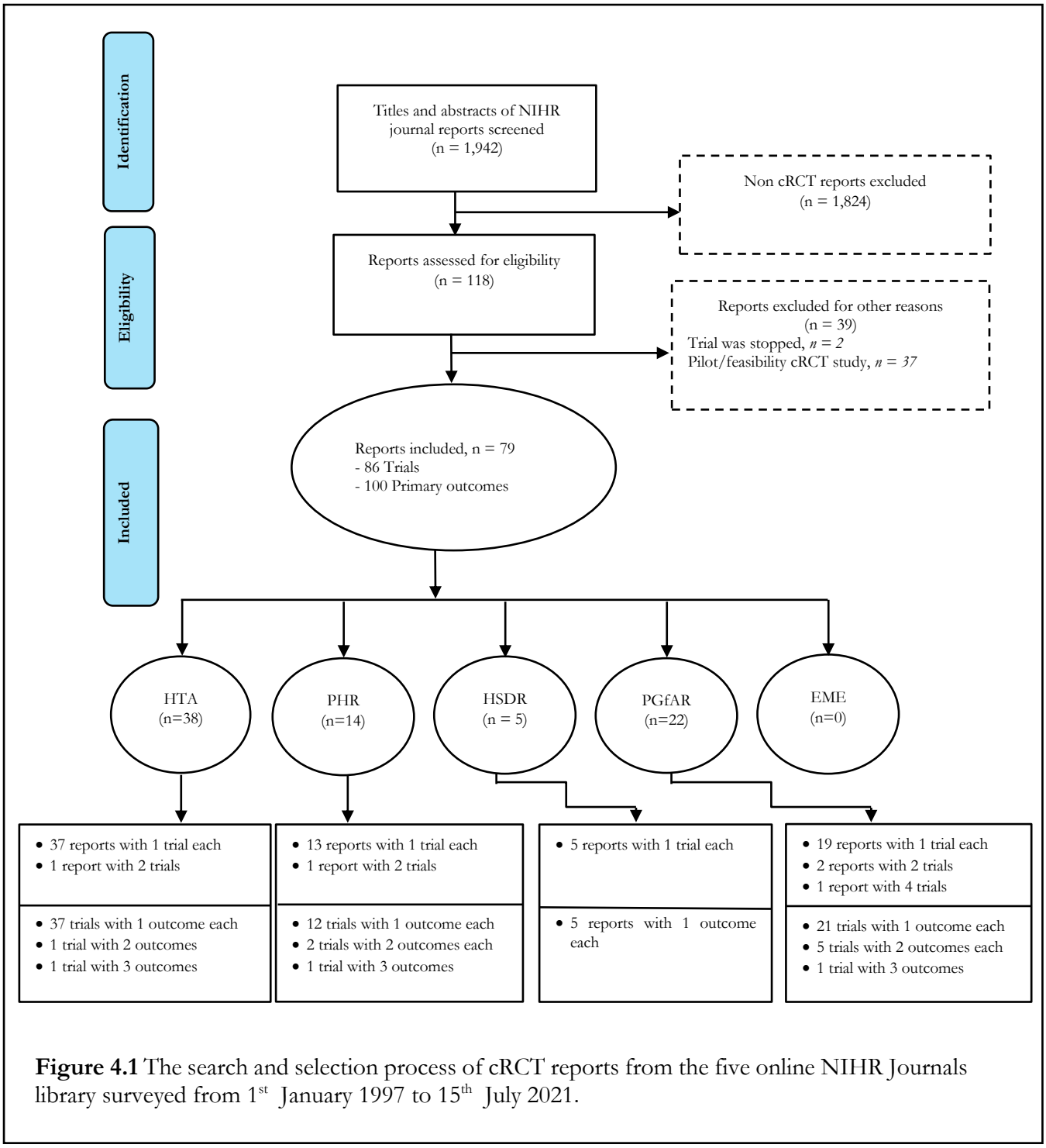


Figure 4.1 The search and selection process of cRCT reports from the five online NIHR Journals library surveyed from 1st January 1997 to 15th July 2021.

Table 4.1 Characteristics of cluster randomised controlled trials published in the NIHR Journals Library, from 1st January 1997 to 15th July 2021

Characteristic	n	%
NIHR journal the cRCT was reported in (N = 79[+])		
HTA	38	48
PHR	14	18
HSDR	5	6
PGAR	22	28
EME	0	0
Trial design (N = 86)		
Parallel	73	85
Factorial	7	8
Crossover	2	2
Others[*]	4	5
Number of trial arms (N = 86)		
2	69	80
3	10	12
2 x 2	4	5
2 x 2 x 2	2	2
2 x 6	1	1
Clinical area (N = 86)		
Cancer/ Oncology	8	9
Mental health (including neurosciences/psychiatry/psychology/dementia)	21	25
Orthopaedics/ Rheumatology/ Musculoskeletal (including back pain)	2	2
Obstetrics and gynaecology	2	2
Primary care	6	7
Cardiovascular	1	1
Gastrointestinal	2	2
Respiratory	1	1
Stroke	4	5
Diabetes	6	7
Dermatology (including ulcers)	1	1
Others[†]	32	37
Setting (N = 86)		
Hospital	4	5
General practice	25	29
Mixed	3	3
Community	3	3
Others[‡]	51	59
Levels of data (N = 86)		
2	85	(99)
3	1	(1)
Trial registration (N = 86)		
ISRCTN	78	91
NCT	2	2
Not reported	6	7

Table 4.1 Characteristics of cluster randomised controlled trials published in the NIHR Journals Library, from 1st January 1997 to 15th July 2021 (cont'd)

Characteristic	n	%
Type of intervention (N = 86)		
Therapy	8	9
Behaviour change technique	4	5
Complex intervention	17	20
Education	12	14
Exercise	3	3
Information and communication technology	3	3
Medical device	2	2
Screening	2	2
Training	17	20
Others[§]	18	21
Type of control (N = 86)		
Usual care	86	100
Are patients blinded? (N = 86)		
Yes	8	9
No	78	91
Any form of a pilot study[a] (N = 86)		
Yes	72	84
No	14	16
Geographical region (N = 86)		
Multiple regions	54	63
Regional	32	37

[a]These are internal pilot studies carried out within the main trials, they are different from the external pilot/feasibility studies mentioned initially in the text.
[+]79, the total number of journal reports included, which reported the results of 86 cRCTs (79 reports included the results of 86 cRCTs)
[*]Partial factorial and step-wedged trials
[†]Insomnia, Paediatrics, Youth bullying, and other aggressive behaviours, Traumatic brain injury, Autism spectrum disorders, Prehospital emergency care, Obesity, Epilepsy, Oral health, End of life care, Children fruit and vegetable intake, Alcohol abuse, Physical activity, Psychosocial work environments, Relationship, and Sexuality Education, Illicit drug use, Smoking prevention, Social and emotional wellbeing of children, Dating and relationship violence, Emergency admission to hospital, Care for older people, Multimorbidity, Abdominal surgery, care of people with long-term conditions, care planning in secondary care mental health services and Psychosis, eating disorder, injuries in under-fives children, patient involvement in safety, psychosis, care planning in secondary care mental health services.
[‡]Care homes, Nursing homes, Clinics, NHS trust, Residential services, Stroke rehab unit, Children centre, Paediatrics diabetes clinic, Schools, Ambulance services, Dental practice, Stroke Services.
[§]Telephone triage, strategies to increase screening, financial incentive, invitation letter, leaflet, behavioural approaches, questionnaire, redesigned care model, health promotion, operational protocol, implementation package, time.

Overall, 95% of the primary analyses used recognised methods to adjust for clustering in their analyses, and 5% did not, they ignored clustering and used standard statistical methods such as the *Chi-square test, standard linear, logistic, and Poisson regressions* (Turner *et al.*, 2000; Morgan *et al.*, 2004; Perez *et al.*, 2016; Gates *et al.*, 2017; Wykes *et al.*, 2018). Continuous outcomes were dichotomised in some studies to enable the use of logistic regression. The trial hypothesis was "superiority" in all the cluster trials except for Heller et al. (2014) which was a non-inferiority trial. **Table 4.2** also shows that most trials recruited and followed up the cohort of participants until the end of the trial; this often leads to missing data due to loss to follow-up (88%, 76/86).

Although 92% of the trials acknowledged the occurrence of missing data, most of them only analysed complete cases (84%). Imputation of missing outcome data was done for just 16% of the trials reviewed (Lamb *et al.*, 2012; Iliffe *et al.*, 2014; Heller *et al.*, 2017; Killaspy *et al.*, 2017; Moniz-Cook *et al.*, 2017; Raine *et al.*, 2017; Snooks *et al.*, 2017; Humphrey *et al.*, 2018; Thompson *et al.*, 2018; Mouncey *et al.*, 2019; Salisbury *et al.*, 2019; Gaughran *et al.*, 2020; Surr *et al.*, 2020).

4.5.3 Planned recruitment targets of participants and clusters

Recruitment characteristics are summarised in **Table 4.3**, 67% (58/86) of cRCTs achieved their planned final individual participant recruitment target, and 87% of the trials achieved $\geq 80\%$ of their final individual participant recruitment target, this indicates successful recruitment to the final targeted sample size for most of the cluster trials. This also applies to the original cluster recruitment target, with 89% of the trials successfully recruiting (and randomising) $\geq 80\%$ of their original targeted number of clusters.

Table 4.2 Characteristics of determinants of (and) statistical methods used for analysing the primary outcomes in cluster trials

Characteristics	n	%
Type of follow-up RCT (N = 86)		
Closed cohort follow-up	76	88
Open cohort follow-up	4	5
Cross-sectional	4	5
Repeated cross-sectional	2	2
The data type of primary outcome (N = 100)		
Continuous	65	65
Binary	28	28
Counts	5	5
Time to event	2	2
Method of adjusting for clustering (N = 100)		
Cluster-level analysis:		
Standard generalized linear model	2	2
Individual-level analysis:		
Generalized linear mixed model	80	80
Robust standard errors	7	7
Generalized estimating equations	6	6
Clustering not accounted for:		
Statistical hypothesis test statistic – chi-square	1	1
Standard generalized linear model	4	4
Specific statistical model (N = 100)		
Linear regression	63	63
Logistic regression	25	25
Relative sensitivity	1	1
Negative binomial regression	2	2
Analysis of proportions	1	1
Cox Proportional Hazards model	2	2
Poisson regression	4	4
Weibull regression model	1	1
Chi-square test	1	1
Random component of GzLMM (N = 80)		
Random intercept	76	95
Shared frailty	1	1
Random intercept and slope (repeated measures)	3	4
Correlation structure in GEE1 (N = 6)		
Exchangeable correlation	5	83
Correlation structure not reported	1	17

N = Total number of trials; n = counts observed in each level of a category; RCT = randomised controlled trial; GzLMM = generalized linear mixed model; GEE1 = generalized estimating equations. Not reported means that the information of interest was not considered and/or provided in the trial.

Table 4.3 Planned participants and clusters recruitment to targets in cluster trials

Characteristics	n	%	Mean (SD)	Median	Range	IQR
Original individual participant target sample size (N = 84[b])						
≤ 300	11	13	10035 (31357)	1250	136 - 250000	550 - 4466
301 – 600	11	13				
601 – 900	13	15				
901 – 1200	3	4				
1201 – 1500	11	13				
1501 – 1800	3	4				
>1800	29	35				
Not reported	3	3				
Final individual participant target sample size (N = 84[b])						
≤ 300	11	13	9372 (30173)	1212	136 - 250000	534 – 4258
301 – 600	11	13				
601 – 900	14	17				
901 – 1200	5	6				
1201 – 1500	11	13				
1501 – 1800	3	4				
>1800	27	32				
Not reported	2	2				
Original individual participant target sample size met (N = 86)						
Yes	57	66				
No, but 80% met	14	16				
No and <80% met	9	11				

[b]Two studies were excluded because the original and final targets were expressed in person-years of observation and not the specific number of participants (Perez *et al.*, 2016; Gulliford *et al.*, 2019). N = Total number of trials; n = counts observed in each level of a variable; SD = standard deviation; IQR = interquartile range. Not reported means that the information of interest was not considered and/or provided in the trial.

Table 4.3 Planned participants and clusters recruitment to targets in cluster trials (Cont'd)

Characteristics	N	%	Mean (SD)	Median	Range	IQR
Final individual participant recruitment target met (N = 86)						
Yes	58	67				
No, but 80% of target met	17	20				
No and <80% of target met	6	7				
Not reported	5	6				
Revised original individual participant target sample size (N = 86)						
Yes, upward	13	15				
Yes, downward	9	10				
Yes, direction not reported	4	4				
No	61	71				
Original cluster recruitment target met (N = 86)						
Yes	68	79				
No, but 80% met	9	11				
No, and <80% met	1	1				
Not reported	8	9				

[b]Two studies were excluded because the original and final targets were expressed in person-years of observation and not the specific number of participants (Perez *et al.*, 2016; Gulliford *et al.*, 2019). N = Total number of trials; n = counts observed in each level of a variable; SD = standard deviation; IQR = interquartile range. Not reported means that the information of interest was not considered and/or provided in the trial.

4.5.4 Cluster and sample size characteristics

In **Table 4.4**, the cluster and sample size characteristics of the included trials are summarised. The design effect if not reported was calculated using the formula; $1 + (\bar{n} - 1) \times ICC$ or $1 + [(CV^2 + 1)\bar{n} - 1] \times ICC$ for equal and unequal/varying cluster sizes respectively, where CV is the coefficient of variation, and $\bar{n}$ is the average cluster size. This is possible if the ICC and average cluster size are reported. The median number of clusters randomised was 44 (IQR, 25 – 74), the minimum was 7 clusters randomised (Heller *et al.*, 2014), and the maximum was 922 clusters randomised, occurred in a trial with households as the clusters (Harris *et al.*, 2018).

A reasonable proportion of the randomised clusters were retained throughout the follow-up period, with a median of 43 clusters (IQR, 25 – 69) included in the analysis which is quite close to the number of clusters randomised. Also, for the number of subjects recruited/randomised the median was 1,184 (IQR, 597 – 3653) while the median number of subjects included in the analyses was 870 (IQR, 441 – 2356).

In the planning stage, 38% (33/86) of the planned ICCs used in the sample size calculations fell in the 0.03 – 0.05 range. The median planned ICC in the sample size calculation was 0.05 (IQR, 0.026 – 0.07). The observed ICCs from the analysed primary outcomes in the trials have a median value of approximately 0.02 (IQR, 0.001 – 0.060) with most of the reported ICCs occurring in the -0.02 to 0.02 range (**Table 4.4**). After excluding two trials that were analysed at the cluster level, we found that 42% (42/100) of the observed ICCs from the primary analyses of the primary outcomes were not reported. Thirty-one percent of the observed ICCs were not reported before the publication of the CONSORT 2010 statement extension for cluster randomised trials compared to 44% after its publication (**Table 4.4**). One study carried out pair-matched randomisation using a minimization technique; however, they analysed their primary outcomes at the individual level (Thompson *et al.*, 2018). Pair matching of clusters reduces the population heterogeneity at the cluster level which could result in a negligible ICC from the analysed primary outcome and improve the statistical efficiency of the trial (Ivers *et al.*, 2011; Campbell *et al.*, 2012). Not reporting the observed ICC for the analysed primary outcomes contradicts the CONSORT 2010 reporting guidelines for cluster trials, which recommends that authors should report *"a coefficient of intracluster correlation (ICC or k) for each primary outcome"*. The minimum observed ICC value appears to be an outlier (-0.02) and was found in Heller et al. (2014).

There is a somewhat trend of not reporting the observed ICC for the analysed primary outcomes, which is more evident after the CONSORT 2010 statement extension for cluster randomised trials was published than before (**Figure 4.2**). Before the publication of the CONSORT 2010 statement extension for cluster randomised trials, and in the years that trials were carried out, 2003, 2005, and 2011 recorded nonreporting of the observed ICCs for the analysed primary outcomes (20%, 100%, and 50%, respectively). However, after the publication of the CONSORT reporting guidelines, almost every year aside from 2013, some of the observed ICCs for the analysed primary outcomes were not reported, ranging from 28% to 90% (**Figure 4.2**).

Similarly, **Table 4.4** shows that a higher proportion did not report their observed ICCs from analysed primary outcomes after the publication of the CONSORT 2010 reporting guidelines compared to the proportion that did not before its publication (44% vs 31%).

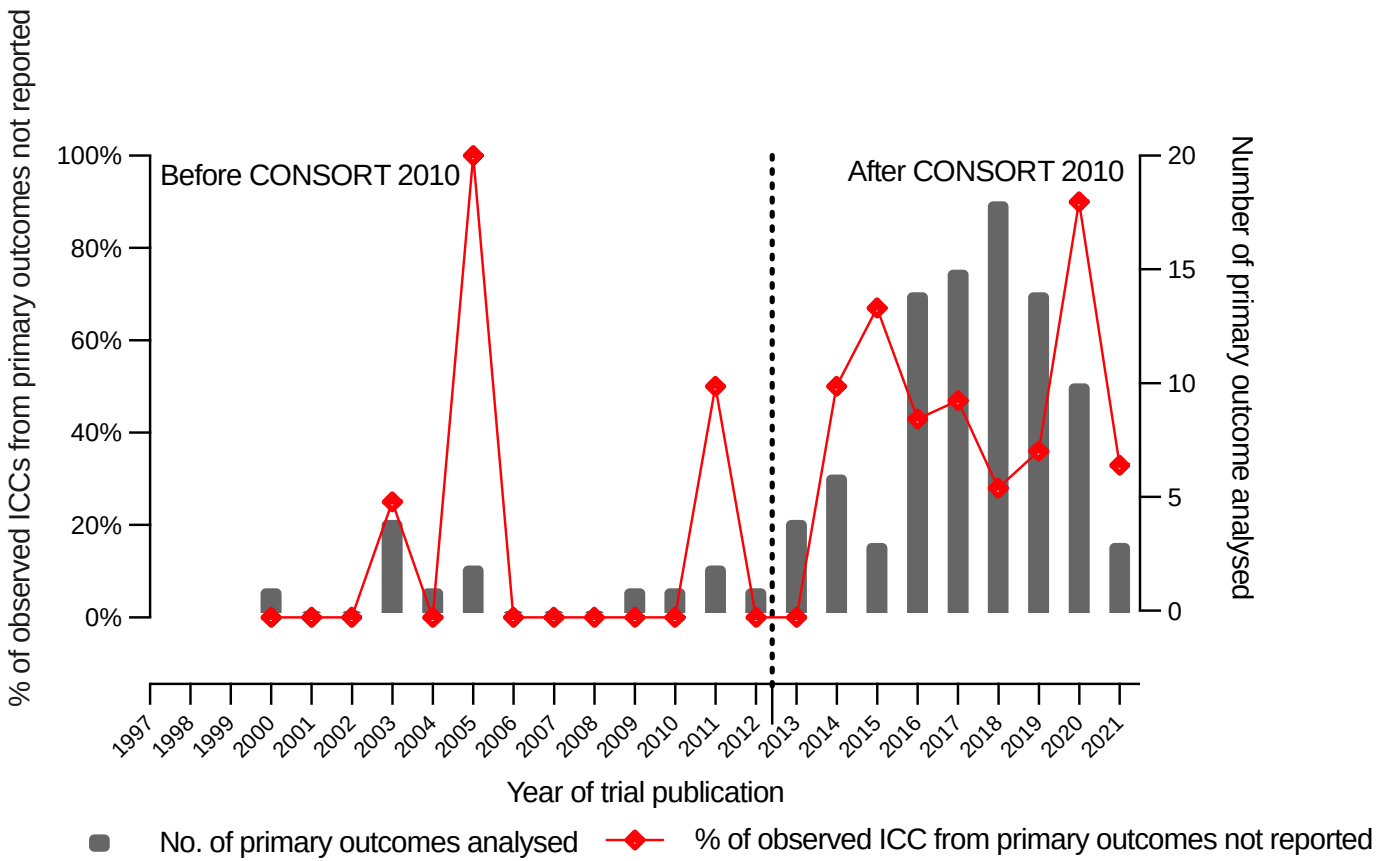


Figure 4.2 Plot comparing the trend of not reporting the observed ICCs of analysed primary outcomes in cRCTs before and after CONSORT 2010 statement with the first published cRCT in NIHR Journals library recorded in 2000.

Table 4.4 Cluster and sample size characteristics of the trials included in the review

Characteristics	n	%	Mean (SD)	Median	Range	IQR	90th	95th
No. of clusters randomised (N = 86)								
4 – 10	2	2	77 (121)	44	7 – 922	25 – 74	123	274
11 – 20	11	13						
21 – 50	40	47						
51 – 100	21	24						
101 – 200	5	6						
>200	7	8						
No. of clusters analysed (N = 86)								
0 – 10	2	2	76 (118)	43	7 – 864	25 – 69	121	274
11 – 20	12	14						
21 – 50	40	47						
51 – 100	21	24						
101 – 200	4	5						
>200	7	8						
No. of subjects recruited (N = 84[b])								
≤ 300	7	8	15348 (48315)	1184	141 – 265434	597 – 3653	15766	146538
301 – 600	14	17						
601 – 900	11	13						
901 – 1200	9	11						
1201 – 1500	9	11						
1501 – 1800	3	4						
>1800	29	34						
Not reported	2	2						
No. of subjects analysed (N = 84[b])								
≤ 300	15	18	14367 (48419)	870	42 – 264325	441 – 2356	14831	150540
301 – 600	15	18						
601 – 900	13	15						
901 – 1200	5	6						

[b]Two trials were excluded because the analysed subjects were measured in person-years. N = Total number of trials; n = counts observed in each level of a category; SD = standard deviation; 90th =90th percentile; 95th =95th percentile. Not reported means that the information of interest was not considered and/or provided in the trial.

Table 4.4 Cluster and sample size characteristics of the trials included in the review (cont'd)

Characteristics	n	%	Mean (SD)	Median	Range	IQR	90th	95th
No. of subjects analysed (N = 84[b])								
1201 – 1500	5	6						
1501 – 1800	2	2						
>1800	25	30						
Not reported	4	5						
Planned ICC for sample size (N = 86)								
0.00 – 0.02	18	21	0.065 (0.082)	0.05	0.0002 – 0.5	0.0258 – 0.0700	0.12	0.15
>0.02 - 0.05	33	38						
>0.05 - 0.08	9	11						
>0.08 - 0.11	8	9						
>0.11 - 0.14	2	2						
>0.14	6	7						
Not reported	10	12						
Planned design effect (N = 86)								
0.00 – 2.99	47	55	4.5 (8.90)	1.96	1.03 – 70.5	1.384 – 4.600	7.00	9.30
3.00 – 5.99	12	14						
6.00 – 8.99	10	12						
9.00 – 11.99	1	1						
≥12	3	3						
Not reported	13	15						
Observed ICC of analysed primary outcome (N = 100)								
-0.02 to 0.02	35	35	0.06 (0.12)	0.02	-0.02 to 0.63	0.0010 – 0.0600	0.22	0.23
>0.02 - 0.07	9	9						
>0.07 - 0.12	3	3						
>0.12 - 0.17	6	6						
>0.17 - 0.22	2	2						
>0.22	3	3						
Not reported	42	42						

[b]Two trials were excluded because the analysed subjects were measured in person-years. N = Total number of trials and/or primary outcomes; n = counts observed in each level of a category; SD = standard deviation; 90th =90th percentile; 95th =95th percentile. Not reported means that the information of interest was not considered and/or provided in the trial.

Table 4.4 Cluster and sample size characteristics of the trials included in the review (cont'd)

Characteristics	n	%	Mean (SD)	Median	Range	IQR	90th	95th
Average number of subjects analysed (N = 84[b])			NA	21	1 – 20917	10 – 71	246	1845
≤ 10	20							
11 – 50	36							
51 – 90	6							
91 – 130	4							
131 – 170	2							
171 – 210	1							
211 – 250	1							
>250	14							

[b]Two trials were excluded because the analysed subjects were measured in person-years. N = Total number of trials and/or primary outcomes; n = counts observed in each level of a category; SD = standard deviation; NA = Not Applicable; 90th = 90th percentile; 95th = 95th percentile. Not reported means that the information of interest was not considered and/or provided in the trial.

Table 4.5 Comparison of the non-adherence in the reporting of observed ICC for each primary outcome before and after the CONSORT 2010 statement

	Year of publication		
	Before 1997 – 2012	**After 2013 – 2021**	**All 1997-2021**
Number of trials	11	75	86
Number of primary outcomes	13	87	100
Number of primary outcomes with the observed ICC not reported (%)	4 (31)	38 (44)	42(42)

4.6 Discussion

This review was carried out to investigate the statistical methods used for analysing cluster randomised controlled trials in practice, to this end I surveyed publicly funded cluster trials funded by the NHIR. Most of the trials used appropriate/recognised statistical methods to adjust for clustering in the main analyses of the primary outcomes from the trials (95%, 95/100). Few (5%) ignored clustering and used standard statistical methods that assumed independence among outcomes from participants in a cluster to analyse the outcome data. This approach is not encouraged as it could lead to smaller SE and consequently, an increased value of the test statistic, smaller P-value, narrower CI, and possibly increase the Type I error rate compared with the statistical methods that allow for clustering. If this happens to be the case, misleading conclusions and decisions will be made, this could have detrimental impact on public health.

The GzLMM was the most popular choice in adjusting for clustering and is more popular than GEE1 (80% vs. 6%). For the GzLMMs applied to outcome data with two levels (i.e., trial participants nested within clusters), the cluster unit is usually incorporated as a random intercept to account for clustering. If the primary outcome was measured more than once or the level of clustering is more than two levels, then GzLMM with random intercept and random slope is ideal. Four trials that used the GEE1 method assumed an exchangeable working correlation structure in the primary analysis (Morrell *et al.*, 2009; Dormandy *et al.*, 2010; Kitchener *et al.*, 2016; Harrington *et al.*, 2019) while one trial did not report the correlation structure that was assumed (Heller *et al.*, 2014).

Fiero et al. (2016) conducted a systematic review that focused more on the handling of missing data than on the statistical methods used for analysing cluster trials and found similar results to ours. They found that most of the trials analysed their primary outcome using GzLMM, and the

cluster unit was modelled as the random intercept to account for clustering. Also, they found that all 14(100%) of the trials that used GEE1 to account for clustering assumed an exchangeable correlation structure, which is similar to the findings of this current review (5/5, 100%; one study did not report their correlation structure (Heller *et al.*, 2014)). Overall, they found that a lower proportion of 79% (68/88) of the trials accounted for clustering compared to our review which observed a higher proportion of 95% (95/100).

It is worth noting that while the use of appropriate statistical methods was high, none of the trials considered the recent potentially improved statistical methods developed in other study designs where clustered data do arise, such as GEE2, QIF, ALR, and TMLE. Three of these recent methods (GEE2, QIF and ALR) are improvements over the standard GEE1 method for estimating the regression coefficients in the model (Turner, 2017). I theorised that this gap in knowledge regarding the available and appropriate methods for analysing outcome data from cRCTs (Chapter 3), and the actual ones used in practice (Chapter 4) could be because there is sparse knowledge on how to implement them. Additionally, the comparative advantages of the newer methods over the classical methods (GzLMM and GEE1) have not been comprehensively investigated in the context of cRCTs. This led to the selection of two classical methods (GzLMM and GEE1) and two emerging methods (GEE2 and QIF) to be further investigated in Chapter 7 (using real-world data) and Chapter 9 (using simulated data).

The two classical methods selected to be further evaluated were the two most studied methods in the literature of cRCTs (**Table 3.1**), and also the two most applied methods in practice (**Table 4.2**). The statistical properties of these methods have been comprehensively evaluated to a reasonable extent, and they have been around for a longer period, hence I tagged them the "classical methods". The statistical properties of these two methods would be evaluated against the two selected newer/emerging methods (i.e., QIF and GEE2). QIF was selected based on being the most studied newer method (**Table 3.1**), indicating a plausible high rate of interest on it by researchers. Research indicating that QIF could be a potential alternative to GEE1 were mostly conducted in the context of longitudinal studies (Qu, Lindsay and Bing, 2000; Qu and Song, 2004; Odueyungbo *et al.*, 2008; Song *et al.*, 2009). For example, Qu, Lindsay and Bing (2000) showed that QIF is more efficient than GEE1 (i.e., less variable) when the working correlation is misspecified, especially when the correlation among outcomes across clusters is substantial.

This result was replicated using a real-world data from a longitudinal study (Odueyungbo *et al.*, 2008). The literature supporting this claim in the context of cRCTs is sparse (Qu and Song, 2004; Yu, Li and Turner, 2020), while some studies showed that the opposite is the case when the number of clusters is small (Westgate, 2012; Westgate and Braun, 2012). Lastly, to the best of my knowledge, no study has evaluated the performance of QIF (a PAM) against any CSM (like GzLMM). These plausible reasons prompted the need to comprehensively evaluate QIF against GEE1 and GzLMM in the context of cRCTs. Similarly, GEE2 appears to be a promising alternative to GEE1 especially when the association among outcomes is also of interest, it's of complex structure and depends on the cluster size (Yan and Fine, 2004; Crespi, Wong and Mishra, 2009). In situations like this, GEE2 is likely to improve inference (Crespi, 2016). GEE2, just like QIF has not been comprehensively evaluated against GEE1 or GzLMM in the context of cRCTs. The above plausible reasons justify the choice of the four methods (GzLMM, GEE1, QIF and GEE2) for further comparative evaluation of their statistical properties.

The results of our study revealed that the number of clusters randomised in a cRCT could be as large as 922 in a study where the clusters were households (Harris *et al.*, 2018) and as few as 7 clusters (Heller *et al.*, 2014). This result is different from the findings of (Arnup *et al.*, 2016) who focused on cluster randomised crossover trials, one reason for choosing a crossover design is if the number of the prospective clusters is small. In their study, the lowest number of clusters randomised was 7 while 25% of the number of clusters randomised was below 25.

In practice, "treatment as usual" is mostly used when assessing the effect of non-pharmacological interventions (86/86, 100%). As revealed in our results, most times it is impractical to conduct studies where the participants are blinded to the experimental arms they are allocated to. However, to some extent, masking is achieved by blinding either the person randomising the subjects, the assessor, and/or the statistician that will analyse the data. To carry out a robust cluster trial it is preferable to conduct an internal pilot/feasibility study (84%, 72/86) to assess the viability of the items/phases of the trial, such as the data collection tools, the understanding (and safety) and acceptance of the intervention by the participants, and the ability to recruit to target before proceeding with the main trial. Recruiting participants (for each cluster) into a trial seems not to be a problem in cRCTs, particularly when compared to RCTs (**Table 4.6**). In 87% of the cluster trials, researchers were able to recruit ≥ 80% of their final planned participants' recruitment targets.

Table 4.6 Comparing the ability to recruit to target the number of participants between cRCTs and RCTs, using results of previous studies that reviewed RCTs

Review	McDonald et al. (2006)	Sully, Julious and Nicholl (2013)	Walters et al. (2017)	**This study**
Recruitment period	**1994 – 2002**	**2002 – 2008**	**2004 - 2016**	**1997 - 2021**
Number of trials in the study	**N = 122 RCTs**	**N = 73 RCTs**	**N = 151 RCTs**	**N = 86 cRCTs**
Recruited 100% of original target	38 of 122 (31%)	40 of 73 (55%)	61 of 151 (40%)	**57 of 86 (66%)**
Original target was revised	42 of 122 (34%)	14 of 73 (19%)	52 of 151 (34%)	21[c] of 86 (24%)
Original target revised upward	6 of 42 (14%)	5 of 14 (36%)	11 of 52 (21%)	**12 of 21 (57%)**
Original target revised downward	36 of 42 (86%)	9 of 14 (64%)	41 of 52 (79%)	9 of 21 (43%)
Recruited 80% of original target	67 of 122 (55%)	57 of 73 (78%)	95 of 151 (63%)	**71 of 86 (83%)**
Recruited 100% of revised target	19 of 42 (45%)	10 of 14 (71%)	28 of 52 (54%)	**16 of 21 (76%)**
Recruited 80% of revised target	34 of 42 (80%)	13 of 14 (93%)	48 of 52 (92%)	**21 of 21 (100%)**
Extended their recruitment	65 of 122 (54%)	33 of 73 (45%)	49 of 151(32%)	**11 of 86 (13%)**

Source: Adapted (and modified) from Walters et al. (2017). [c]Was supposed to be 25 trials but 2 trials did not report their original target that was revised, and another two trials did not report their final revised target and the number of participants recruited respectively, they were excluded since comparison cannot be done.

This result also applies to the number of clusters recruited, where 76% of the trials were able to recruit ≥ 80% of their planned cluster recruitment target (**Table 4.6**). The comparison of the ability of cRCTs and RCTs to recruit to their target the number of participants, in terms of recruiting 100% of the original participant target, cluster trials seem more successful than RCTs (66% vs 55%). This is confirmed by the fact that in cluster trials the originally planned sample sizes are rarely revised (24%), and tend to be revised upward (57%, 12/21) rather than downward (43%). When compared to RCTs, the number (and percentage) of upward revisions of planned participant recruitment was higher in cluster trials (57% vs. 36%). Even with the most upward revisions, cluster trial recruitment periods are rarely extended to meet recruitment targets compared to RCTs (13%, 11/86 vs. 54%, 65/122).

This review results also showed that in practice the completely randomised parallel-group cluster trial design is the most used design involving two treatment arms in its simplest form. This cluster design is easy to set up, implement, and analyse (Campbell and Walters, 2014a). The results indicated that all the trials reviewed, except one were superiority trials involving contrasting treatment arms. For the sample size calculation, our results indicated that the median assumed ICC value, used in the calculation, was 0.05 while the mean was 0.065. However, we observed that the ICC assumed in the sample size calculation could be as low as 0.0002 (**Table 4.4**). Results also indicated that a disappointing trend of not reporting the observed ICC for each primary outcome is still happening. About 4 out of 10 of the observed ICCs from the analysed primary outcomes in cRCTs are not being reported. The implication of not reporting the ICC cannot be overemphasized, the ICC is an important item in designing/planning future cluster trials as it is needed for sample size calculation. It is reasonable to make it available for researchers planning to undertake a similar study. The importance of reporting the ICC was reemphasized by the development of a framework specifying how and what should be reported in association with the ICC to facilitate understanding and the planning of future cluster trials (Campbell, Grimshaw and Elbourne, 2004).

Surprisingly, this has occurred more in recent times despite the availability and publicity of the CONSORT 2010 statement extension for cluster randomised trials. Although the year 2005 had the highest percentage of this disappointing practice (100%), with only two analysed primary outcomes. It is worth noting that this was also after the publication of the CONSORT 2004 statement extension for cluster randomised trials (Campbell, Elbourne and Altman, 2004). Ivers

et al. (2011) assessed the impact of the CONSORT 2004 statement extension for cluster randomised trials on the quality of reporting and study methodology, and one of the criteria compared was the "reporting of an estimated intracluster correlation". They found only 18% of the 300 manuscripts reported an ICC estimate, and 22% vs. 14% before and after the CONSORT 2004 statement respectively. This result indicated a decline in the practice of reporting observed ICC which is like this current review.

This review found a 13% relative increase (change in non-adherence before and after CONSORT 2010 statement) in non-adherence to the CONSORT reporting guidelines with regards to reporting the observed ICC for each primary outcome analysed, using CONSORT 2010 statement as the basis for comparison (**Table 4.5**). One of the potentials explanations to this declined in the reporting quality/adherence to CONSORT extension to cluster trials regarding the observed ICC could be because the developers of the guidelines have not sensitised the research community well enough. One way to do this is ensure that subsequent update is broadly published in several medical journals (Ivers *et al.*, 2011). It has been recommended that developers of CONSORT extension to cluster trials should work beyond the publication of the guidelines and provide structures to assist editors and researchers in the proper use of it (Ivers *et al.*, 2011). Some of the problems of not doing this are situations where authors are using the wrong guidelines to report cRCTs (which means not reporting some items peculiar to cRCTs).

Furthermore, another potential cause of the increase in not reporting the observed ICC could be that the editors are not strict (or due to oversight) in implementing/monitoring their policy regarding the use of appropriate reporting guidelines. It is also likely that the editors and peer reviewers are not strict about the reporting of certain aspect of the trials like the observed ICC. In other words, the authors could have implemented the item but failed to clearly report it, and this is not being picked up by the editorial team. For example, it could be that the editorial team are not too keen about the reporting of the observed ICC, may be because it has no direct impact on the results obtained, or conclusions made in the study. However, reporting the observed ICC is of great important for planning and designing future studies (Campbell *et al.*, 2012). Lastly, it has been found that studies conducted under clinical settings (like primary cares) are more likely to adhere to the reporting guidelines compared those conducted in non-clinical settings (like communities) (Ivers *et al.*, 2011).

CONSORT statements extensions for cluster randomised trials are published to facilitate improved quality reporting of cluster trials. If used properly, they are supposed to help in the understanding, assessing, and replicating of cluster trials by all stakeholders of clinical trials. Hence all authors intending to write up the report for their cluster trial(s) should make good use of the updated CONSORT 2010 statement extension for cluster randomised trials.

The observations made in this review are that, in practice, there are important issues in cRCTs that are still being ignored or handled inadequately, or not reported. Firstly, missing data is not adequately handled most of the time. The majority (79/86, 92%) of the studies reviewed acknowledged the existence of missing data, which is obvious due to the inevitable loss of follow-up in a closed cohort follow-up study, however, the majority still went ahead to analyse only available observations (84%). Although most researchers check the robustness of their findings when missing data is not technically handled by conducting sensitivity analysis, however, if missing data is properly handled it might improve the inferences made in a study.

Secondly, there appears to be a slow uptake of newer statistical methods developed in other study designs where clustered data can arise, such as the QIF, TMLE, and ALR which are acclaimed to be better in certain situations than the popular methods used currently for analysing clusters trials. It would be ideal if these methods were publicised by methodologists of cluster trials so that researchers can use them when necessary to make optimal inferences (Turner, 2017).

4.7 Limitations

Firstly, the articles included in the review was sourced from only one database, the NIHR Journal Library which could be a recipe for publication bias. To circumvent this issue, I ensured that all the articles that was identified and meet the inclusion criteria were included instead of a random sample. Also, it is worth noting that the reports published in the NIHR Journals Library mostly published independently as result articles in other journals and stored in other databases. Hence, reports included in this review represent a collection of articles from several journals/databases that are independent of the NIHR Journals Library.

Secondly, the articles included reported studies conducted only in the UK. In other words, none of the articles reported studies conducted outside of the UK. Consequently, the findings of this current study are not generalisable to cluster randomised controlled trials conducted outside the

UK. Additionally, the results are only generalisable to studies conducted in the UK with similar characteristics.

4.8 Summary

In this chapter, the statistical methods used in practice for analysing outcome data from cRCTs were successfully identified. While some of the methods are appropriate and adequately account for clustering in the outcome data from cRCTs, some ignored clustering.

The most commonly used analytical methods in practice were GzLMM and GzLM – typifying the conditional and marginal regression models, respectively. The specific models were intercept random effects, regression with robust SEs, and GEE1. The basic summary statistics of some parameters used for planning and conducting the simulation study in Chapters 8 and 9 were obtained from the results of this practice review, such as cluster sizes, observed ICCs, and number of clusters.

Furthermore, these two most used analytical approaches (GzLMM and GEE1), and some other classical approaches have been described in Chapter 2 (Section 2.8.3). These approaches included cluster-level analysis, and RMRSE. Chapter 3 was a step further in identifying more appropriate methods in the literature that have been used or can be used to analyse outcome data from cRCTs. In Chapter 3, twenty-seven unique and appropriate methods for analysing cRCTs were identified, which included GzLMM (with MLE/REML), GEE1, Bayesian methods (with MCMC), t-test, QIF, PQL, permutation test, Wilcoxon rank-sum test, adjusted chi-square test, and QIF as the most common methods used. Similarly, some of the parameters needed for the simulation study conducted in Chapters 8 and 9 were obtained from this methodological scoping review. For example, information on the common values of the following parameters were obtained; types of outcomes, cRCT designs, analytical methods of missing data, and the median (and IQR) number of runs often used in simulation studies in the cRCT literature.

From the results of these two reviews conducted in Chapters 3 and 4, it became apparent that gaps exist in the literature of the design and analysis of cRCTs. For example, there is a substantial gap between the methods that are available in the literature for analysing cRCTs, and the methods that are actually used in practice. In Chapter 3, I identified some newer analytical methods (GEE2 and QIF) for analysing cRCTs, however, none of these methods was used in practice (Chapter 4). I

theorised that there are several reasons behind this. Firstly, it could be due to the lack of practical guidance on how to implement these newer methods in the context of cRCTs. Secondly, may be due to the lack of comprehensive studies showing the comparative advantages of these newer methods over the classical methods, in the context of cRCTs. Lastly, it could be because of the lack of readily available commands in common statistical computing platforms for the implementation of these newer methods.

To tackle these identified gaps, I selected four analytical methods for further investigation. Two classical methods (GzLMM and GEE1) and two newer methods (QIF and GEE2) were selected. First, GzLMM (with MLE/REML) and GEE1 were selected primarily because they are the most studied methods in the literature of statistical methods for analysing cRCTs (Chapter 3), indicating the amount of interest on the methods by researchers. Second, the results from the practice review of Chapter 4 showed that GzLMM and GEE1 are the two most used methods in practice (**Table 4.1**), and it is a common knowledge that their properties are well established in the literature to warrant calling them classical methods in the case (Murray *et al.*, 2004; Turner, 2017). On the other hand, QIF and GEE2 were selected as the newer/emerging methods because they are more recently proposed, and their properties are yet to be comprehensively established in the literature. Furthermore, they are acclaimed to be promising alternatives to GEE1 (Turner, 2017), however, their advantages over GEE1 are yet to be comprehensively evaluated in the context of cRCTs (Westgate, 2012). Hence, the purpose of the further studies is to compare the properties of the two classical methods, GzLMM and GEE1, that have been established as the common choice both in the literature and practice (Chapter 3 and 4, respectively) against the two acclaimed promising newer alternatives; QIF and GEE2 to establish superiority in different situations. The rest of the chapters of this thesis will be dedicated to achieving this.

Chapter 6 presents further technical descriptions of the four selected classical methods mentioned above. Chapter 7 presents the results of applying these selected methods to real-world outcome data from four cRCTs. One of the limitations of using real-world data is that it only presents a specific scenario according to how the trial was designed. A simulation study was planned and presented as a working protocol in Chapter 8. The simulation study was conducted, and the results were reported in Chapter 9.

Chapter 5

Research questions, aim, and objectives

5.1 Introduction

The literature of cRCTs was successfully reviewed in previous Chapters 3 and 4. The review in Chapter 3 focussed on identifying appropriate methods available in the literature that can be used or have been used to analyse outcome data from cRCTs. While Chapter 4 was on the methods that have been used in practice to analyse outcome data from cRCTs. These two reviews helped identify the gaps in knowledge – concerning the comparison of methods for analysing outcome data from cRCTs. Chapter 2 described the classical methods for analysing outcome data from cRCTs with examples of their applications. Namely: cluster-level summary measures; RMRSE; CSMs – GzLMMs with model parameters estimated by MLE/REML; PAMs – GzLMs with model parameters estimated by GEE1.

The review of statistical methods in Chapter 3 identified more statistical methods, in addition to the "classical methods" mentioned above, that can be used to analyse outcome data from cRCTs. Among these methods identified QIF is the most studied newer method and is acclaimed to be a promising alternative to GEE1 (Qu, Lindsay and Bing, 2000). To the best of my knowledge, in the context of cRCT, this claim has not been proven comprehensively. Only four comparative studies exist for GEE1 and QIF, three of these studies carried out their simulations using moderate to high ICC values ranging from 0.01 to 0.7 and was based on the efficiency of the methods (Westgate, 2012; Westgate and Braun, 2012, 2013). The most recent of the four was not comprehensive (Yu, Li and Turner, 2020). Yu, Li and Turner (2020) assumed only one level for some of their simulation parameters (i.e., fixed), for example, cluster sizes were assumed to be 25, and the number of clusters was 100 (50 per treatment arm). In addition, none of the studies compared QIF (population average model) to any CSM like GzLMM with parameters estimated by MLE/REML.

Chapter 4 reviewed the cRCT literature for methods used in practice to analyse outcome data from cRCTs, and found that the most common methods used in the 86 publicly funded cRCTs (with 100 primary outcomes) in the UK were mainly the "classical" statistical methods, with the most popular being the GzLMMs (80%, 80/100), followed by regression with robust SEs 7%, GzLMs with model parameters estimated by GEE1 (6%), aggregate summary statistics methods 2%, and 5% did not account for clustering. None of these cRCTs used newer methods like QIF. Thus, there appears to be a gap in knowledge regarding the comprehensive comparative evaluation of GEE1 and QIF, and a need to compare QIF and GzLMMs - one of the most used methods to analyse cRCTs.

5.2 Chapter aim

This chapter aims to present the research questions, aim, and objectives of this thesis based on the gaps in knowledge identified previously in Chapters 3 and 4.

5.3 Research questions

The research questions of interest are based on the main purpose of this research, which is to comparatively evaluate the statistical methods for analysing outcome data from cRCTs. The questions are:

- What are the appropriate, and available methods in the literature used for analysing outcome data from cRCTs?
- What statistical methods are used in practice for analysing outcome data from cRCTs?
- What criteria could be used to decide the appropriateness of the identified methods?
- How well do the selected methods perform, when compared?

The SISAQOL review (Coens *et al.*, 2020) for the analysis of outcome quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials recommended a set of essential and highly desirable criteria for defining appropriate statistical methods for patient-reported outcome (PRO) analysis. These criteria have been modified/adapted for cluster RCTs. An "appropriate" statistical method for analysing cluster RCTs would be one with:

1) The ability to handle clustered data (correlated outcome data within a cluster).

2) The ability to do a comparative test (statistical significance test).

3) The ability to produce interpretable treatment effects (estimation of treatment effect and associated uncertainty (can the model produce an estimate with an associated confidence interval?)).

4) The ability to adjust for covariates, including baseline individual and cluster level covariates.

5) The ability to handle missing data with the least restrictions.

5.4 Research aim

The overall aim of this research is to identify, describe, and compare the selected different statistical methods that can be used to analyse outcome data from cRCTs and make recommendations about the most appropriate method of analysis in different scenarios. This aim is further broken down into the following specific objectives below.

5.5 Research objectives

The objectives to be achieved in this thesis are:

- To identify what statistical methods are described in the literature for analysing cRCTs.
- To identify what statistical methods are used in practice, by reviewing reports of cRCTs published in the NIHR journals library, to identify what statistical analyses were used.
- To identify the key statistical properties or criteria for an appropriate statistical method for analysing cRCTs.
- To apply the criteria created to the statistical methods found in the NIHR Journals Library and methodological reviews to produce a list of methods to evaluate.
- To briefly describe the selected methods for analysing cRCTs.
- To apply the selected statistical methods to various real-world outcome data from cRCTs and evaluate their performance.
- To use simulated continuous outcomes to compare the selected statistical methods for analysing cRCT by evaluating their performances based on their long-run statistical properties.

- To make recommendations for the most appropriate method of analysis based on the findings of this research.

5.6 Summary

Chapter 5 recalls most of what was explained in Chapter 1 – research questions, aim, and objectives. The remaining chapters of this research will be dedicated to achieving this aim and objectives. While Chapters 3 and 4 addressed the objectives of identifying appropriate methods that can be used for analysing cRCTs available in the literature and the methods that have been used in practice, respectively, Chapter 6 addresses the objective of describing the selected methods to be evaluated.

The next Chapter 7 presents the results of applying the four selected methods (GzLMM, GEE1, GEE2, and QIF) to outcome data from four cRCTs. But, to reach a valid conclusion on the superiority of the methods a simulation study was conducted in Chapters 8 and 9. A simulation study provides several scenarios that could arise in a cRCT designed to mimic clinical trials in primary care and community trials. Chapter 10 discussed the findings of this research and how they compared to the findings of other studies, the strengths and contributions of the research, the research limitations, implications and recommendations, conclusions, and issues to be explored in the future.

Chapter 6

Further descriptions of statistical methods

6.1 Introduction

Classical approaches for analysing outcome data from cRCTs are well-studied in the literature (Zeger and Liang, 1986; McCulloch, 1997; Ukoumunne, 2002; Heo and Leon, 2005; Walters, Morrell and Slade, 2011; Campbell and Walters, 2014a), some of which were described in Chapter 2. These approaches included CLA, GzLMM, RMRSE, and GEE1. However, the literature on emerging/newer methods is sparse, and this could affect their routine application (Turner, 2017). This chapter provides an in-depth technical description of the four selected statistical methods: GzLMM, GEE1, GEE2, and QIF. These include the underlying theories of the methods, brief derivations, and the statistical software packages used to implement the methods. The first two methods are considered as classical methods and the last two, as emerging methods.

Two reviews were conducted which led to the selection of these methods in Chapters 3 and 4. Chapter 3 summarised the results from a review to identify the available and appropriate methods in the literature that could be or have been used to analyse outcome data from a cRCT. This was a methodological scoping review of the cRCT literature, conducted by applying systematic search techniques to the online bibliography databases of MEDLINE, EMBASE, PsycINFO, CINAHL, and SCOPUS. Briefly, under CSMs the MLE was the most studied method (18/112) while for PAM it was the GEE1 (23/112), complete results are presented in **Table 3.1**. Among the emerging methods, QIF and GEE2 were identified as the most promising alternatives to GEE1. It has been shown that QIF and GEE2 outperformed GEE1 in certain situations; in the context of longitudinal study design and where the association structure is of interest, respectively (Prentice and Zhao, 1991; Qu, Lindsay and Bing, 2000). But not much is known about these two emerging methods (QIF and GEE2) in the context of cRCT design.

For example, QIF was found to perform better than GEE1 when their correlation structures were misspecified in the context of longitudinal design (Qu, Lindsay and Bing, 2000), and when there were outliers in the outcome data at both the cluster and individual participant levels of a cRCT

(Qu, Lindsay and Bing, 2000; Qu and Song, 2004). Similarly, GEE2 has been recommended for studies where the dependence among subjects in a cluster is of great interest (Prentice and Zhao, 1991) or clustering is highly heterogeneous across clusters (Crespi, Wong and Mishra, 2009).

Chapter 4 presented the results of reviewing the NIHR Journals Library to find what methods are used in practice to analyse outcome data from cRCTs. The two selected classical methods named above happen to be the two most used methods. However, there was no recorded use of the two emerging/alternative methods, QIF and GEE2. For these acclaimed alternatives to be recommended for routine application, they must be well understood and shown to be relatively better than the already existing classical methods. Hence, these four selected statistical methods – GzLMM, GEE1, GEE2, and QIF are subjected to the comparative evaluation of their statistical properties to ascertain their superiority in different situations. Some of the statistical properties of a good method are unbiasedness, consistency, efficiency, and sufficiency.

6.2 Chapter aim

The primary aim of this chapter is to describe the fundamental theories and assumptions underlying the four selected methods. Additionally, the statistical packages used in implementing the methods are provided.

6.3 Generalized linear model

The GzLM is an extension of the general linear model (GLM) and some of its special cases include simple LM, ANCOVA, binomial, Probit, and Poisson models. The unification of these traditional regression models under GzLM is made possible through link functions, $\eta(.)$. A link function establishes a linear relationship between covariates and the outcome variable. Examples of some common link functions are the identity, log-linear, Probit, logit, and complementary log-log. The suitable link functions for each data type (and their distribution) are summarised in **Table 6.1** below. The theories of LM and GzLM provide the theoretical basis to establish the other methods to be discussed.

Table 6.1 Some common link functions for different types of outcome data that follows the exponential family distribution

Link function	Type of outcome (Support)	Distribution of outcome	Notation	Canonical link
Identity	Continuous (-∞,+∞)	Normal	$No(\mu, \sigma^2)$	$\eta = \mu$
Logit	Binary $(0\ and\ 1)$	Binomial	$Bi(n,p)$	$\eta = log\left\{{}^{p}/_{(1-p)}\right\}$
Probit	Binary $(0\ and\ 1)$	Binomial	$Bi(n,p)$	$\eta = \Phi^{-1}(\mu)$
Log	Counts (0,1, …,∞)	Poisson	$Po(\mu)$	$\eta = log\ (\mu)$
Complementary log-log	Binary $(0\ and\ 1)$	Binomial	$Bi(n,p)$	$\eta = log\ \{-log\ (1-p)\}$
Reciprocal	Positive continuous (0, +∞)	Gamma	$Ga(a,b)$	$\eta = \mu^{-1}$

One of the advantages of the GzLM is that a homogenous residual variance assumption is not necessary, GzLM uses a variance function characterised by the mean to account for unequal variances. However, GzLM assumes that the residuals are independent, which is often not the case in cRCTs. Hence, estimates from a GzLM in cRCTs could lead to poor SE estimates, since $corr(\varepsilon_{ij}, \varepsilon_{ij'}) \neq 0$ but rather $corr(\varepsilon_{ij}, \varepsilon_{ij'}) = \rho(x_{ij}, x_{ij'}; \boldsymbol{P})\ \forall\ j \neq j'$, where $\boldsymbol{P}$ is the true correlation matrix to be approximated by a "working" correlation matrix, $\boldsymbol{R}(\alpha)$. In simple words, the outcomes are correlated within the clusters.

These assumptions of the GzLM and LM make it unrealistic to directly use them to analyse outcome data from cRCTs. Recall that model equation (2.12), $\eta\big(E(\boldsymbol{Y}_i)\big) = \boldsymbol{X}_{pij}^{T}\boldsymbol{\beta}_p$, is the mathematical representation of the GzLM. Obtaining the parameter estimates of (2.12) requires the use of an appropriate estimator, such as the MLE which is discussed further in Section 6.5. As asserted initially, the parameter estimates from GzLM could be less efficient when the correlation is substantial, therefore approaches that explicitly account for correlation should be considered. Specifically, all the methods selected to be investigated further met the required criteria of an appropriate method for analysing a cRCT (See, Section 5.3). In the next sections, the methods are further described.

6.4 Generalized estimating equations (GEEs)

To circumvent the limitation of the GzLM – not accounting for correlation among pairs of outcomes in a cluster, the marginal generalized linear model (mGzLM) was proposed in the context of longitudinal analysis, and extension to the cRCT design is straightforward (Song, 2007). The methods under this class formulate quasi-likelihoods that are analogues to score equations of

the correlated responses of the i^{th} cluster responses, $\boldsymbol{Y}_i$, hence are considered semi-parametric estimators. The parameter estimates of the fixed effects from GzLM and mGzLM are equivalent in interpretation (Hubbard *et al.*, 2010), the difference lies in the parameter estimate of their SEs for the fixed effects component. In mGLMs, the correlation between pairs of outcomes is considered when estimating the SEs. Liang and Zeger (1986) proposed the popularised class of generalized estimating equations, $U_i(\boldsymbol{\beta})$, of equation (2.15) that accounts for the correlation using the working covariance matrix $\boldsymbol{V}_i$ of (2.16) which is characterised by the working correlation matrix $\boldsymbol{R}_i(\alpha)$. Two commonly used $\boldsymbol{R}_i(\alpha)$s in cRCTs, in matrix format are given as:

Independence working correlation matrix/structure,

$$\boldsymbol{R}_i(\alpha) = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} \tag{6.1}$$

where $\boldsymbol{R}_i(\alpha)$ is an $n_i \times n_i$ identity matrix, the zeros on the off-diagonal elements indicate that any random pairs of outcomes within the i^{th} cluster have zero (no) correlation, hence outcomes are independent. When (6.1) is plugged into (2.6) of the covariance function of $\boldsymbol{Y}_i$ it gives rise to the independent working covariance structure, thus GzLM and mGzLM have the same SE estimates. The second $\boldsymbol{R}_i(\alpha)$ is the most assumed and ideal working correlation in cRCT design, which is

Exchangeable working correlation matrix/structure,

$$\boldsymbol{R}_i(\alpha) = \begin{bmatrix} 1 & \alpha & \dots & \alpha & \alpha \\ \alpha & 1 & \dots & \alpha & \alpha \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha & \alpha & \dots & 1 & \alpha \\ \alpha & \alpha & \dots & \alpha & 1 \end{bmatrix} \tag{6.2}$$

Equation (6.2) characterises the exchangeable/compound symmetric working covariance structure representing a situation where the correlation of any random pairs of outcomes taken from a cluster is equal to α. The exchangeable working correlation structure directly accounts for the correlation in the outcome data, and α quantifies the correlation. Going forward only three methods of mGzLMs built on the framework of the GEEs are further discussed; GEEs typify the mGzLM. These methods are GEE1(Liang and Zeger, 1986), GEE2 (Yan and Fine, 2004), and QIF (Qu, Lindsay and Bing, 2000).

6.4.1 First-order generalized estimating equations (GEE1)

This class of GEE focuses on modelling the mean parameters and treats the correlation parameters as a nuisance – the correlation is described but not modelled. It is commonly referred to as the first-order GEE and denoted as GEE1. Note that GEE1 draws strength from the linear exponential family distribution (Ziegler, 2011). To illustrate this, let $\boldsymbol{Y}_i$ be as defined previously and let us drop the index i for simplicity so $\boldsymbol{Y}_i$ is the same as $\boldsymbol{Y}$, if the marginal probability density function (or probability mass function for a discrete distribution) of $\boldsymbol{Y}$ can be expressed as belonging to the rich class of the linear exponential family distribution given as

$$f(\boldsymbol{Y};\boldsymbol{\theta},\boldsymbol{\phi}) = exp[(\boldsymbol{\theta})'\boldsymbol{Y} + b(\boldsymbol{Y},\boldsymbol{\phi}) - c(\boldsymbol{\theta},\boldsymbol{\phi})] \tag{6.3}$$

where $\boldsymbol{\theta}$ is a vector of natural parameters to be estimated, $\boldsymbol{\phi}$ is the fixed nuisance parameters assumed to be a positive definite matrix, $b(.)$ and $c(.)$ are known measurable functions. The first and second moments (the mean and variance) of $\boldsymbol{Y}$ can be solved by taking the partial derivative of the log of the moment generating function (MGF), $c(\boldsymbol{\theta},\boldsymbol{\phi})$, parametrised in the mean as (Ziegler, 2011).

$$E(\boldsymbol{Y}) = \hat{\boldsymbol{\mu}} = \frac{\partial \log\big(c(\boldsymbol{\mu},\boldsymbol{\phi})\big)}{\partial \boldsymbol{\mu}}$$
$$var(\boldsymbol{Y}) = \hat{\sigma}^2 = \frac{\partial^2 \log\big(c(\boldsymbol{\mu},\boldsymbol{\phi})\big)}{\partial \boldsymbol{\mu}\partial \boldsymbol{\mu}^{\boldsymbol{T}}} \tag{6.4}$$

(6.5)

To demonstrate how to obtain the mean parameter estimates while ignoring the nuisance parameter, for a continuous outcome, let y be a univariate outcome that follows the Normal distribution $N(\mu,\sigma^2)$. The first step is to express the distribution of y as a distribution belonging to the exponential family given as

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}}$$

$$f(y;\mu,\sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2}$$

$$= \frac{y\mu - 1/2\,\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$$

$$f(y;\mu,\sigma^2) = \frac{y\mu}{\sigma^2} - \frac{1}{2}\left\{\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right\} - \frac{\mu^2}{2\sigma^2} \tag{6.6}$$

with $\theta = \mu$, $b(y,\phi) = -\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right\}$, $c(\theta,\phi) = -\frac{\mu^2}{2\sigma^2}$ and $\phi = \sigma^2$.

Hence, the first and second moments, the mean and variance respectively, of y can be obtained by partially differentiating the log of the MGF, $c(\theta,\phi)$ as done in (6.4), we have

$$E(y) = \hat{\mu} = \frac{\partial c(\theta,\phi)}{\partial\theta} = \frac{\mu^2}{2} = \mu$$

$$Var(y) = \frac{\partial^2 c(\theta,\phi)}{\partial\theta\partial\theta^T} = \phi\frac{\partial\mu}{\partial\theta} = \sigma^2, \text{ since } \phi = \sigma^2 \tag{6.7}$$

(6.8)

For discrete/categorical outcomes, let y follow a binomial distribution $Bi(n,p)$, for fixed n and $p \in [0,1]$

$$f(y;n,\mathrm{p}) = \binom{n}{y}p^y(1-p)^{n-y} = \exp\{\log(p)\times y + log\binom{n}{y} + (n-y)\log(1-p)\}$$
$$= exp\left\{\, ylog\left(\frac{p}{1-p}\right) + log\binom{n}{y} + n\,log(1-p)\right\} \tag{6.9}$$

Let $\theta = \log\left(\frac{p}{1-p}\right) \equiv e^\theta = \left(\frac{p}{1-p}\right) \Rightarrow p = {e^\theta}/{1+e^\theta}$ and $1-p = 1/1+e^\theta$

Hence, expressing (6.9) in its natural parameter form becomes

$$f(y;\theta,\phi) = exp\left\{\theta y + log\binom{n}{y} - n\,log\left(1+e^\theta\right)\right\} \tag{6.10}$$

where $\phi = 1$, $b(y,\phi) = log\binom{n}{y}$, and $c(\theta,\phi) = -n\log\left(1+e^\theta\right)$

The estimating algorithm of GEE1 is akin to (6.7) where the mean parameter is of interest while the nuisance parameter is not. Recall that the GEE1 formulation is $U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \Delta_i$ and depends on two parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, but again only $\boldsymbol{\beta}$ are estimated. The nuisance parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ are unknown but estimated from the data. Under some mild regularity conditions $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\phi}}$ are $N^{\frac{1}{2}}$ consistent, $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ with Gaussian residuals $N^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ that have a zero mean and covariance matrix $\boldsymbol{\xi}$, given as

$$\boldsymbol{\xi} = \lim_{\boldsymbol{N \to \infty}} N \left(\sum_{\boldsymbol{i=1}}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1} \left\{ \sum_{\boldsymbol{i=1}}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} cov(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right\} \left(\sum_{\boldsymbol{i=1}}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1} \tag{6.11}$$

where $\boldsymbol{D}_i = \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)$, hence the estimator $\hat{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$ is specified by estimating $cov(\boldsymbol{Y}_i)$ from the data using $(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\,(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T$, and plugging in all the other estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\alpha}}$, and $\widehat{\boldsymbol{\phi}}$. Although, $\hat{\boldsymbol{\xi}}$ is robust to the misspecification of the correlation structure, in some circumstances, it could suffer some loss in efficiency. Since most of the components of (6.11) are obtained from the data, it is apparent that high variability in the outcome caused by a small sample size, and high dependence among outcomes are two circumstances that would affect the estimates from GEE1 severely, especially when the assumed correlation structure is misspecified, and the true correlation is substantial (Liang and Zeger, 1986; Leyrat *et al.*, 2018; Thompson *et al.*, 2022).

In summary, GEE1 produces consistent estimates of the intervention effect provided that the mean model is correct, and data are missing completely at random regardless of whether the correlation structure is misspecified but could suffer loss in efficiency if the working correlation structure does not closely approximate the true correlation structure, especially if the correlation is large, and the cluster size varies (Liang and Zeger, 1986). Also, a loss in efficiency can occur when there are few clusters, the robust SE does not provide full protection over incorrect working correlation structure in this situation (Leyrat *et al.*, 2018). These disadvantages of the GEE1 are the reason for the proposal of GEE2 and QIF that are discussed next.

6.4.2 Second-order generalized estimating equations (GEE2)

In Section 6.4.1, $\boldsymbol{\phi}$ was treated as a nuisance parameter. But, in GEE2 they are estimated simultaneously with the mean parameters to improve the asymptotic efficiency of the parameter estimates. The class of regression models under GEE2 attempts to leverage the major drawback of GEE1 – possible loss in efficiency when the correlation structure is misspecified, especially when the correlation among outcomes is substantial (Prentice and Zhao, 1991; Yan and Fine, 2004).

The second-order pseudo maximum likelihood (PML2) is based on the quadratic exponential family of distributions. GEE2 is based on the PML2 inference, so it is ideal to explain the quadratic exponential family distribution. If $\boldsymbol{y^T}$ is a $n \times 1$ transposed vector of the response values and $\boldsymbol{s} = (y_1^2, y_1y_2, \dots, y_1y_n, y_2^2, y_2y_3, \dots, y_n^2)^T$, let $\boldsymbol{\mu}$ be the mean vector and $\boldsymbol{V(\beta, \alpha)}$ be the associated $n \times n$ covariance matrix of the $\boldsymbol{y^T}$. If the joint density of the $\boldsymbol{y^T}$ can be expressed as

$$f(\boldsymbol{y};\, \boldsymbol{\theta}, \boldsymbol{\phi}) = exp[\boldsymbol{\theta^T y} - c(\boldsymbol{\theta}, \boldsymbol{\phi}) + b(\boldsymbol{y}) + \boldsymbol{\phi^T s}] \tag{6.12}$$

equation (6.12) is said to be a n-dimensional quadratic exponential family distribution in the natural parameter form and can be parameterised in the mean and covariance matrix given as

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{V}) = exp[a(\boldsymbol{\mu}, \boldsymbol{V})^T \boldsymbol{y} + c(\boldsymbol{\mu}, \boldsymbol{V}) + b(\boldsymbol{y}) + \gamma(\boldsymbol{\mu}, \boldsymbol{V})^T \boldsymbol{s}] \tag{6.13}$$

where *a, b, c, and* γ are some known functions with γ having a $n(n+1)/2$ dimension. From (6.13), the "quadratic" in "quadratic exponential family" distribution is established by representing the function, $\gamma(\boldsymbol{\mu}, \boldsymbol{V})^T \boldsymbol{s}$, with $\boldsymbol{y^T D y}$ for a symmetric matrix $\boldsymbol{D(\mu, V)}$ since

$$\boldsymbol{y}^T \boldsymbol{D} \boldsymbol{y} = \sum_{\tilde{n}=1}^{n} \sum_{\tilde{n}'=1}^{n} y_{\tilde{n}} y_{\tilde{n}'} [\boldsymbol{D}]_{\tilde{n}\tilde{n}'} = \sum_{\tilde{n}}^{n} y_{\tilde{n}}^2 [\boldsymbol{D}]_{\tilde{n}\tilde{n}} + 2 \sum_{\tilde{n}'>\tilde{n}} y_{\tilde{n}} y_{\tilde{n}'} [\boldsymbol{D}]_{\tilde{n}\tilde{n}'} = \gamma^T \boldsymbol{s} \tag{6.14}$$

while $\gamma = ([\boldsymbol{D}]_{11}, 2[\boldsymbol{D}]_{12}, \dots, 2[\boldsymbol{D}]_{1n}, [\boldsymbol{D}]_{22}, 2[\boldsymbol{D}]_{23}, \dots, [\boldsymbol{D}]_{nn})^T$. Hence, $\boldsymbol{D}$ is the coefficient matrix of a quadratic term $\boldsymbol{s}$(in $\boldsymbol{y}$) characterise by $\boldsymbol{\mu}$ and $\boldsymbol{V}$. So, let's assume that the natural

parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are stacked in one column and the variables $\boldsymbol{y}$ and $\boldsymbol{s}$ in another column, the first two moments of an outcome data are obtained simultaneously as

$$E((\boldsymbol{y}^T, \boldsymbol{s}^T)^T) = \frac{\partial \gamma(\boldsymbol{\theta}, \boldsymbol{\phi})}{\partial(\boldsymbol{\theta}^T, \boldsymbol{\phi}^T)^T} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{V} \end{pmatrix}$$

$$Var((\boldsymbol{y}^T, \boldsymbol{s}^T)^T) = \begin{pmatrix} Var(\boldsymbol{y}) & cov(\boldsymbol{y}, \boldsymbol{s}) \\ cov(\boldsymbol{s}, \boldsymbol{y}) & Var(\boldsymbol{s}) \end{pmatrix} = \frac{\partial(\boldsymbol{\mu}^T, \mathbf{V}^T)}{\partial(\boldsymbol{\theta}^T, \boldsymbol{\phi}^T)^T} = \begin{pmatrix} \frac{\partial \boldsymbol{\mu}}{\partial(\boldsymbol{\theta})^T} & \frac{\partial \boldsymbol{\mu}}{\partial(\boldsymbol{\theta})^T} \\ \frac{\partial \boldsymbol{\mu}}{\partial(\boldsymbol{\phi})^T} & \frac{\partial \boldsymbol{V}}{\partial(\boldsymbol{\phi})^T} \end{pmatrix}$$

(6.15)

For example, if y has a univariate Normal distribution as given in (6.6), and ϕ is substituted by $V = \sigma^2$, then y belongs to the quadratic exponential family distribution of (6.12) given as

$$a(\mu, \mathrm{V}) = \frac{\mu}{\sigma^2}, c(\mu, \mathrm{V}) = -\frac{1}{2}\left\{\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2)\right\}, b(y) = 0, and\ \gamma(\mu, \mathrm{V}) = -\frac{1}{2}\frac{1}{\sigma^2}$$

(6.16)

The GEE2 borrows strength from the quadratic exponential family described above and does estimate the mean and nuisance parameters (as called in GEE1) simultaneously using (6.15). Several estimators based on the GEE2 framework have been proposed in the literature of clustered data (Prentice, 1988; Prentice and Zhao, 1991; Hall and Severini, 1998; Ziegler *et al.*, 2000; Yan and Fine, 2004), however, Yan and Fine (2004) proposed a GEE2 variant that uses separate link functions to model the mean, scale, and correlation parameters and the corresponding sets of estimating equations are solved simultaneously. This is known as the three-estimating equations (3EE) GEE2 which is used in this research.

To establish the model specification, let $\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}\ and\ \boldsymbol{X}_{3i}$ be the $n_i \times p, n_i \times r$ and $\frac{n(n+1)}{2} \times q$ design matrices for the mean, the scale, and the correlation parameters of the vector of outcomes $\boldsymbol{Y}_i$, respectively. The specific link function for the mean, the scale, and correlation parameters to $\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}\ and\ \boldsymbol{X}_{3i}$ is given as

$$\eta_1(\boldsymbol{\mu}_i) = \boldsymbol{X}_{1i}\boldsymbol{\beta}$$

$$\eta_2(\boldsymbol{\phi}_i) = \boldsymbol{X}_{2i}\boldsymbol{\varphi}$$
$$\eta_3(\boldsymbol{\rho}_i) = \boldsymbol{X}_{3i}\boldsymbol{\alpha}$$

(6.17)

where $\boldsymbol{\mu}_i$ is a $n_i \times 1$ mean vector specified by $\boldsymbol{\beta}$, $\boldsymbol{\phi}_i$ is a $n_i \times 1$ scale vector specified by $\boldsymbol{\varphi}$ and $\boldsymbol{\rho}_i$ is a $\frac{n_i(n_i+1)}{2} \times 1$ pairwise correlation vector specified by $\boldsymbol{\alpha}$. The unified corresponding set of estimating equations to be solved simultaneously is

$$U_i(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}) = \begin{array}{l} \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{V}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) = 0 \\ \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\phi}_i}{\partial \boldsymbol{\varphi}}\right)^T \boldsymbol{V}_{2i}^{-1}(\boldsymbol{Z}_i - \boldsymbol{\phi}_i(\boldsymbol{\varphi})) = 0 \\ \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\rho}_i}{\partial \boldsymbol{\alpha}}\right)^T \boldsymbol{V}_{3i}^{-1}\left(\boldsymbol{S}_i - \boldsymbol{\rho}_i(\boldsymbol{\alpha})\right) = 0 \end{array}$$

(6.18)

where $\boldsymbol{Z}_i$ is the $n_i \times 1$ vector of the scales, $\boldsymbol{S}_i$ is the $\frac{n_i(n_i+1)}{2} \times 1$ vector of the pairwise correlations, $\boldsymbol{V}_{1i}$ and $\boldsymbol{V}_{2i}$ are the working covariance matrices of $\boldsymbol{Z}_i$ and $\boldsymbol{S}_i$ respectively. The GEE2 estimator of (6.18) requires the specification of the first four central moments of the outcome vector (mean, variance, skewness, and kurtosis). Yan and Fine (2004) proposed a way to avoid the problem of convergence caused by specifying higher-order moments and it is implemented in their *geese* function (Yan, 2002) in the R package: *geepack* (Højsgaard, Halekoh and Yan, 2005). To be specific, the third and fourth moments can be specified as functions of the first and second moments, thereby avoiding the direct estimation of higher-order moments (Prentice and Zhao, 1991). The GEE2 estimator consistently estimates the mean parameters $\boldsymbol{\beta}$ regardless of whether the scale and correlation structures are wrong; the estimates for scale parameters $\boldsymbol{\varphi}$ are consistent regardless of whether the working correlation is misspecified, but provided that the mean and scale structures are correct.

The major merit of the 3EE variant of the GEE2 estimator is that it allows for separate covariates in the mean, scale, and correlation structures to be adjusted for, this is important when

investigating heterogeneity across clusters or the treatment arms such as modelling multiple forms of clustering. Where each cluster or treatment arm presents a different degree of correlation α_i among subjects, possibly due to cluster sizes and covariates imbalance. Taking this into account may improve efficiency, instead of assuming a constant correlation value across clusters or treatment arms (Crespi, Wong and Mishra, 2009). The solutions of (6.18) are obtained iteratively by alternating between a modified Fisher scoring algorithm and the moment estimation method.

6.4.3 Quadratic inference function (QIF)

Similarly, to GEE2, QIF was proposed to circumvent a major disadvantage with GEE1, which is the loss in efficiency due to the misspecification of the correlation structure. But compared to GEE2, QIF does not require the specification of the third and fourth moments (as they impose additional constraints). The QIF estimator avoids the direct use of the working correlation matrix in its algorithm. Instead, it uses a linear combination of basis matrices and some constants to replace the inverse of the working correlation matrix. Hence, QIF is more robust to misspecification of the working correlation matrix compared to GEE1, providing better protection against incorrect working correlation structure. With this, QIF produces more efficient parameter estimates compared to GEE1 (Qu, Lindsay and Bing, 2000). However, if the working correlation matrix is not misspecified, the efficiency of the parameter estimates from GEE1 and QIF are equivalent (Qu, Lindsay and Bing, 2000; Song *et al.*, 2009).

Let $\boldsymbol{Y}_i$, $\boldsymbol{X}_i$, $\boldsymbol{\mu}_i$, and $\boldsymbol{V}_i$ be the same as defined previously. In the QIF equation, the inverse of $\boldsymbol{R}$ which is a component of the covariance matrix $\boldsymbol{V}_i$ is approximated using a linear combination of a set of several basis matrices: $\boldsymbol{R}_h^{-1} \approx k_h \boldsymbol{I}_h + \cdots + k_m \boldsymbol{M}_m, (h = 0, \ldots, m)$; $\boldsymbol{I}_h$ is the identity matrix, $\boldsymbol{M}_m$ are known basis matrices and $\boldsymbol{k}_m$ are unknown constants that need to be estimated. For exchangeable and autoregressive working covariance matrices, $h = 1\ and\ 2$ would suffice, respectively (Qu, Lindsay and Bing, 2000; Westgate and Braun, 2013).

A basis for a vector space V occurs when any element $v \in V$ can be written as a finite linear combination of the elements of a set B of vectors $\{b_1, \ldots, b_n\}$, that is $v = a_1 b_1 + \cdots + a_n b_n$, where a's are referred to as the components or coordinates of v w.r.t B. Using this new information, we can rewrite the estimating equations of the GEE1 as extended score equations given as

$$\bar{g}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\boldsymbol{\beta}) \approx \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{G}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) \\ \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{G}_i^{-1/2}\boldsymbol{M}_1\boldsymbol{G}_i^{-1/2}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) \\ \vdots \\ \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{G}_i^{-1/2}\boldsymbol{M}_m\boldsymbol{G}_i^{-1/2}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) \end{pmatrix}$$

(6.19)

where $g_i(\boldsymbol{\beta})$ is the score vector of each cluster, the constants $\boldsymbol{k}_m$ are considered a nuisance and are not included. The QIF estimator uses the generalized method of moments (GMM) (Hansen, 2010) to optimally combine the multiple estimating equations in (6.19). The estimate $\widehat{\boldsymbol{\beta}}$ is obtained by minimising the weighted length of $\bar{\boldsymbol{g}}_N$ using GMM and given as

$$\widehat{\boldsymbol{\beta}} = arg\ min_{\boldsymbol{\beta}}\ \bar{\boldsymbol{g}}_N^T \boldsymbol{\Sigma}_N^{-1}\ \bar{\boldsymbol{g}}_N$$

(6.20)

where $arg\ min_{\boldsymbol{\beta}}$ is the argument of the minimum of $\boldsymbol{\beta}$ that minimises $\bar{\boldsymbol{g}}_N^T \boldsymbol{\Sigma}_N^{-1}\ \bar{\boldsymbol{g}}_N$. As expected, the true covariance matrix $\boldsymbol{\Sigma}_N$ is replaced by the estimated covariance matrix $\boldsymbol{C}_N$ in (6.20), with its inverse $\boldsymbol{C}_N^{-1}$ representing a weighting function. Thus, the QIF estimator is defined as

$$Q_N(\boldsymbol{\beta}) = N\bar{\boldsymbol{g}}_N^T(\boldsymbol{\beta})\mathbf{C}_N^{-1}\ \bar{\boldsymbol{g}}_N(\boldsymbol{\beta})$$

(6.21)

where $\boldsymbol{C}_N = (1/N^2)\sum_i^N g_i(\boldsymbol{\beta})g_i^T(\boldsymbol{\beta})$, $\boldsymbol{C}_N^{-1}$ is the main reason behind QIF's efficiency advantage because it weights the information each i^{th} cluster contributes to the estimating equation, clusters with large variation are given less weight than the ones with small variation. The estimates $\widehat{\boldsymbol{\beta}}$ are obtained iteratively using the Newton – Raphson algorithm (Qu, Lindsay and Bing, 2000), and given as

$$\widehat{\boldsymbol{\beta}} = arg\ min_{\boldsymbol{\beta}}\ \bar{\boldsymbol{g}}_N^T \mathbf{C}_N^{-1}\ \bar{\boldsymbol{g}}_N$$

(6.22)

The QIF models of Section 7.4 were fitted using the SAS 9.4 macro: *qif*. There is also an R version called the *qif* package, but it requires that no cluster should have a single observation.

6.5 Linear mixed model (LMM)

The LMM was established based on the theories of simple LM (conditioned on fixed covariates) and Normally distributed random effects models. The LMM is a natural choice for analysing continuous outcome data from cRCTs where the random effects vary between clusters, and their shared common factors induce within-cluster dependence between pairs of outcomes after conditioning on fixed covariates $\boldsymbol{X}_{ij}$. A simple case of an LMM is given in (2.11) and elaborated in (2.12) with the response having a conditional expectation given as

$$\left(E\left(\boldsymbol{Y}_{ij}|\boldsymbol{X}_{ij},\boldsymbol{\tau}_i\right)\right) = \left(\boldsymbol{\mu}_{ij}\right) = \beta_0+\beta_1 x_{1i} + \cdots + \beta_p x_{pij} + \tau_i \tag{6.23}$$

Thus, LMM is conditioned on fixed and random effects components. The random effects are assumed to follow a Normal distribution $\tau_i \sim (0, \sigma_b^2)$. The random effects from the clusters τ_i and the residual error for each participant ε_{ij} are assumed independent across clusters, and conditional on the covariates in the model.

6.6 Generalized linear mixed model (GzLMM)

GzLMM is formulated based on a combined theory from GzLM and LMM. As an extension to both GzLM and LMM, it improves the flexibility of both models combined. Firstly, the theory from GzLM allows GzLMM to relax the assumption that outcome data follows a Normal distribution, while that of the LMM allows the dependence among outcomes from subjects within a cluster. With these combined theories, the GzLMM can model outcome data that are not necessarily Normally distributed and exhibit some form of dependence, where the mean response is a linear combination of the unknown parameters using a link function of the mean (McCulloch and Searle, 2000). Note that the GzLMM also borrows strength from the exponential family distributions.

The scope of this thesis only covers a simple case of GzLMM – where only the fixed covariates effects and random cluster effects are examined. This is known as the random intercept model, and the cluster unit is included in the model as a random intercept term. Recall model equation (2.9) that represented the GzLMM in Chapter 2 given as

$$\eta\left(E\left(\boldsymbol{Y}_{ij}|\boldsymbol{X}_{ij},\boldsymbol{\tau}_i\right)\right)=\eta\left(\boldsymbol{\mu}_{ij}\right)=\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pij}+\tau_i$$

where $\boldsymbol{\mu}_{ij}$ is the vector of the expected values of the continuous outcome data, $\boldsymbol{Y}_{ij}$, $\eta(.)$ is a canonical link function that linearises the model in the parameters, $\boldsymbol{\beta}=(\beta_0,\beta_1\ldots\beta_p)$ is the fixed but unknown parameters of the model that would be estimated, τ_i is the random effects term for the cluster unit which is often assumed to follow a Normal distribution with a mean 0 and a variance σ_b^2. Thus, the properties of the cluster random effects could be summarised as

$$E(\tau_i)=0\ \forall\ i \tag{6.24}$$

$$var(\tau_i)=E[(\tau_i-E[\tau_i])^2]=E[\tau_i^2]=\sigma_b^2 \tag{6.25}$$

and

$$cov(\tau_i,\tau_i')=0 \text{ for } \tau_i\neq\tau_i'$$

It is worth noting that a linear PAM with an exchangeable covariance matrix is the same as a linear mixed model with a random intercept. The coefficients of models with full likelihoods such as GzLM, LM, GzLMM, and LMM can be estimated using several estimators. One of the most common estimators is the MLE described below. The joint likelihood of the outcome data y_{ij} is given as

$$\begin{aligned}l(\boldsymbol{\beta},\tau_i\,;y_{ij})&=\prod_{i=1}^{N}f(y_{ij})\\&=\prod_{i=1}^{N}\int\prod_{j=1}^{n_i}f(\tau_i,\boldsymbol{\beta})g(\tau_i;\sigma_b^2)\partial\tau_i\end{aligned} \tag{6.26}$$

where $l(.)$ is the full likelihood function for y_{ij}, $f(.)$ is the probability function for y_{ij}, $g(.)$ is the Normal probability density function for the unobservable random term τ_i and $\boldsymbol{\beta}$. The estimates using MLE are obtained by taking the first derivatives of the log of $l(.)$ for each parameter, while the second derivatives produce the SEs. However, equation (6.26) does not have an analytical closed-form solution due to the high dimensional integral involved. However, for LMM a closed-form solution is possible because the marginal density of y_{ij} is a Normal density. In general, integrating (6.26) becomes difficult as the dimension of the integrals increases due to the complex nature of the function $f(\tau_i,\boldsymbol{\beta})g(\tau_i;\sigma_b^2)$.

The most effective way of circumventing this problem is to use approximation algorithms to obtain an approximate solution. One such algorithm is the numerical integration of definite integrals, quadrature methods typify the numerical integrations. The adaptive Gauss-Hermite quadrature (AGHQ) is a special case of quadrature methods and applies the quadrature rules efficiently to obtain optimal approximation. The AGHQ performs well for a wide range of cluster sizes and ICCs (Fitzmaurice *et al.*, 2008).

Let $z \in \mathbb{R}$ be a random variable with density function $f(z)$, and $\psi(z)$ is a known function that would be integrated alongside, hence the Gauss-Hermite quadrature rule is

$$\int_{-\infty}^{\infty} f(z)\psi(z)\, dz \approx \sum_{j=1}^{d} f(z_j)w_j \tag{6.27}$$

where d is the total quadrature points, w_j are quadrature weights, and z_j are the evaluation points. Equation (6.27) is an alternative approach of using a weighted sum to approximate the solution of integrating a function with high integrals. The right-hand side equals the left-hand sides when $\psi(z)$ is continuous and $f(z)$ is a polynomial with degree $\leq 2d-1$. The values of w_j and z_j for few levels of d can be obtained from (McCulloch and Searle, 2000) or from statistical/mathematical software given as

$$z_j = ith \text{ zero of } H_n(z) \tag{6.28}$$

$$w_j = \frac{2^{n-1}n!\sqrt{\pi}}{n^2\left[H_{n-1}(z_j)\right]^2} \tag{6.29}$$

where $H_n(z)$ is the Hermite polynomial of degree n. Theoretically, in GzLMM, the parameters of the fixed effects and random effects components are obtained simultaneously. However, technically, the MLE estimates the fixed effects component first (ignoring the random effects component), then plugs the estimates into the algorithm to estimate the random effects component. This process is repeated until optimal estimates are obtained. But, ignoring the random effects component in the first step causes the MLE parameter estimates of SEs to be biased since it means ignoring the variations in the fixed effects component, which could be substantial when the sample size is small (McNeish and Stapleton, 2016; Leyrat *et al.*, 2018; Thompson *et al.*, 2022).

Also, the MLE does not adjust for the degrees of freedom (DoF) lost in estimating the parameters of the fixed effects component (McNeish and Stapleton, 2016). Hence, the MLE is likely to produce SEs that are too small, resulting in smaller P-values, and inflated Type I error rates, especially when there are few clusters. An alternative likelihood-based estimation method is the REML which can be utilised to circumvent these problems. For large sample sizes, these problems are not noticeable, and the estimates from MLE and REML are approximately the same. But for cRCTs with small number of clusters, the problems are more pronounced (Leyrat *et al.*, 2018; Thompson *et al.*, 2022).

REML first transforms the outcome data to remove the fixed effects, before estimating the random effects component. Then, it applies a generalized least squares estimator to obtain the estimates of the fixed effects component within its algorithm. Put differently, REML obtains the estimates of the fixed effects and random effects components separately, starting with the random effects component (McNeish and Stapleton, 2016). To appropriately adjust for the loss in the DoF we applied the Satterthwaite correction on the DoF when analysing outcome data from cRCTs with a small sample, consequently obtaining the correct P-values and CIs (Leyrat *et al.*, 2018).

The following are the reasons why the MLE was used as the primary estimator instead of REML, given its limitations:

- The MLE has a wider applicability, and provides consistent estimates using flexible approach, even in cases where some model's assumptions are violated (McCulloch, 1997; McCulloch and Neuhaus, 2011).
- The MLE has an asymptotic property, that is, its results get better as the sample size increases. In ideal situation, it is a good practice to use adequate sample size in a study making the MLE a sufficient estimator in most situations (McNeish, 2017).
- MLE is appropriate for simple random effects models, which is the case this current research(McCulloch, 1997).

The GzLMM models in Chapter 7 (Section 7.4) were fitted using the SAS 9.4 procedure; *PROC GLIMMIX* and *glmer* command in lme4 R's package. The Informed Choice trial (O'Cathain *et al.*, 2002) had a small number of clusters (10 clusters). REML was fitted to its dataset in Section 7.4.3 using SAS *PROC MIXED* with mixed syntax.

6.7 Comparison between the methods

The theoretical properties of the four different methods are compared in **Table 6.2**, and some of these properties are discussed below.

For ILA (using CSMs and PAMs) there are situations where they are equivalent in interpretation. A random intercept LMM typifying a CSM is equivalent to a PAM with an exchangeable working correlation structure for models with collapsible link functions (Ritz and Spiegelman, 2004; Hubbard et al., 2010; Campbell and Walters, 2014b), but both methods are not consistent (i.e., produce biased estimate) when the cluster sizes are informative (Neuhaus and McCulloch, 2011). Cluster sizes are informative if they have an impact on the outcome variable or the random effects due to differing values (Neuhaus and McCulloch, 2011). In other words, component of the outcome variable or random effects are affected by differing cluster sizes (caused by missing data or unequal allocation). Theoretically, the random intercept LMM and PAMs with an exchangeable working correlation estimate different parameters in the case of noncollapsible link functions, and their estimates are biased if the cluster sizes are informative.

In terms of efficiency, concerning the size of the SE of the estimated treatment effect, the GEE1 takes into account the correlation among outcomes within clusters, this improves its efficiency (see, **Table 6.2**, row 8). Although, GEE1 produces a consistent estimate of the intervention effect (and its SE) when the mean model is correct, and missing data are missing completely at random even if the correlation structure is misspecified (Liang and Zeger, 1986). However, GEE1 suffers some loss in efficiency if the working correlation structure is not close to the true structure, especially when the true correlation is large and/or the sample size is small. When the number of clusters is small (which is a recipe for imbalance) the robust SE estimator of GEE1 does not provide full protection over incorrect working correlation structure, causing GEE1 to have reduced efficiency in regard to the estimates of the SEs of the intervention effect (Liang and Zeger, 1986; Leyrat *et al.*, 2018; Thompson *et al.*, 2022).

This disadvantage of the GEE1 is the reason GEE2 and QIF were developed to improve the GEE1's efficiency. GEE2 achieves this by explicitly modelling the mean and correlation parameters simultaneously, using two separate sets of estimating equations. Also, if the mean and association among responses are of interest, GEE2 is more likely to produce efficient inferences for the mean and correlation parameters than GEE1, especially if the correlation in the outcome

is substantial and the sample size is small (Prentice, 1988; Prentice and Zhao, 1991; Yan and Fine, 2004; Crespi, Wong and Mishra, 2009). QIF is another alternative to GEE1 that uses a different strategy to estimate the working correlation parameter, therefore minimising the impact of its misspecification. Studies have proved this advantage of the QIF in the context of a longitudinal study (Qu, Lindsay and Bing, 2000; Odueyungbo *et al.*, 2008; Song *et al.*, 2009). Their results showed that QIF is more efficient than GEE1 when the true correlation is large and misspecified. Several authors have shown that this claim might not necessarily hold when there are few clusters and/or there is cluster and covariate imbalance between treatment arms (Westgate, 2012; Westgate and Braun, 2012, 2013).

The MLE is one of the most common estimators of GzLMM, and it's known to be consistent and efficient when the distributional assumptions that are made are correct. One such assumption is that the random cluster effects are Normally distributed. Previous studies had overstated the impact of misspecifying the distribution of the random effects on MLE (Agresti, Caffo and Ohman-Strickland, 2004; Litière, Alonso and Molenberghs, 2008). However, a recent study has shown that the MLE is quite robust to the impact of misspecifying the random effects in most situations considered previously (Neuhaus and McCulloch, 2011), even when the cluster size is informative (Neuhaus and McCulloch, 2011).

The goodness-of-fit of a statistical model is a crucial part of building an optimal regression model for practical uses. Appropriate goodness-of-fit methods for CSMs have been extensively studied in the literature whereas goodness-of-fit methods for PAMs are few. The early goodness-of-fit methods for GEE-based models involve partitioning the covariates space into separate groups and then calculating their score statistics which are approximately Chi-square distributed (Barnhart and Williamson, 1998; Horton *et al.*, 1999). This strategy is an extension of that of Tsiatis (1980) and Hosmer and Lemesbow (1980) for uncorrelated outcomes. This strategy was found to produce different results in different statistical software because the partitioning is subjective to the software used (Hosmer *et al.*, 1997), and this problem may likely extend to population average models (Pan, 2001).

Pan (2001) proposed a goodness-of-fit method for PAMs that mimics Akaike's Information Criterion (AIC) known as the Quasi-likelihood information criterion (QIC). While the AIC is based on maximum likelihood, QIC is based on quasi-likelihood under independent working covariance in GEE1. The results of the simulation study of the authors showed that the AIC was

more efficient than the proposed QIC, however, the performance of the QIC was remarkable. The author did not clearly state if this criterion applies to GEE2 but noted that using the GEE2 approach to estimate the scale parameter included in their criterion is difficult. A goodness-of-fit method exists for GEE2 in McCullagh and Nelder (1989). To the best of my knowledge, the method is not available in standard statistical packages.

If the regression model includes a binary outcome variable with at least a continuous covariate, the application of goodness-of-tests that are based on Chi-square distribution is technically impossible. The partitioning of the continuous covariate would result in a situation where the total number of the distinct groups is bigger than the sample size. For this reason, Pan (2002) developed two tests; the Pearson chi-square G and the unweighted sum of squares U tests which are based on the Normal distribution with means and variances (using unstructured working correlation).

QIF's goodness-of-fit method is based on an objective function that is approximately chi-square distributed with appropriate DoF. It shares similar asymptotic properties to that of the likelihood ratio test, which is negative twice the log-likelihood $[-2 \times (\log(l(.))]$ (Qu, Lindsay and Bing, 2000). This is one of the advantages QIF has over GEE1 (Qu, Lindsay and Bing, 2000; Odueyungbo *et al.*, 2008; Song *et al.*, 2009). The QIF's objective function can be constructed from models with a correlation structure different from independence, but unlike the GEE1's QIC which is only based on independent working covariance (Pan, 2001).

Table 6.2 Similarities and differences in the methodological properties of the four selected statistical methods for analysing cRCTs

S/NO	Feature	GzLMM	GEE1	GEE2	QIF
1	**Covariate adjustment**	Allows adjustment for both cluster-level and individual-level covariates using an appropriate link function.	Same as GzLMM	Same as GzLMM	Same as GzLMM
2	**Adjustment for clustering**	Clustering is accounted for via a random effects term with its coefficient and that of fixed effects term estimated simultaneously via a mean model.	The structure of clustering is described using a working covariance matrix (characterised by the working correlation parameter) and it is specified separately from the mean model.	A separate set of estimating equations and link functions are used to model the mean and correlation parameters, thereby explicitly explaining the source of the cluster-level variations.	Avoids the direct use of the correlation parameter in its algorithm and instead uses a linear combination of the product of basis matrices and some constants.
3	**Assumption on the distribution of the cluster-level random effects**	Most times in GzLMM it is assumed that the cluster-level random effects follow a parametric distribution, and Normal distribution is a common choice.	As a semi-parametric method, it does not assume any distribution for the cluster-level random effects.	Same as GEE1	Same as GEE1
4	**Multiple forms of clustering**	Accommodates multiple forms of correlation to be investigated by incorporating them as random effects in the mean model.	Allows multiple forms of correlation but through a complex procedure of including higher forms of clustering as fixed effects in the mean model.	Same as GEE1	Same as GEE1
5	**Assumption of missing data mechanism required to obtain consistent parameter estimates**	Missing completely at random and missing at random.	Missing completely at random	Same as GEE1	Same as GEE1
6	**Heterogenous correlation**	Flexible in modelling complex correlation structures using multiple random effects variables.	Not flexible in modelling data with complex correlation structure.	More flexible than GEE1 by using a separate equation, link function, and covariates for the correlation parameter.	Same as GEE1
7	**Improvement in efficiency (i.e., smaller SE of the estimate of the treatment effect)**	Gain in efficiency by including random effects components in the mean model to account for correlation among outcomes in a cluster, especially when the correlation is large.	Gain in efficiency by using a "working covariance matrix" which accounts for the effect of correlation among outcomes in a cluster and treats it as a nuisance.	More gain in efficiency compared to GEE1 by explicitly modelling the effect of the correlation among outcomes with a separate equation that allows covariates adjustment. This provides some protection against misspecification of the correlation structure.	Firstly, it uses a different strategy that protects against the misspecification of the correlation structure. Secondly, it weights the information contributed by each cluster using an empirical weighting matrix, clusters with large variation are given less weight and vice versa. It is acclaimed that these two features increase its gain in efficiency compared to the GEE1..
8	**Moment specification**	First and second-order moments are to be specified.	First and second-order moments are to be specified.	The first four order moments[1], but the third and fourth can be specified as a function of the first two moments since a working correlation is being used.	Same as GEE1
9	**Approximation technique**	Laplace/Adaptive Gauss-Hermite Quadrature[2]	Modified Fisher scoring algorithm	Alternate between the Modified Fisher scoring algorithm and the method of the moment.	Newton-Raphson algorithm
10	**Goodness of fit**	All the model selection criteria that are based on maximum likelihood theory are applicable, such as the LRT, AIC, and BIC.	Uses a modification to the AIC based on a quasi-likelihood theory known as QIC (and QICu[3]) for model and working correlation selections.	Same as GEE1	Provides an objective function that follows a chi-square distribution (which is analogue to the likelihood ratio test).
11	**Availability in selected statistical software, function(package)**	R = glmer(lme4) and SAS = glimmix(proc)	R = glmgee(geepack) and SAS = genmod(proc).	R = geese(geepack) only	R = qif(qif) and SAS = qif(macro)

GzLMM: Generalized linear mixed model; GEE: Generalized estimating equations; QIF: Quadratic inference function; LRT = likelihood ratio test; AIC = Akaike information criteria; BIC = Bayesian information criteria; QIC = Quasi-likelihood independence criterion.

1. The first four order moments of the outcome of interest are the mean, variance, skewness, and kurtosis.
2. Adaptive Gauss-Hermite Quadrature equals the Laplace approximation when the quadrature point/node is 1. Other techniques do exist.
3. QICu is a variant of QIC that allows for the correlation in the data but is not adequate for selecting a working correlation structure (Pan, 2001).

6.8 Summary

In summary, Chapter 6 is an extension to Chapter 2 where classical methods of analysing cRCTs were described. The statistical approaches discussed in Chapter 2 include CLA, RMRSE, GEE1, and GzLMM. However, given the methodological review of methods available in the literature for analysing outcome data from cRCTs in Chapter 3, two emerging methods were identified, GEE2 and QIF, although GEE2 was not directly identified from the review. The practice review of Chapter 3 identified GzLMM and GEE1 as the most used methods in practice. Hence, GzLMM (with MLE/REML), GEE1, GEE2, and QIF were selected for further evaluation in this research.

These four selected methods have been described in more detail in this current Chapter 6. Furthermore, the functions (and statistical package) used to implement the four statistical methods are mentioned, and these methods were also compared to provide a better understanding of their similarities and differences. In general, the parameter estimates from these methods have different interpretations (and sometimes magnitude). Nonetheless, there are circumstances when their parameter estimates are equivalent. For example, with a continuous outcome, the parameter estimates of the intervention effect from an LMM with random intercept and a GEE1 with compound symmetric working correlation structure are equivalent, even when the cluster sizes are informative concerning outcomes.

In Chapter 7, these four methods were applied to outcome data from four cRCTs. The case studies cover some common settings in cRCTs, the focus of Chapter 7 is to observe the performance of the methods across the four cases. It would be of interest to ascertain if the theoretical attributes of the methods match up with their practical attributes given the unique features of the case studies and the limiting nature of finite sample sizes.

Chapter 7

Statistical methods for analysing cRCTs – an empirical analysis of four cRCT datasets

7.1 Introduction

This Chapter presents the results from the application of the four selected statistical methods (GzLMM, GEE1, GEE2, and QIF) to four cRCT datasets. The four case studies measured both continuous and binary outcomes. These four case studies have arisen from the work of my thesis supervisor as applied medical statisticians in clinical trial research, and the case studies are adequate for the current research being conducted. Statistical methods are known to behave differently when applied to real-world data with finite sample sizes. This justifies the need to conduct this empirical study to observe the statistical behaviours of the four selected methods in practice. In Chapter 2, traditional statistical methods for analysing outcome data from cRCTs were briefly described. They include the cluster-level summary measures in conjunction with t-test and regression models, RMRSE, GEE1, and GzLMM.

The methodological review in Chapter 3 identified more available and appropriate methods (27 methods) for analysing outcome data from cRCTs. The most studied traditional methods are GEE1 (23/112), MLE (18/112), Bayesian methods with MCMC (13/112), REML (11/112), and t-test (7/112). Among the newer/emerging methods, QIF was the most studied method (5/112) (**Table 3.1**). It is worth noting that after investigating the literature, we found that QIF and GEE2 are the two most prominent alternatives to the traditional GEE1 method (Turner, 2017). The review of statistical methods used in practice to analyse outcome data from cRCTs in Chapter 4 showed that no trial used GEE2 or QIF. Instead, traditional methods like GzLMM (80%), RMRSE (7%), and GEE1 (6%) were preferred. One of the plausible reasons for this low usage of these alternatives (i.e., GEE2 and QIF) could be because the comparative advantages of these methods have not been convincingly demonstrated to warrant their routine application in cRCTs. With this research gap in mind, this Chapter presents the results from the application of the four selected methods – GzLMM (with MLE/REML), GEE2, and QIF. The focus is to investigate how these methods behave concerning the unique features of each of the four cRCTs.

7.2 Chapter aim

This chapter aims to present the results from the application of the four selected methods (GzLMM, GEE1, GEE2, QIF) to four real-world datasets from cRCTs with different features like different finite sample sizes.

7.3 Software

The statistical packages used to apply the specific methods are summarised in **Table 7.1**. SAS (version 9.4) and R (version 4.2.1) were used to analyse the outcome data from the four cRCTs using the four selected methods. Two of the methods, GzLMM (with MLE/REML) and QIF, were fitted using SAS while GEE1 and GEE2 were fitted using R. The SAS syntax and R codes for fitting all the statistical models to one case study (the PoNDER trial) with continuous and binary outcomes are provided in Appendix 4.

The initial plan was to fit all the models using free and open software such as R, but it was observed that the *qif* command in R's *qif package (CRAN - Package qif (r-project.org)*, last assessed on 20th June 2022*)* could not fit QIF model to trials with clusters of size one, PoNDER and Bridging the Age Gap trials have clusters of size one. The error message suggests that it is a problem of the incompatibility of the matrices in the matrix multiplication procedure. So, SAS was used instead, and it was able to overcome the problem. Also, the *lmer* function of *lme4* package for fitting LMMs in R does not have AGHQ as an option but *glmer* does. However, the SAS procedure for fitting both LMM and GzLMM: *GLIMMIX,* has an AGHQ option for both continuous and binary outcomes.

Hence, GzLMMs were fitted using the *GLIMMIX* procedure in SAS and the quadrature point (nodes) was set to ten for the AGHQ algorithm. Higher nodes increase the complexity of the AGHQ procedure but produce more reliable results than lower nodes (Handayani *et al.*, 2017) The *GLIMMIX* procedure does not produce a value for the ICC, so we calculated it using the estimates of the between and within cluster variances from *PROC GLIMMIX* output. GEE1 models were fitted using the *geeglm* function (from R's *geepack* package) with an exchangeable correlation structure and robust SEs, and so was GEE2 using the *geese* function. It is worth noting that for the GEE2 models, no covariate was included in the working correlation and scale structures. Recall from Section 6.4.2 that GEE2 does allow for covariates in the working correlation and scale model

equations which could be different (or the same) as the ones included in the mean model equation. GEE2 uses different link functions other than the one specified for the mean model equation to achieve this. These are the specifications that differentiate GEE2 from GEE1. QIF models were fitted using the *qif* macro in SAS. The link function for the mean model was either identity for continuous or logit for a binary outcome, for the scale model it was the identity, and for the correlation model, it was the modified Fisher's z transformation.

Table 7.1 Summary of the statistical software used in the analyses of the four cRCT datasets

Model	Statistical software	Procedure/ Package	Function
MLE	SAS	GLIMMIX	*glimmix*
REML	SAS	MIXED	*mixed*
QIF	SAS	QIF (macro)	*qif*
GEE1	R	GEEPACK	*geeglm*
GEE2	R	GEEPACK	*geese*

7.4 Analysis strategies

The sample size characteristics of all the case studies are summarised using frequency and percentage. All the models were fitted using only complete cases from the datasets. Among the case studies, the range of the missing data was from 0% to 7%, which is negligible, hence no sensitivity analysis was conducted. In clinical trials, it is a common strategy to fit both unadjusted and adjusted models containing different numbers of covariates.

The unadjusted model contains only the indicator variable x_{1ij} for the randomised treatment arms as a covariate. While the adjusted model contains other known prognostic factors $\boldsymbol{X}_{pij}^{T}$ (with the treatment arm indicator inclusive), such as baseline outcome values, age, and sex. There are several known benefits from adjusting for prognostic covariates in an adjusted analysis, such as protection against imbalance in baseline participant prognostic covariates among groups (Kahan *et al.*, 2014), increased power and precision for linear models (Hauck, Anderson and Marcus, 1998; Kahan *et al.*, 2014; Samsa and Neely, 2018), to obtain an estimate of the intervention effect that has a closer individual level interpretation, and to account for special features of the study design like stratification and subgroup consideration (Campbell *et al.*, 2012). A study used simulations to show that adjusting for prognostic and non-prognostic covariates led to increased and reduced power, respectively (Kahan *et al.*, 2014).

Unadjusted and adjusted analyses were conducted using each method. In each analysis a P-value < 0.05 represents a statistically significant result. The Informed Choice trial had a few clusters (ten clusters). Studies with few clusters have a higher risk of imbalance in covariates and outcomes, among the treatment arms (Westgate, 2012; Westgate and Braun, 2012, 2013; Leyrat *et al.*, 2018; Samsa and Neely, 2018). Hence, for a study with a continuous outcome and clusters ≤ 20, small sample corrections are required to maintain the nominal 5% Type I error and reasonable power, whereas, for GEE1 with clusters ≤ 40, small sample correction is highly recommended (Leyrat *et al.*, 2018). Similarly, if the study measured a binary outcome and the number of clusters randomised is ≤ 30 if a GzLMM is used then a small sample correction should be applied to the degrees of freedom (DoF), which is the number of clusters minus cluster-level parameters estimated (Thompson *et al.*, 2022).

The Informed choice trial had ten clusters. To analysis the outcome data from the trial a GzLMM was fitted using both MLE and REML estimators. No small sample correction is compatible with the MLE (with AGHQ), so REML was used with Satterthwaite (SAT) correction applied to correct the model's DoF. Corrections on the DoF of the parameter estimates only affect the CIs and P-values, but the point estimate of the intervention effect (and its SE) remains the same as that of the uncorrected version (Leyrat *et al.*, 2018). For GEE1, Fay and Graubard (FG) correction was applied to correct the robust SEs of the parameter estimates, which will consequently affect only their CIs and P-values (Fay and Graubard, 2001). All the corrections used are available in R and SAS, but for GEE2 and QIF there were not readily available or easily implementable corrections in standard statistical packages that I am aware of at the time of authoring this thesis.

Table 7.2 Summary of the sample size of the four cRCTs case studies

Trial	No. of clusters	No. of clusters missing	No. of subject	Average cluster size	(Min, Max) cluster size	Median cluster size	Missing n (%)
PoNDER	101	1	2659	27	(1, 101)	21	35 (1)
Informed Choice	10	0	1547	155	(74, 308)	145	108 (7)
Bridging the Age Gap	43	0	748	18	(1, 73)	16	36 (5)
NOSH	92	0	9207	100	(12, 333)	75	0 (0)

7.5 Analysis of PoNDER trial

Data: The PoNDER cRCT (Morrell *et al.*, 2009) aimed to assess the effect of two psychologically informed interventions by health visitors on postnatal depression (PD) in postnatal women who have recently given birth. The descriptive statistics of the PoNDER trial size are presented in

Table 7.2. One hundred and one general practices (clusters) in the Trent region of England were included in the trial. The general practices were randomised in a 2:1 ratio to the intervention group (n=63 clusters) or the control group (n=38 clusters). Health visitors in the intervention clusters were trained to identify depressive symptoms at six to eight weeks postnatally using the Edinburgh postnatal depression scale (EPDS) and were also trained in providing psychologically informed sessions based on cognitive behavioural or person-centred principles for an hour a week for eight weeks. Health visitors in the control group provided usual care.

The primary outcome was the score on the EPDS at six months follow-up. The EPDS consists of ten questions and generates a score on a 0 to 30 scale with higher scores indicating a great risk of depression. For the PoNDER trial, this outcome was dichotomised into a binary outcome of EPDS score < 12 vs. ≥ 12 with women with a score of 12 or more classified as "at risk" of PD. One hundred (n=63 intervention, n=37 control) clusters and n=2659 new mothers (1745 Intervention: 913 Control) provided valid primary outcome data at 6 months. Also, one of the secondary outcomes in the PoNDER trial "the mean EPDS score at six months" was used as a continuous outcome in this study. In the original study, both outcomes were analysed using GEE1 and an exchangeable correlation structure with robust standard errors.

Results: The mean age of all the women in the control and intervention groups was the same (32±5yrs, respectively), and the maximum age across all women was 46 years. The proportion of women with EPDS score ≥ 12 at 6 months was 16% (150/914) in the control arm and 12% (205/1745) in the intervention arm. For the other outcome "the mean EPDS score at six months" it was 6.4 (SD = 5.0) vs 5.5 (SD = 4.9) for the control vs the intervention arms, respectively. It is worth noting that for both outcomes, smaller is better. The results for the unadjusted intervention effect from the analysis of the continuous primary outcome are slightly different among the models except for GEE1 and GEE2 which were the same both in size and direction (i.e., -0.98). After adjustments were made for the baseline EPDS 6 weeks score, living alone, previous history of major life events, and previous history of PD, the intervention effect became the same across the models (mean difference, -0.78) except for QIF (-0.84).

The standard errors of the intervention effect estimates are the same across the models, ranging from 0.25 to 0.28 for the unadjusted models and 0.20 to 0.21 for the adjusted models. The intervention effect estimates across all the models were significant as evidenced by the small P-values (<0.05) and the confidence intervals which excluded zero. Similar results were obtained

from the binary primary outcome analysis, the odds ratio was 0.67 in all the unadjusted models and adjusted models, except for QIF (0.66 and 0.62 respectively), and all were significant as well, suggested by the small P-values and confidence intervals excluding one (**Table 7.3**). These results are graphically compared using forest plots in **Figure 7.1**, in the plots all the point estimates for the intervention effect and the associated 95% CIs are to the left-hand side of zero favouring the intervention arm. The distance between the left and right whiskers that indicate the 95% CIs are the same for all the methods

Table 7.3 Summary of the results obtained from analysing the PoNDER trial data with the four different statistical methods (N = 2659)

		Continuous outcome[1]				Binary outcome[2]			
Parameter	**Type of modelling**	**GzLMM**	**GEE1**	**GEE2**	**QIF**	**GzLMM**	**GEE1**	**GEE2**	**QIF**
Intervention effect[3]	Unadjusted	-0.97	-0.98	-0.98	-0.94	0.67	0.67	0.67	0.66
	Adjusted*	-0.78	-0.78	-0.78	-0.84	0.67	0.67	0.67	0.62
SE	Unadjusted	0.25	0.28	0.28	0.28	0.13	0.14	0.14	0.14
	Adjusted*	0.20	0.21	0.21	0.20	0.13	0.13	0.13	0.13
P-value	Unadjusted	0.0002	0.0005	0.0005	0.0009	0.0025	0.0032	0.0032	0.0019
	Adjusted*	0.0001	0.0001	0.0001	<0.0001	0.0019	0.0019	0.0019	0.0001
95% CI	Unadjusted	-1.47 to -0.47	-1.53 to -0.43	-1.53 to -0.43	-1.50 to -0.39	0.51 to 0.86	0.51 to 0.87	0.51 to 0.87	0.51 to 0.86
	Adjusted*	-1.17 to -0.39	-1.18 to -0.38	-1.18 to -0.38	-1.24 to -0.44	0.52 to 0.86	0.52 to 0.86	0.52 to 0.86	0.48 to 0.79
ICC	Unadjusted	0.0167	0.0191	0.0382	0.0191	0.0167	0.0063	0.0126	0.0063
	Adjusted*	0.0077	0.0081	0.0162	0.0081	<0.0001	-0.0018	-0.0036	-0.0018
Number of subjects	Unadjusted	2659	2659	2659	2659	2659	2659	2659	2659
	Adjusted*	2624	2624	2624	2624	2624	2624	2624	2624
Number of clusters	Unadjusted	100	100	100	100	100	100	100	100
	Adjusted*	100	100	100	100	100	100	100	100

* Model adjusted for EPDS score at 6 weeks, living alone (no or yes), previous history of major life events (no or yes), and any previous history of postnatal depression (no or yes). SE = Standard error; CI: Confidence interval; ICC: Intracluster correlation coefficient. GzLMM: Generalized linear mixed model; GEE: Generalized estimating equations; QIF: Quadratic inference function.

1. EPDS score at 6 months postnatally. The EPDS is scored on a 0 to 30 scale with higher scores indicating a greater risk of PND.
2. Dichotomised EPDS score at 6 months postnatally of < 12 or >=12.
3. The intervention effect for the continuous outcome is the difference in the mean 6-month EPDS scores between the intervention and control groups; with a negative mean difference favouring lower scores (better outcomes) in the intervention group. The intervention effect for the binary outcome is the odds ratio for an EPDS score of 12 or more in the intervention group compared to the control group with an odd ratio <1 favouring better outcomes (lower odds of PND) in the intervention group.

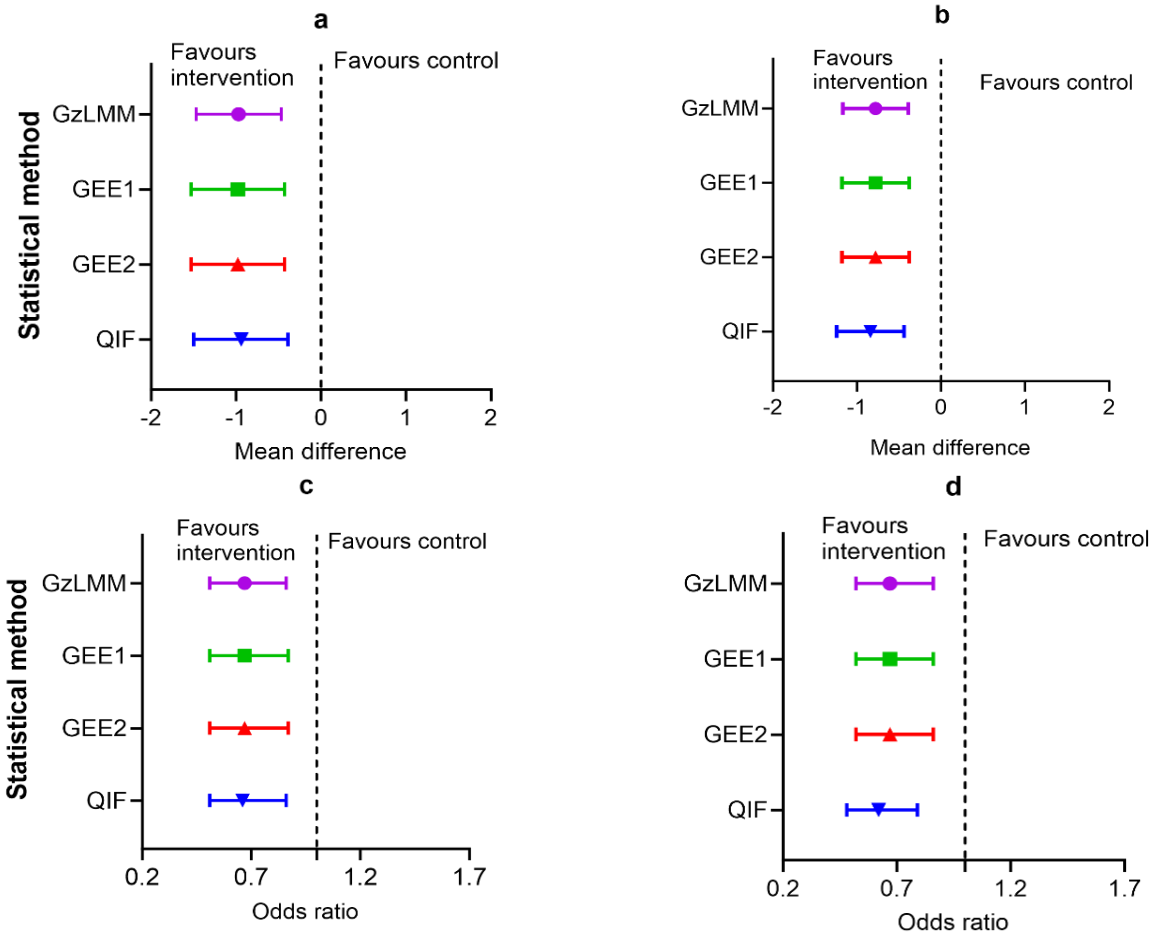


Figure 7.1 Forest plots showing the intervention effect estimate and its associated 95% CI for each of the four statistical methods applied to outcome data from PoNDER cRCT; plots a & b are the unadjusted and adjusted models for continuous outcome, and c & d are for binary.

7.6 Analysis of Bridging the Age Gap Trial

Data: Bridging the Age Gap trial (Wyld *et al.*, 2021) investigated the effects of two decision support interventions (DESI) to support treatment choices in older women (aged≥70 years) with operable breast cancer. Forty-three breast cancer units (clusters) in England and Wales were included in the trial. The breast cancer units were randomised to have access to the DESI (Intervention group n=21 clusters) or to continue with usual care (Control group n=25 clusters). The DESI comprised an online algorithm, booklet, and brief decision aid to inform choices between surgery plus adjuvant endocrine therapy versus primary endocrine therapy, and adjuvant chemotherapy versus no chemotherapy.

The primary outcome was the global health status/quality of life (QoL) score (questions 29 and 30) on the cancer-specific patient-reported outcome of the European Organisation for the Research and Treatment of Cancer (EORTC) QoL questionnaire (QLQ)-C30 at 6 months post-baseline. The EORTC QLQ-C30 global health status/QoL scale is scored on a 0 to 100 scale with a higher score representing a better QoL. Forty-three clusters (n=19 intervention, n=24 control), and 748 patients (359 Intervention: 389 Control) provided valid primary outcome data at 6 months. The primary endpoint was a continuous outcome "Global health status quality of life score" measured 6 months after diagnosis and was analysed using GEE1 with sandwich (robust) SEs and an exchangeable working correlation structure. The total number of participants included in the trial is 748 distributed across the forty-three clusters and the cluster sizes ranged from 1 to 73. The complete description of the trial size is provided in **Table 7.2**.

Results: The mean global health status/quality of life (QoL) score at the 6-month follow-up was 68.9 (SD=19.6) for the control arm, and 69.0 (SD=19.5) for the intervention arm. The results from the analysis of the continuous outcome are summarised in **Table 7.4** and graphically shown in **Figure 7.2**. Bridging the Age Gap trial had a moderate number of clusters (43 clusters) with 748 patients in total. The unadjusted models appear to produce differing estimates of the intervention effect ranging from a mean difference of -0.28 to 0.12 but became stable after the baseline QoL variable (*ql scale*) was adjusted for; the mean difference became 1.71 for all the models except for QIF (1.46).

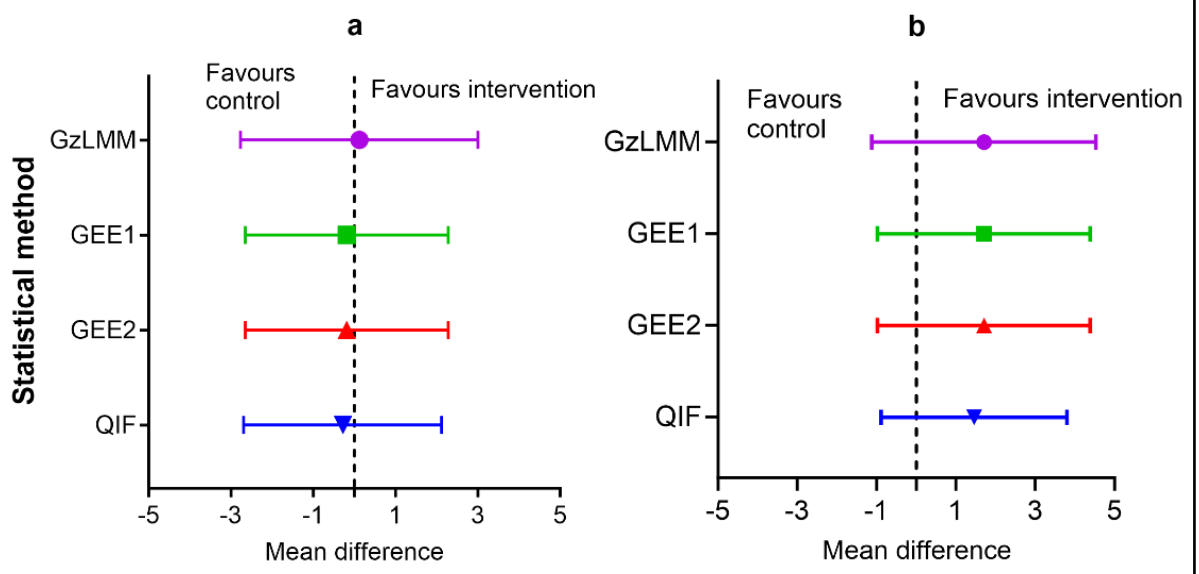


Figure 7.2 Forest plots showing the intervention effect estimate and its associated 95% CI for each of the four statistical methods, and plots a & b are the unadjusted and adjusted model with the continuous outcome, respectively.

However, all the estimates of the intervention effect across models were not significant (i.e., P > 0.05). The SEs are the same for the adjusted models (1.40) except for QIF (1.20). All the SE estimates from QIF were lesser compared to the other three methods, lesser SE is indicative of better precision when the method is not biased towards the null (Morris, White and Crowther, 2019). However, lesser SEs from QIF should be interpreted with caution, because QIF produced different estimates of the intervention effect compared to the other three models which could be indicative of biasedness. For the unadjusted models, when the ICC was negative the intervention effect estimates for GEE1, GEE2, and QIF were all negative. Finally, none of the estimates of the intervention effect was significant (i.e., P > 0.05).

Table 7.4 Summary of the results from different models on outcome data from Bridging the Age Gap trial with a continuous primary outcome[1] (N = 748)

	Unadjusted model				Adjusted model**			
Parameters	**GzLMM**	**GEE1**	**GEE2**	**QIF**	**GzLMM**	**GEE1**	**GEE2**	**QIF**
Intervention effect[2]	0.12	-0.19	-0.19	-0.28	1.71	1.71	1.71	1.46
SE	1.43	1.26	1.26	1.23	1.40	1.37	1.37	1.20
P-value	0.9343	0.8818	0.8810	0.8175	0.2294	0.2127	0.2127	0.2230
95% CI	-2.77 to 3.00	-2.65 to 2.28	-2.65 to 2.28	-2.69 to 2.12	-1.12 to 4.53	-0.98 to 4.39	-0.98 to 4.39	-0.89 to 3.80
ICC	<0.0001	-0.0068	-0.0135	-0.0068	0.0042	0.0028	0.0056	0.0028
Number of subjects	748	748	748	748	712	712	712	712
Number of clusters	43	43	43	43	43	43	43	43

** Model adjusted of global QoL baseline values. SE = Standard error; CI: Confidence interval; ICC: Intracluster correlation coefficient; GzLMM: Generalized linear mixed model; GEE: Generalized estimating equations; QIF: Quadratic inference function

1. Global QoL score on the EORTC-C30 at 6 months post-baseline. The EORTC-C30 Global scale is scored on a 0 (poor) to 100 (good health) scale.
2. The intervention effect for the continuous outcome is the difference in the mean 6-month Global QoL scores between the intervention groups; with a positive mean difference favouring higher scores (better outcomes) in the intervention group.

7.7 Analysis of Informed Choice trial

Data: The Informed Choice (IC) study (O'Cathain *et al.*, 2002) was aimed at investigating the impact of a set of ten pairs of evidence-based leaflets – The Midwives' Information and Resource Service (MIDIRS) and NHS Centre for Reviews and Dissemination informed choice leaflets through a survey. The study was designed to cover eight of the ten MIDIRS decision points in everyday maternity care. Conducted in twelve large maternity units in Wales, the maternity units were grouped into ten clusters. Pairs of clusters were randomly assigned to the intervention and control arms based on their annual numbers of deliveries, to achieve balance.

The primary objective was to improve the management of women during pregnancy and childbirth, by assessing the effect of an intervention that promotes informed choice. The primary binary outcome was the change in the proportion of women who reported exercising informed choice (yes or no). For illustration, one of the secondary outcomes "the average of the women's levels of knowledge" on the ten topics covered in the survey was used as a continuous outcome in this current study. Knowledge of the topics was assessed on a 1 (poor) to 10 (good) scale. Two samples of different women were surveyed the antenatal and postnatal samples. The antenatal sample is made up of all women who reached 28 weeks' gestation period during six weeks and were receiving antenatal care in any setting. The questionnaire used for the cohort covered three decision points that the women may have encountered. The postnatal sample was made up of all women who delivered live babies within six weeks. A questionnaire that covered the remaining five decision points was used to survey the women postnatally.

The postnatal sample had a total of 3,288 women, who were cross-sectionally surveyed before (n=1,741) and after the intervention was administered (n=1,547). **Table 7.2** presents the descriptive summary of the trial size. However, to demonstrate the fitting of the statistical methods in this study only the follow-up (i.e., after the intervention) postnatal sample was used and reported. Furthermore, the matched paired clusters were unmatched and analysed to prevent the reduction of the already few clusters (to 5 unpaired clusters from 10 paired), which is too small. A simulation study shows that when the relevant matching variables are not certainly known, and the number of matched pairs is small (<10), and unmatched study design and analysis is more powerful than a matched design and analysis. However, in a case where matching study design was used and the pairs are small, then the power of the study could be attained by conducting an unmatched analysis using the matched data (Diehr *et al.*, 1995).

Additionally, I intended for this current analysis to be consistent with the original analysis by the original researchers, which was an unmatched analysis with an matched data(O'Cathain *et al.*, 2002). Only women who delivered in all settings and above the age of 16 years were included. Random effects models (i.e., GzLMM) were used to analyse the outcomes in the original study. The unique feature of this cRCT is in the small number of clusters that were randomised, which was ten clusters. So small sample corrections described in Section 7.4 were applied.

Results: The IC trial had a small number of clusters (10 clusters) with many subjects (1547 participants). In the intervention arm, 59% (477/816) of the women reported having exercised informed choice while using the maternity service compared to 57% (346/612) in the control arm. The mean knowledge of the 10 topics covered in the survey for those in the intervention arm was 3.6 (SD = 1.62) compared to the 3.3 (SD = 1.60) of the control arms. The individual level covariates in the adjusted models are the age of the mother, the age at which the mother left education, parity, and the delivery style, and no cluster level covariate was adjusted for.

The results of the unadjusted and adjusted models from the analysis of the continuous and binary outcomes are presented in **Table 7.5** and visualised in **Figure 7.3** For the continuous outcome, the unadjusted intervention effect was the same for the three models (mean difference = 0.20, SE = 0.11) but different for QIF (0.03, SE = 0.05) (**Figure 7.3a**). Similarly, the adjusted intervention effects were the same 0.22 (SE = 0.1) for all the models except QIF 0.05 (SE = 0.02) (**Figure 7.3b**). The parameter estimates of the intervention effect for the QIF models are far more inconsistent with the observed data (difference in mean score = 0.3). The unadjusted analysis was not significant for all the models (i.e., $P > 0.05$). The adjusted analysis was significant (i.e., $P < 0.05$) for all the methods except for GzLMM.

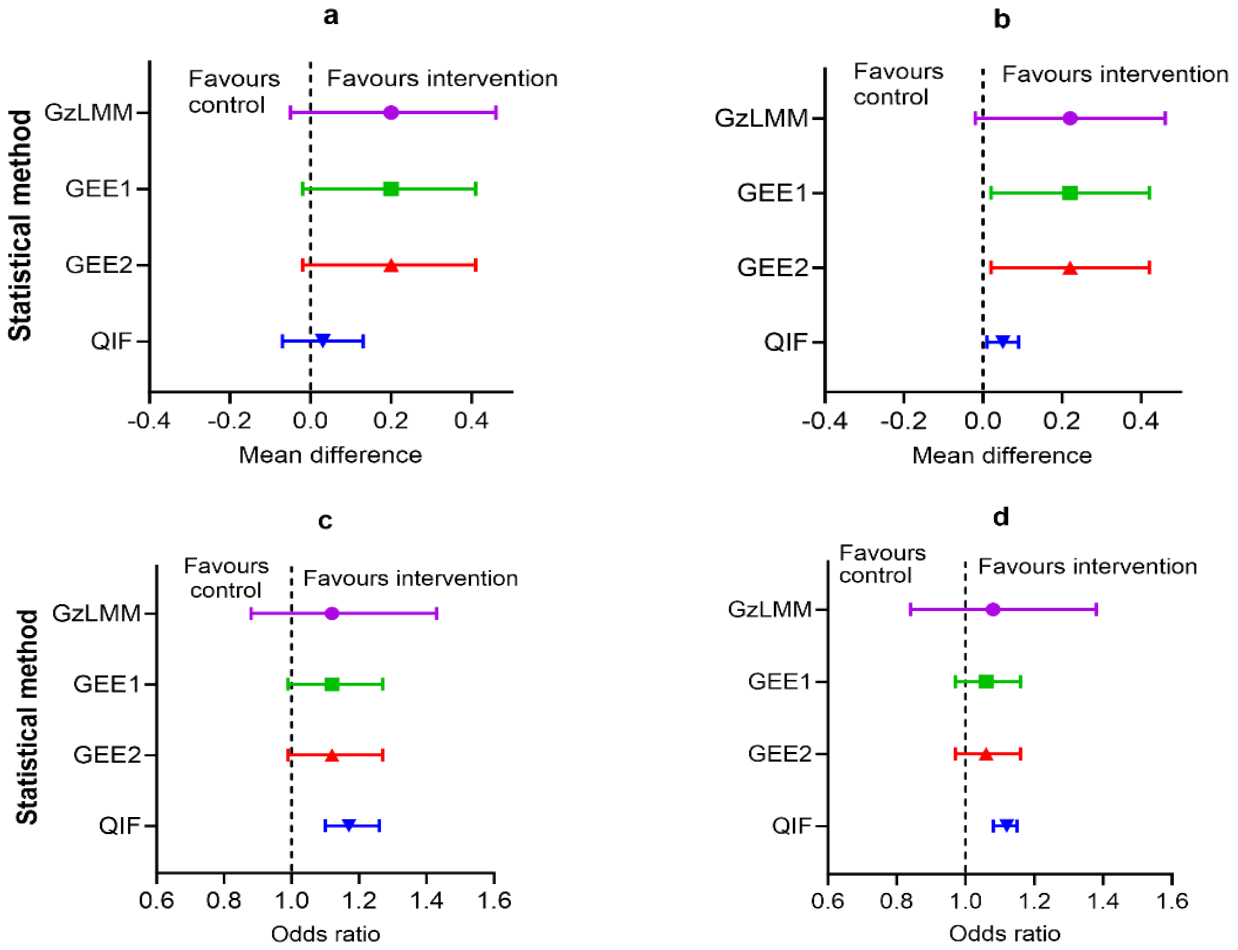


Figure 7.3 Forest plots showing the intervention effect estimate and its associated 95% CI for each of the four statistical models, plot a & b are the unadjusted and adjusted models for a continuous outcome, and c & d are that of the binary outcome.

Table 7.5 Summary of the results obtained from analysing the data from the IC postnatal trial with the different statistical methods (N = 1547)

Parameters	Type of modelling	Continuous outcome[1]				Binary outcome[2]			
		GzLMM	**GEE1**	**GEE2**	**QIF**	**GzLMM**	**GEE1**	**GEE2**	**QIF**
Intervention effect[3]	**Unadjusted**	0.20	0.20	0.20	0.03	1.12	1.12	1.12	1.17
	Adjusted* **	0.22	0.22	0.22	0.05	1.08	1.06	1.06	1.12
SE	**Unadjusted**	0.11	0.11	0.11	0.05	0.11	0.06	0.06	0.04
	Adjusted* **	0.10	0.10	0.10	0.02	0.11	0.05	0.05	0.07
P-value	**Unadjusted**	0.1030	0.0730	0.0731	0.5306	0.3178	0.0647	0.0647	<0.0001
	Adjusted* **	0.0676	0.0324	0.0324	0.0158	0.5206	0.2175	0.2175	<0.0001
95% CI	**Unadjusted**	-0.05 to 0.46	-0.02 to 0.41	-0.02 to 0.41	-0.07 to 0.13	0.88 to 1.43	0.99 to 1.27	0.99 to 1.27	1.10 to 1.26
	Adjusted* **	-0.02 to 0.46	0.02 to 0.42	0.02 to 0.42	0.01 to 0.09	0.84 to 1.38	0.97 to 1.16	0.97 to 1.16	1.08 to 1.15
ICC	**Unadjusted**	0.0042	0.0027	0.0055	0.0027	0.0000	-0.0029	-0.0058	-0.0029
	Adjusted* **	0.0029	0.0018	0.0036	0.0018	0.0000	-0.0036	-0.0072	-0.0032
Number of subjects	**Unadjusted**	1534	1534	1534	1534	1485	1485	1485	1485
	Adjusted* **	1474	1474	1474	1474	1439	1439	1439	1439
Number of clusters	**Unadjusted**	10	10	10	10	10	10	10	10
	Adjusted* **	10	10	10	10	10	10	10	10

***Model adjusted for mother's age, age mother left education, parity, and delivering style. SE = Standard error; CI: Confidence interval; ICC: Intracluster correlation coefficient; GzLMM: Generalized linear mixed model; GEE: Generalized estimating equations; QIF: Quadratic inference function.

1. Knowledge of informed choice leaflets score at 8 weeks postnatally. Knowledge is scored on a 0 to 10 scale with higher scores indicating a greater knowledge of the leaflets.
2. Proportion of women who answered "yes" to the question "Have you had enough information and discussion with midwives or doctors to make a choice together about all the things that happened during maternity care?" with the options "yes," "partly," "no," "there was no choice," and "did not apply."
3. The intervention effect for the continuous outcome is the difference in the mean 6-week knowledge scores between the intervention and control groups; with a positive mean difference favouring (better outcomes) in the intervention group. The intervention effect for the binary outcome informed choice (yes or no) is the odds ratio for yes to overall informed choice in the intervention group compared to the control group with an odd ratio >1 favouring better outcomes (higher odds of an informed choice) in the intervention group.

Similarly, for the primary binary outcome, the unadjusted odds ratio of women who reported exercising informed choice in the intervention arm compared to the control arm was the same for all the models (odds ratio = 1.12, SE = 0.10 to 0.11) except for QIF (1.17, SE = 0.04). The adjusted odds ratio for all the models was the same (odds ratio = 1.1, SE = 0.10 to 0.11). The odds ratios for the unadjusted and adjusted intervention effect were not significant for all the models except QIF which was highly significant (P < 0.0001) (see, **Table 7.5**).

Table 7.6 Small sample size corrections applied to outcome data from Informed Choice cRCT with ten clusters

		Continuous outcome[1]				Binary outcome[2]			
Method	**Type of modelling**	**Intervention effect**	**SE**	**P-value**	**95% CI**	**Intervention effect**	**SE**	**P-value**	**95% CI**
$GzLMM_{Sat}$	**Unadjusted**	0.20	0.11	0.1371	(-0.09, 0.52)	1.12	0.11	0.4796	(0.29, 4.31)
	Adjusted* **	0.22	0.10	0.0930	(-0.05, 0.52)	1.08	0.11	0.6234	(0.27, 4.26)
$GEE1_{FG}$	**Unadjusted**	0.20	0.11	0.1853	(-0.13, 0.53)	1.12	0.06	0.3229	(0.79, 1.61)
	Adjusted* **	0.22	0.10	0.1086	(-0.06, 0.50)	1.06	0.05	0.5495	(0.80, 1.38)

***Model adjusted for mother's age, age mother left education, parity, and delivering style. SE = Standard error; CI: Confidence interval; GzLMM: Generalized linear mixed model; GEE: Generalized estimating equations; QIF: Quadratic inference function; Sat: Satterthwaite; FG: Fay & Graubard.

1. Knowledge of informed choice leaflets score at 8 weeks postnatally. Knowledge is scored on a 0 to 10 scale with higher scores indicating a greater knowledge of the leaflets.
2. Proportion of women who answered "yes" to the question "Have you had enough information and discussion with midwives or doctors to make a choice together about all the things that happened during maternity care?" with the options "yes," "partly," "no," "there was no choice," and "did not apply."

The results of applying small sample corrections are summarised in **Table 7.6**. When compared to the results from the uncorrected version in **Table 7.5** the differences lie in the P-values and 95% CIs of the treatment effect estimates, for both the continuous and binary outcomes. The corrected P-values are bigger, and the CIs are wider (**Table 7.6**).

7.8 Analysis of Nourishing Start for Health (NOSH) trial

Data: The NOSH trial assessed the effect of an area-level financial incentive (shopping vouchers) on breastfeeding in new mothers and babies in areas with low breastfeeding prevalence (Relton *et al.*, 2018). Ninety-two electoral ward areas (the clusters) in England were included in the trial with baseline breastfeeding prevalence at 6 to 8 weeks postnatally of less than 40%. The areas were randomised to the financial incentive plus usual care (n = 46 clusters) or usual care alone (n = 46

clusters). **Table 7.2** shows that all 92 clusters provided breastfeeding outcome data on 9,207 mother-infant pairs (4973 in the NOSH group, 4324 in the control group). The primary outcome was the electoral ward area-level 6 to 8 weeks breastfeeding prevalence, as assessed by clinicians at the routine 6 to 8 weeks postnatal check. This was derived from the number of new mothers who were breastfeeding or not at 6 weeks in each local authority area/cluster. A cluster-level approach was used to analyse the primary outcome after obtaining a summary measure for each cluster. Specifically, a weighted multiple linear regression model was used in the original study.

Results: Overall, 36% (1869/4973) of mothers in the forty-six clusters of the NOSH group were breastfeeding at 6 weeks compared to 30% (1299/4324) in the 46 clusters of the control group. The statistical analysis adjusted for cluster-level baseline prevalence and local government area as covariates. For the NOSH case study, only a binary outcome was measured. The results from the unadjusted and adjusted analysis are presented in **Table 7.1** and graphically presented in **Figure 7.4**. The odds ratios that the mothers were breastfeeding at the end of the trial were the same for all the unadjusted (odds ratio = 1.40) and adjusted (odds ratio = 1.30) intervention effect estimated from the four methods and were statistically significant. This was similar for the SEs = 0.08 for all the unadjusted models and 0.07 for all the adjusted models, except for GEE2 (0.05). The ICCs from the unadjusted and adjusted GEE2 models were quite different from the other methods. The ICC ranged from 0.02 to 0.04 for the unadjusted analysis and 0.004 to 0.02 for the adjusted analysis (see, **Table 7.1**).

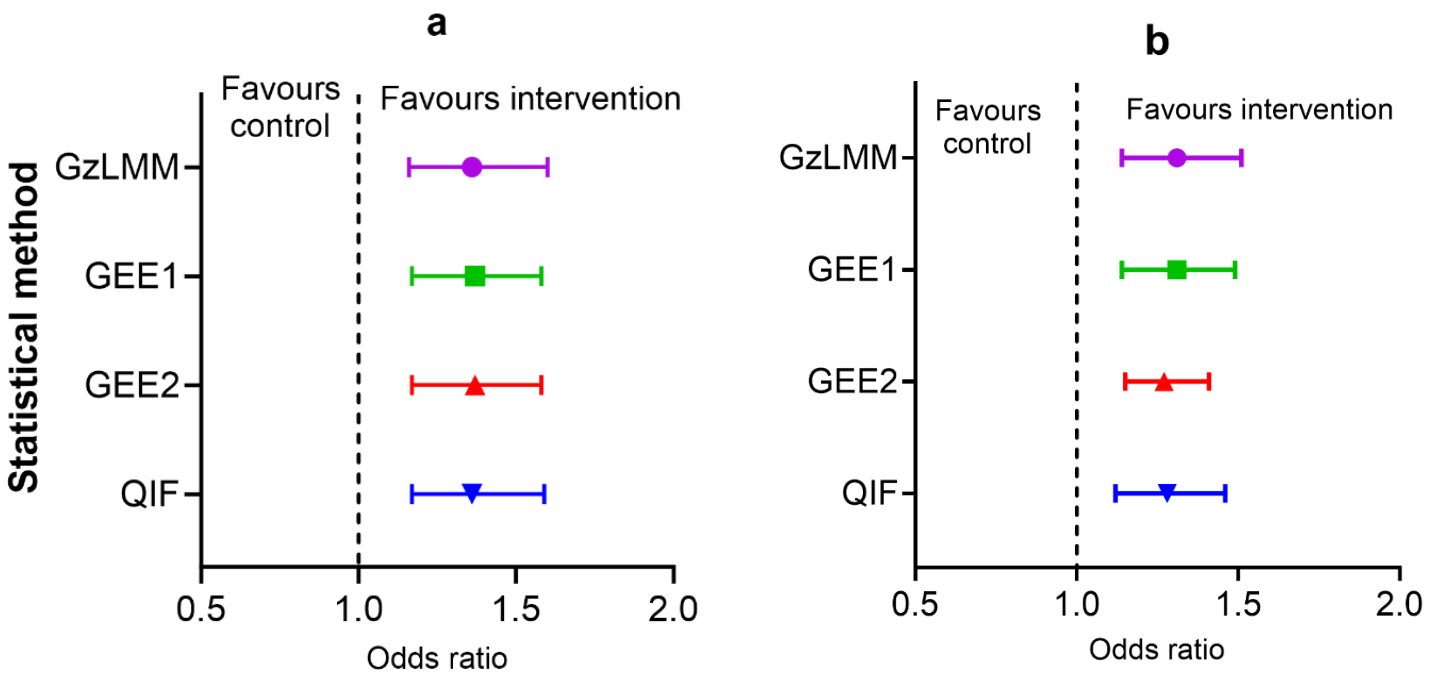


Figure 7.4 Forest plots showing the intervention effect estimate and its associated 95% CI from each of the four statistical methods, plots a & b are the results of the unadjusted and adjusted analyses with the binary outcome respectively.

Table 7.1 Summary of the results obtained from analysing the binary outcome data from NOSH trial[1]

	Unadjusted model				Adjusted model†			
Parameters	**GzLMM**	**GEE1**	**GEE2**	**QIF**	**GzLMM**	**GEE1**	**GEE2**	**QIF**
Intervention effect[2]	1.37	1.36	1.36	1.36	1.31	1.31	1.27	1.28
SE	0.08	0.08	0.08	0.08	0.07	0.07	0.05	0.07
P-value	0.0002	<0.0001	<0.0001	0.0009	0.0002	<0.0001	<0.0001	0.0002
95% CI	1.16 to 1.60	1.17 to 1.58	1.17 to 1.58	1.17 to 1.59	1.14 to 1.51	1.14 to 1.49	1.15 to 1.41	1.12 to 1.46
ICC	0.0262	0.0192	0.0383	0.0192	0.0162	0.0098	0.0042	0.0098
Number of subjects	9207	9207	9207	9207	9207	9207	9207	9207
Number of clusters	92	92	92	92	92	92	92	92

†The statistical models were adjusted for the cluster-level baseline breastfeeding rate and local government area. SE = Standard error; CI: Confidence interval; ICC: Intracluster correlation coefficient; GzLMM: Generalized linear mixed model; GEE: Generalized estimating equations; QIF: Quadratic inference function

1. The binary outcome was if the mother was breastfeeding her baby at 6 weeks postnatally (response value = 1) or not (response value = 0).
2. The intervention effect for the binary outcome is the odds for breastfeeding at 6 weeks postnatally in the NOSH intervention group compared to the odds of breastfeeding in the control group with an odds ratio >1 favouring better outcomes (higher odds of breastfeeding) in the intervention group.

7.9 Discussion

In this Chapter, four different methods for analysing outcome data from cRCTs with clustering in the treatment arms have been applied. The GzLMM, GEE1, GEE2, and QIF were applied to four cRCT case studies with different features, the focus is to demonstrate their implementation and evaluate their use in practice. To the best of my knowledge, this study is the first study to compare these four methods in the context of cRCTs. The initial intention was to use a free and open-source software package to analyse the datasets such as R, but I resorted to using R to fit only GEE1 and GEE2. SAS macro "QIF" was used for QIF because its sister version in R could not fit the QIF models to datasets of trials with a cluster size of one (i.e., only one outcome was observed in the cluster). The PoNDER and Bridging the Age Gap trials had clusters with one observed outcome only. This was communicated to one of the authors of both software packages, Peter X.K. Song, through email correspondence and Song promised to investigate it. It is worth noting that in cases where QIF can be used in R and SAS – where the cluster sizes are greater than one participant, the results from the two statistical packages are identical. Similarly, the results from R and SAS are identical for GEE1 and GzLMM in cases studied.

The case studies considered have estimates for the ICC which are similar to the values reported in trials in primary care (Adams *et al.*, 2004) and community trials (Ukoumunne *et al.*, 1999). All had an ICC less than 0.05 and three studies had an ICC less than 0.02. This indicates there was little clustering of outcomes as would be expected in primary care and community based clinical trials(Ukoumunne *et al.*, 1999; Adams *et al.*, 2004). Three studies had negative estimates for the ICC, from GEE and QIF methods. Theoretically, the ICC is bounded between 0 and 1. But in practice, negative ICCs can be realised from real-world data with finite samples. The GzLMM model truncates the ICC to zero instead of producing a negative ICC (Nelder and Wedderburn, 1972), but that is not the same for the other three PAMs – GEE1, GEE2, and QIF (Eldridge, Ukoumunne and Carlin, 2009). Our results confirmed this, only the marginal models produced negative ICCs (**Table 7.3**, **Table 7.4**, and **Table 7.5**).

Upon reading the documentation of the functions for fitting the population average models, *geeglm (for GEE1), geese (for GEE2)* functions in *R* and the *qif* macro in SAS we could not ascertain which of the estimators (i.e., equation

(2.1) or (2.7)) that could have been used in computing their ICC values. However, it is more likely that the population average models could be using equation (2.7) or a method similar to it, which could be the reason why negative ICC estimates were obtained. From a sample survey perspective, sampling error due to finite sample cluster size compared to the population cluster size, which is assumed to be infinite, could be the cause of negative ICC estimates (Eldridge, Ukoumunne and Carlin, 2009). Another reason is when there are large discrepancies in the allotment of trial resources within the clusters, this would cause large variations in the observed outcomes (Campbell and Walters, 2014a).

Results from this Chapter showed that estimates for the intervention effect, its SE, P-value, and 95% CI were the same for GEE1 and GEE2 models in almost all cases, they only differ in their estimates for the ICC. This means that in situations where only the mean structure is of primary interest, both methods fit the same models regardless of whether the correlation parameter was modelled or considered as a nuisance within the methods formulations, however, in GEE2 models the ICC parameter is explicitly modelled which could be a recourse to producing a more consistent ICC estimate compared to GEE1(Yan and Fine, 2004; Crespi, Wong and Mishra, 2009).

If the observed ICC is expected to be large, it is recommended that models that allow for heterogenous correlation structure should be considered, such as GEE2, because it is likely to improve inferences (Crespi, Wong and Mishra, 2009). This happens to be the major merit of Yan & Fines' 3EE GEE2 model over GEE1(Yan and Fine, 2004). The study of Balemi and Lee (2009) supported the results regarding GEE1 and GEE2. They used asymptotic expansions to show that the complexity introduced in the formulation of GEE2 by specifying separate estimating equations to model the association structure does not necessarily improve inference, especially when the mean parameters are of primary interest (Balemi and Lee, 2009). This result is consistent with that of Liang, Zeger and Qaqish (1992) that used simulations to show that GEE1 is adequate for estimating mean parameters when they are of interest, but they recommended the use of GEE2 when the association parameter (α) is also of primary interest and/or the number of clusters in small. In this current study, the performance of the statistical methods with respect to the mean parameters is of primary interest, hence, GEE2 was dropped from further investigations since it does not have any special benefit compared to the standard GEE1, at least in these cases being studied.

The four case studies present some key features of the cRCT design. The impact of these key features on the estimates from the four statistical methods is evident in the results obtained. For example, the PoNDER trial was conducted in a primary care setting and hence had a large sample size (both in the number of clusters and cluster sizes, 100 clusters with an average cluster size of 26). Hence, the unadjusted and adjusted estimates of the intervention effect from the four different methods are slightly different for the continuous outcome but were the same for the binary outcome. The odds ratios obtained possibly show the noncollapsible feature of the logistic regression model (with a logit link) – where including a baseline covariate changes the size of the estimate of the intervention effect, if the covariate is a strong predictor of the outcome, even if it is not related to the treatment conditions (Daniel, Zhang and Farewell, 2021). Since in this case the estimated intervention effect did not change upon inclusion of the covariates in the adjusted analysis, it would indicate that the covariates in the adjusted logistic model are not related to the outcome.

On the aspect of hypothesis testing, the conclusions were the same using any of the four methods which were consistent with that of the original analysis; a significant benefit of training health visitors to adequately manage women with postnatal depressive symptoms (i.e., favouring the intervention arm) (Morrell *et al.*, 2009). The ICC estimates were small; ICCs from the marginal logistic models were negative. This result is consistent with the findings of Adams et al. (2004), who reanalyse datasets from thirty-one cRCTs conducted within primary care settings and provided ICC estimates for several common variables, their median unadjusted ICC was 0.01 while the adjusted was 0.005. These results for the intervention effect estimate conformed to that of previous simulation studies, these studies found that both cluster-specific models (GzLMM) and population average models (e.g., GEE1) produced similar results for cRCTs that have many clusters and small ICC with binary (Heo and Leon, 2005) or continuous outcomes (Leyrat *et al.*, 2018). Hence, for large trials with weak correlation among outcomes, any of the four models – GzLMM, GEE1, GEE2, and QIF is optimal. Therefore, the choice of which model to use would be based on other factors like the aim of the research.

Bridging the Age Gap trial had a moderate sample size (43 clusters with an average cluster size of 18), and small ICC estimates. When the ICC estimate was negative, the estimates of the intervention effect were also negative across the population average models. A negative ICC is difficult to interpret since by the definition of the ICC it should be constrained between 0 and 1. This implies that negative ICC could be assumed to indicate no clustering among outcomes within

clusters, however, it is still preferable to use methods that account for clustering within clusters. Across the four methods, the unadjusted intervention effect estimates were unstable ranging from -0.28 to 0.12 but became stable (mean difference = 1.78) after the baseline outcome covariate was adjusted for, except for QIF (1.46) which also had the smallest SE estimates. This elucidates the importance of accounting for relevant prognostic factors in clinical trials, especially baseline outcome values (Samsa and Neely, 2018).

This was similar for the SEs and the 95% CIs. QIF appears to be slightly more precise than the other methods (i.e., had smaller SEs). However, this result should be interpreted with caution since the estimate of its intervention effect could be biased – methods that are biased toward the null hypothesis often tend to have smaller SEs (Morris, White and Crowther, 2019). Studies have confirmed the possibility of QIF producing biased estimates of the SE for trials with small to moderate clusters (Westgate, 2012; Westgate and Braun, 2012, 2013). Similarly, studies have found that the GzLMM with parameters estimated by REML performs better than GEE1 in maintaining the nominal Type I error rate and power, for continuous (Leyrat *et al.*, 2018) and binary outcomes (Thompson *et al.*, 2022) when the number of clusters is moderate or small. However, all four methods resulted in the same inference which is consistent with that of the original analysis which was "no significant difference in the Global QoL between the control and the intervention arms" (Wyld *et al.*, 2021).

The Informed Choice trial had a few clusters (10 clusters) with a large average cluster size (median cluster size = 145). The original study was a cross-sectional repeated measurement, so the estimate for the intervention effect was the interaction effect term between the treatment group (*group*) and time of measurement (*time*). But for illustration, we used only the "after intervention" postnatal sample. Only individual-level covariates were included in the adjusted models. Three of the methods produced approximately the same estimates which differed from that of QIF, for both continuous and binary outcomes. The most obvious difference lies in the P-values and CIs. For the continuous outcome, the adjusted P-value from GEE1 (and the other population average models) was significant whereas that of the GzLMM was not (**Table 7.5**). This could indicate that the impact of the small number of clusters affected the population average models more than the GzLMM (with identity link and parameters estimated by MLE).

For binary outcome, the unadjusted and adjusted P-values from QIF were significant but that of the other three methods were not. The QIF CI estimates were also narrower than that of the other

methods. This is indicative of a possible inflated test size, and bias in the estimated intervention effect which is consistent with the findings of previous studies by Westgate (2012), and Westgate and Braun (2013). The impact of the interplay between the small number of clusters, covariates, and cluster size imbalance on QIF and GEE1 has been studied. It was found that the QIF was severely affected compared to the GEE1(Westgate and Braun, 2012). A correction was proposed to improve the empirically estimated covariance matrix that causes the QIF to be poorly behaved (Westgate and Braun, 2013). Also, GzLMM was found to perform better than GEE1 in maintaining the nominal Type I error and power in trials with few clusters (≤ 20) for both continuous (Leyrat *et al.*, 2018) and binary outcomes (Thompson *et al.*, 2022). These current results confirmed those previous findings; however, it is more likely that the differing performance of the QIF estimator is due to the small number of clusters. Given these findings, it is likely that the QIF is severely affected by small to moderate numbers of clusters, followed by GEE1 then GzLMM. Although, no simulation study has been carried out to compare these three methods in this regard, to reach a definite conclusion.

Small sample corrections were applied to estimates from GzLMM and GEE1 only. Although there are recommended corrections for GEE2 (Zhang *et al.*, 2023) and QIF (Westgate, 2012), however, they are not available or easy to implement in standard statistical packages, respectively. The corrected results showed changes in the P-values and the 95% CI of the treatment estimates, for binary and continuous outcomes. The corrected P-values are bigger, and the CIs are wider. These findings are consistent with that of previous studies (McNeish *et al.*, 2016; Leyrat *et al.*, 2018; Thompson *et al.*, 2022).

Lastly, for the NOSH trial with only binary primary outcome measured and large sample sizes (92 clusters with an average cluster size of 100). The parameter estimates from the four statistical approaches are approximately the same, hence, their performance was equivalent. A unique finding is that it is only in this case study that the adjusted intervention effect estimated from GEE2 and GEE1 were different (1.27 vs. 1.31) with SEs of 0.05 vs. 0.07, consequently, their 95% CIs were different. The key feature of the NOSH trial which is different from other case studies is that in NOSH only cluster-level covariates were adjusted for, maybe this feature had a differing impact on the GEE1 and GEE2. Further studies are needed to confirm this.

These results depict some patterns in the behaviour of the analytical methods, for example, QIF behave differently when the number of clusters is small or moderate. However, due to the

limitations of using these four case studies it would be unscientific to reach an explicit conclusion regarding the superiority of the methods. These four real-world datasets present limited scenarios that could arise in the design of cRCTs, for example, the Informed Choice trial (O'Cathain *et al.*, 2002) present a scenario with small number of clusters (5 clusters per arm), large average cluster sizes (155 participants), with binary and continuous outcomes measured, and small missing outcome data (7%). These features affected QIF differently – QIF produced smaller estimates for the intervention effect and it's SE, especially the small number of clusters (Westgate, 2012; Westgate and Braun, 2012, 2013). Hence, it would be a hasty conclusion to base the superiority of the methods only on just these scenarios, because, these methods behave different in different scenarios, and several design scenarios are possible in clinical research. Furthermore, another major limitation is that all four cases presented a similar study design – two-arm parallel group cRCT design with primary care and community settings. It is likely that the methods would behave differently when applied to datasets from other study designs like crossover cRCT or Factorial cRCT designs with other settings like care homes. Lastly, the four datasets measured only continuous and binary outcomes. The methods compared are adequate for count outcome data and are likely to behave differently.

In the light of the abovementioned limitations of the real datasets used, the conduct of simulation studies to investigate the operating characteristics of these four methods is imperative. Simulation studies involve generating pseudo-random numbers from computer-designed experiments that mimic different settings of CRCT design (Morris, White and Crowther, 2019). A simulation study involving the combinations of different numbers of clusters, levels of ICC, effect sizes (i.e., the true intervention effect), cluster sizes, types of outcomes, and distribution of the cluster random effects would be ideal for a comprehensive investigation. This will help create different scenarios that are needed to investigate the sole and combined impact of the varied factors on the operating characteristics of the methods. Another possible simulation study that is like the one stated above, but with a focus on extremely small numbers of clusters ($\leq$ 30 clusters), and the methods would include the uncorrected and corrected versions (adjustment for small sample size) of the four methods. This study will determine how well the corrected versions of the methods perform both absolutely and relatively.

7.10 Summary

After describing the traditional methods for analysing outcome data from cRCTs in Chapter 2, a methodological scoping review was conducted in Chapter 3 to identify more available methods in the literature. Twenty-seven unique methods were identified, with the traditional methods included. A review of what statistical methods is used in practice to analyse the outcome from cRCTs in Chapter 4 showed that none of the two identified alternatives to GEE1, GEE2, and QIF were not used in the analysis of the primary outcomes of all the articles included. One plausible reason could be that the advantages of these newer methods have not been comprehensively explored in the context of cRCTs to warrant their routine applications. Based on these findings from the reviews, four statistical methods were selected for further comparative evaluations. The technical descriptions of these methods are available in Chapter 6, while this current Chapter 7 reports the results of applying the four methods to four datasets from cRCTs. Each of these four case studies has its unique feature. Since one of the objectives of this thesis is to check whether the known theoretical behaviours of these methods conform to their behaviours when applied to real-world data, hence this empirical analysis was necessary. To this end, QIF produces different estimates of the treatment effect with small or moderate numbers of clusters and negative ICC. GEE2 on the other hand, produced same parameter estimates as GEE1. Previous studies using asymptotic expansions (Balemi and Lee, 2009) and simulations (Liang, Zeger and Qaqish, 1992) have shown that the extra complexity of GEE2 does not improve inference, especially when only the mean parameters are of primary interest. Consequently, the GEE2 was dropped from further investigation since this current research focusses on the mean parameters.

Although the empirical analyses reported in this chapter have shown that the four methods can be applied to the outcomes from cRCTs. It is difficult to compare the statistical performance of the four methods, in terms of say bias in the estimation of the treatment effect, precision of the estimators of the treatment effect, coverage of its confidence interval, and Type I and Type II errors without knowing the true treatment effect.

To quote Burton et al. (2006)

"Simulation studies use computer-intensive procedures to assess the performance of a variety of statistical methods concerning a known truth. Such evaluation cannot be achieved with studies of real data alone."

Therefore, the next Chapter 8 presents the protocol for a computer simulation study to further investigate the operating characteristics of the methods in terms of bias in the estimated treatment effect, precision of the treatment effect, coverage of the confidence interval, and Type I and Type II errors comprehensively. The simulation study would generate outcome data covering more scenarios that the four real-world data used in this current chapter did not have. In Chapter 9, the results of the simulation study are presented, and a comprehensive discussion regarding all the findings in the thesis is provided in Chapter 10.

Chapter 8

Comparison of statistical methods for analysing continuous outcome data from cRCTs: protocol of a simulation study

8.1 Introduction

Multilevel statistical models that account for clustering are recommended for individual-level analysis in cRCTs. The methodological review of Chapter 3 identified GzLMM (with parameters estimated by MLE) of CSM and GEE1 of PAM as the most studied methods, while QIF was identified as the most studied new method. The practice review of Chapter 4 also identified GzLMM and GEE1 as the two most used models in practice (Offorha, Walters and Jacques, 2022). The superiority of QIF over GEE1, when the correlation structure is misspecified, has been established in the context of longitudinal studies (Qu, Lindsay and Bing, 2000; Odueyungbo *et al.*, 2008). Some papers have compared the finite sample size performance of QIF to GEE1 in the context of cRCTs with a focus on the efficiency of the methods (Westgate, 2012; Westgate and Braun, 2012, 2013).

Nonetheless, more studies are required to comprehensively evaluate the performance of these acclaimed alternatives against that of the most used methods in cRCTs (e.g., GzLMM and GEE1). For example, to the best of my knowledge, except for a publication on the analysis of real-world cRCT datasets stemming from this current thesis, no other study has compared a CSM (like GzLMM) to QIF (a PAM) in the context of cRCTs. The empirical analysis conducted in Chapter 7, using real-world data did not conclude on the superiority of the methods since the truth is not known, but some obvious patterns exist concerning the behaviour of QIF. For instance, QIF behaved differently when the number of clusters was small or moderate compared to GzLMM, GEE1, and GEE2. There is a recent simulation study comparing QIF to GEE1 in the context of cRCTs, but it was not comprehensive – used only fixed values for the parameters, one value for number of clusters (100 clusters), and single value for cluster size (25 per cluster) (Yu, Li and Turner, 2020). Though Westgate and Braun used parameters with varying level in their simulation

studies, they used only performance measures that assessed the efficiency (smaller SE) of the methods (Westgate, 2012; Westgate and Braun, 2013).

This current research attempts to fill in the gaps identified in Chapters 3 and 4, that led to the research questions, aim, and objectives in Chapter 5. Precisely, the purpose of this current study is to conduct a comprehensive simulation study to evaluate the performances of GzLMM (based on an identity link function and continuous outcome) with parameters estimated by MLE, GEE1, and QIF with compound symmetry/exchangeable working correlation structures and robust SEs for both, in the context of cRCTs with primary care and community settings.

8.2 Chapter aim

As stated above, the primary aim of this chapter is to investigate the performance of the three selected methods (GzLMM, GEE1, and QIF) in the context of cRCTs. The specific objectives to explore are:

1. To evaluate and compare the impact of varying the number of clusters on the consistency, efficiency, and power of the statistical methods for analysing continuous outcome data from ideal (balanced) cRCTs.

2. To evaluate and compare the impact of varying the average cluster size on the consistency, accuracy, efficiency, and power of the statistical methods for analysing outcome data from ideal (balanced) cRCTs with continuous outcomes.

3. To evaluate and compare the impact of varying the ICC on the consistency, accuracy, efficiency, and power of the statistical methods for analysing outcome data from ideal (balanced) cRCTs with continuous outcomes.

4. To evaluate and compare the impact of varying the effect size on the consistency, accuracy, efficiency, and power of the statistical methods for analysing outcome data from ideal (balanced) cRCTs with continuous outcomes.

8.3 Study design

This study is a completely randomised two parallel treatment arms cRCT. This involves the combination of the different levels of the parameters chosen to generate the data generating mechanisms (DGMs). The parameters are described below.

8.3.1 Fixed parameters

Two treatment arms (intervention and control) are assumed for the comparison. The three statistical methods being evaluated are GzLMM (specified by an identity link function) with parameters estimated by MLE, GEE1, and QIF. An exchangeable working correlation structure was used for GEE1 and QIF. Lastly, the number of simulations for each combination (DGM) is 4000.

- **Covariate**

Group (x_{1i}): This is the cluster-level treatment arm indicator variable for the i^{th} cluster, where $x_{1i} = 0$ indicates the control arm and $x_{1i} = 1$ the intervention arm. This is the only covariate included in the models.

8.3.2 Varying parameters

In each run some of the parameters described in this Section will take a different value from the previous run, but within a pre-specified range of values and this will continue till all the levels of each of the parameter is utilised. The combination of these values will give rise to the different DGMs/scenarios.

- **Average cluster size (n_i)**

The average cluster size chosen ranges from 10 to 250 participants per cluster. Specifically, they are 10, 20, 50, 100, 150, and 250 subjects per cluster. These values were chosen to be close to the observed percentiles of the distribution of average cluster size analysed, as depicted in the review of publicly funded cRCTs in UK (Chapter 4; **Table 4.4**). Specifically, from the results of the practice review the 25th, 50th, 75th and 90th percentiles are 10, 21, 71 and 246 participants

respectively. These values also support gradual increase in the parameter values with the intent to provide a robust coverage of the range (i.e., 10 to 250 average participants per cluster).

Additionally, some of these chosen values (i.e., 10, 20, 100 and 150 participants) are also close to the average number of subjects randomised (i.e., 18, 27, 100 and 155 participants) in the four cRCT case studies that were analysed in Chapter 7 (see, **Table 7.2**).

- **Number of clusters (*N*)**

The numbers of independent clusters were chosen to reflect what has been observed in practice (Chapter 4). Offorha, Walters and Jacques (2022) found that the 1st and 3rd quartiles of the numbers of clusters randomised in practice were 25 and 74 with a median of 44 clusters. The most common numbers of clusters randomised in cRCTs fell within these groups: [11 to 20], [21 to 50], and [51 to 100] representing 13%, 47%, and 24% of the total randomised clusters, respectively.

Specifically, from the results of the review of publicly funded cRCTs in the UK (Chapter 4; **Table 4.4**), the minimum number of clusters randomised/analysed was 7, the 25th percentile of the cumulative distribution of numbers of clusters randomised (and analysed) was 25 clusters (25 clusters), 50th percentile was 44 clusters (43 clusters), 75th percentile was 74 clusters (69 clusters), 90th percentile was 123 clusters (121 clusters). Hence, these values 10, 20, 40, 50, and 120 total independent numbers of clusters (i.e., N/2 per treatment arm) were chosen to be close to these percentiles for conducting the simulation study in Chapter 9. Additionally, these values are like those observed in the empirical analysis of datasets from the four cRCTs with small (10 clusters), moderate (43 clusters), and large (100 clusters) numbers of clusters conducted in Chapter 7.

In summary, the total number of clusters to be randomised ranges from 10 to 120 covering the range of the number of clusters (90th percentile) used in most cRCTs conducted in the UK with primary care and community based trials as informed by the results from the review of the NIHR Journal Library in Chapter 4 (**Table 4.4**).

- **Intracluster correlation coefficient (ICC)**

Four ICC ρ values were selected to represent four different degrees of clustering commonly observed in cRCTs with primary care and community settings as indicated by the results of the review of publicly funded cRCTs in the UK conducted in Chapter 4 (**Table 4.4**). The results of

Table 4.4 shows that the 25th, 50th, 75th, 90th and 95th percentiles of the cumulative distribution of observed ICCs were 0.001, 0.02, 0.06, 0.22 and 0.23. Consequently, ICC values ($\rho = 0.001$, 0.01, 0.05, and 0.25) that are close to these percentiles were chosen and used in the simulation study in Chapter 9. For example, $\rho = 0.25$ is the 96.66th percentile of the cumulative distribution of observed ICCs from the results of the review of publicly funded cRCTs conducted within the UK. Hence, these chosen range of ICCs cover those commonly (about 97%) found in publicly funded cRCTs conducted in the UK (Chapter 4, **Table 4.4**)

- **True intervention effect/effect size** (θ)

The effect size θ is specified by β_1 in the regression models, and its range is given as $\beta_1 = \{0, 0.2, 0.3\}$. $\beta_1 = 0$ will aid with the investigation of the ability of the selected statistical methods to control the Type 1 error rate under the null hypothesis, while 0.2 and 0.3 correspond to the 1st quartile and the median standardised target effect sizes that are commonly encountered in RCTs (Rothwell, Julious and Cooper, 2018). Rothwell, Julious and Cooper (2018) found that most RCT studies arrive at the choice of effect size through the review of literature and from previous studies. The results from their review showed that the median standardised target effect size was 0.30 (IQR, 0.20 – 0.38). Therefore, the values of effect sizes ($\theta = 0$, 0.2, 0.3) used in the conduct of the simulation study in Chapter 9 was chosen to reflect these quantiles.

8.4 Number of simulations/repetitions

To determine the number of simulations required for each scenario the primary outcomes will be the Monte Carlo SE of the estimated coverage of the 95% confidence interval for the treatment effect θ, Type I error rate, and Power.

For example, assuming the point estimate of the coverage probability for the 95% confidence interval for the true intervention effect θ, from the simulations is 0.95; then with n_{sim} = 4000 simulations, the Monte Carlo SE (MCSE) of the estimate will be 0.0034 and an approximate 95% CI for this parameter would range from 0.943 to 0.957. This level of precision should be enough to compare the different statistical models. Morris, White and Crowther (2019) provided a formular for the above calculation with regards to the MCSE of coverage probability given in (8.1) below.

$$MCSE_{CPR} = \sqrt{\frac{\widehat{CPr} \times (1 - \widehat{CPr})}{\mathrm{n_{sim}}}} \tag{8.1}$$

Using (8.1) the generated distribution for $\mathrm{n_{sim}}$ and MCSE of the CPr for the 95% CI of $\hat{\theta}$ is presented in **Table 8.1**. It could be observed from **Table 8.1** that $\mathrm{n_{sim}} = 4000$ (shaded row) produced the optimal MCSE of the CPr for the 95% CI of $\hat{\theta}$.

Table 8.1 MCSE of the estimated 95% CI coverage of θ determined by the number of simulations

				Approximate 95% CI	
$\mathrm{n_{sim}}$	**Coverage estimate**	**1 – coverage**	**MCSE of estimate**	**−2SE**	**+2SE**
1000	0.95	0.05	0.0069	0.936	0.964
2000	0.95	0.05	0.0049	0.940	0.960
3000	0.95	0.05	0.0040	0.942	0.958
4000	0.95	0.05	0.0034	0.943	0.957
5000	0.95	0.05	0.0031	0.944	0.956
6000	0.95	0.05	0.0028	0.945	0.956
7000	0.95	0.05	0.0026	0.945	0.955
8000	0.95	0.05	0.0024	0.945	0.955
9000	0.95	0.05	0.0023	0.945	0.955
10000	0.95	0.05	0.0022	0.946	0.954

Similarly, if the Type I error rate is to be considered in determining the number of simulations $\mathrm{n_{sim}}$, then the number of simulations $\mathrm{n_{sim}}$ required to maintain the desired MCSE for the Type I error rate can be derived from the MCSE for the Type I error rate, given as

$$MCSE_{Type\ I} = \sqrt{\frac{\widehat{Type\ I} \times (1 - \widehat{Type\ I})}{\mathrm{n_{sim}}}} \tag{8.2}$$

Table 8.2 MCSE of the estimated Type error I determined by the number of simulations

n_{sim}	Type I error estimate	1 – Type I error	MCSE of estimate	Approximate 95% CI −2SE	Approximate 95% CI +2SE
1000	0.05	0.95	0.00005	0.04991	0.05010
2000	0.05	0.95	0.00002	0.04995	0.05005
3000	0.05	0.95	0.00002	0.04997	0.05003
4000	0.05	0.95	0.00001	0.04998	0.05002
5000	0.05	0.95	0.00001	0.04998	0.05002
6000	0.05	0.95	0.00001	0.04998	0.05002
7000	0.05	0.95	0.00001	0.04999	0.05001
8000	0.05	0.95	0.00001	0.04999	0.05001
9000	0.05	0.95	0.00001	0.04999	0.05001
10000	0.05	0.95	0.00000	0.04999	0.05001

Table 8.2 was generated similarly to **Table 8.1**. With $n_{sim} = 4000$ the MCSE of the Type I error rate seem to be optimal since above that value the MCSE did not improve further. **Table 8.3** was generated in the same manner as **Table 8.1** and **Table 8.2** when MCSE of Power is considered in determining the optimal number of simulations n_{sim}. Using (8.2) with $\widehat{Type\ I}$ substituted with $\widehat{Power}$, as seen before, $n_{sim} = 4000$ seems to result in a reasonable MCSE for Power (shaded row).

Table 8.3 MCSE of the estimated Power determined by the number of simulations

n_{sim}	Power	Type II error = 1 – Power	MCSE of estimate	Approximate 95% CI −2SE	Approximate 95% CI +2SE
1000	0.90	0.10	0.00009	0.8998	0.9002
2000	0.90	0.10	0.00005	0.8999	0.9001
3000	0.90	0.10	0.00003	0.8999	0.9001
4000	0.90	0.10	0.00002	0.9000	0.9000
5000	0.90	0.10	0.00002	0.9000	0.9000
6000	0.90	0.10	0.00002	0.9000	0.9000
7000	0.90	0.10	0.00001	0.9000	0.9000
8000	0.90	0.10	0.00001	0.9000	0.9000
9000	0.90	0.10	0.00001	0.9000	0.9000
10000	0.90	0.10	0.00001	0.9000	0.9000

8.5 Scenarios investigated

This simulation study was designed to depict an ideal cRCT where the number of clusters and ICC are balanced between the treatment arms, and the average cluster size is the same across all clusters. This design is not uncommon among primary care and community based cRCTs.

8.6 Data Generation Mechanism (DGM)

The partial factorial combinations of the different 4 levels of β_1, 5 of N, 6 of n_i and 3 of ρ give a total of 120 possible scenarios to be investigated (**Table 8.4**). For each effect size only two specific levels of n_i are combined with N and ρ, instead of the total 6 levels, hence, this study design is not a full factorially designed simulation study.

Table 8.4 Input varying parameters for the data generating mechanisms (120 scenarios)

Parameter	Symbol	Values	(n)	Source
True intervention effect	β_1	0, 0.2, 0.3	3	(Rothwell, Julious and Cooper, 2018)
Number of clusters	N	10, 20, 40, 50, 120	5	Chapter 4; Table 4.4
Average cluster size	n_i	10, 20, 50, 100, 150, 250	6	Table 4.4; Table 7.2
ICC	ρ	0.001, 0.01, 0.05, 0.25	4	Chapter 4; Table 4.4

For each of the 120 DGMs, 4000 datasets are generated using the data generating model of equation (8.3) below. The values of the input parameters are presented in **Table 8.4**, each chosen single value of the true intervention effect β_1 and that of the number of clusters N will be combined differently into two values of the average cluster size n_i and three values of the ICC ρ to form the DGMs as presented in **Figure 8.1**. Specifically, each branch of the tree diagram of **Figure 8.1** forms a single DGM. The complete simulation process of a single DGM is presented in **Figure 8.2**.

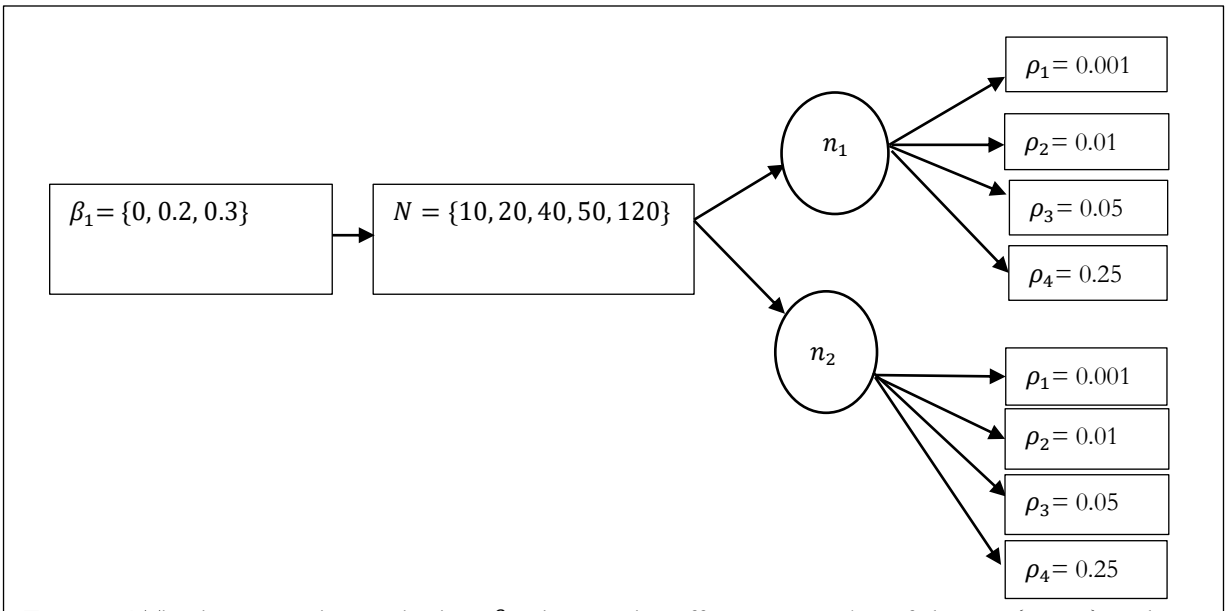


Figure 8.1 The data generating mechanism, β_1= intervention effect; N = number of clusters; $\{n_1, n_2\}$ = cluster sizes; $\{\rho_1, \rho_2, \rho_3, \rho_4\}$ = range for correlation parameter.

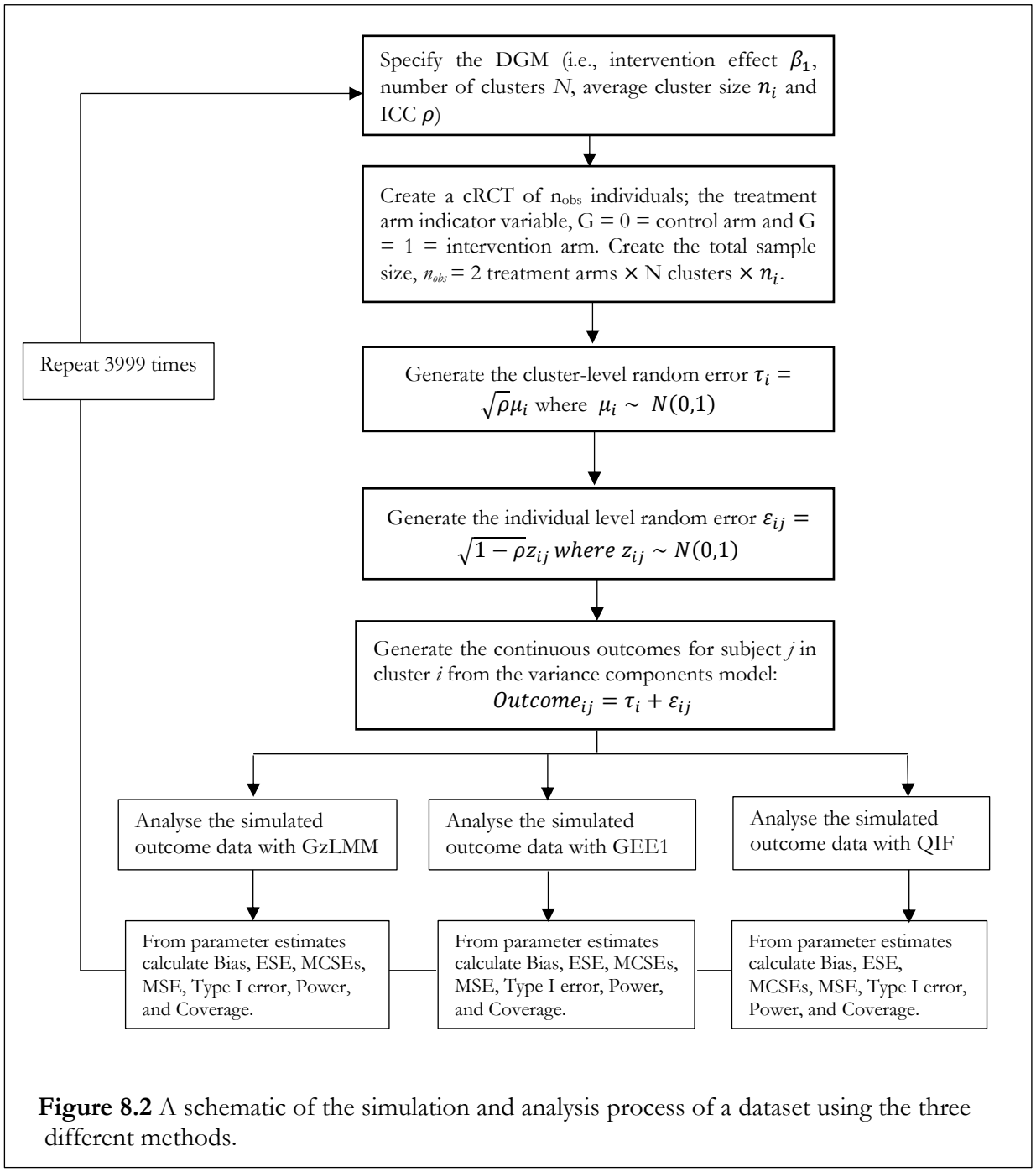


Figure 8.2 A schematic of the simulation and analysis process of a dataset using the three different methods.

8.7 Data generating model

The continuous outcome data to be analysed are generated from the random intercept model of (8.3) below, (8.3) also called the variance component model (Campbell and Walters, 2014a).

$$Outcome_{ij} = \beta_1 x_{1i} + \tau_i + \varepsilon_{ij} \tag{8.3}$$

where $i = 1, \ldots, N$ is the independent clusters, $j = 1, \ldots, n_i$ is the average cluster size of the i^{th} cluster, β_1 is the intervention effect to be estimated, x_{1i} is the treatment arm indicator variable, τ_i is the cluster-level random effects and ε_{ij} the individual-level random effects (**Figure 8.1**).

8.8 Statistical methods

The selection of the three statistical methods was inspired by the findings of the reviews of Chapters 3 and 4, and confirmed by published reviews (Twardella, Bruckner and Blettner, 2005; Turner, 2017; Offorha, Walters and Jacques, 2022). These methods have been described in Chapters 2 and 6, briefly, they are:

8.8.1 GzLMM

The GzLMM (with an identity link function and parameters estimated by MLE) is a popular example of a CSM for analysing continuous outcome data from cRCTs and has been described in Sections 6.5 and 6.6. The MLE as an estimator has been described in Section 6.6. Briefly, GzLMM (with an identity link function applied to the continuous outcome) is a one-stage model that uses a single mean model equation to account for both the fixed effects from covariates and random effects from clusters (and individuals) in the observed outcomes. As a level two model, the individual subjects are nested within the independent clusters. The cluster unit is included in the model as a random intercept to account for clustering among outcomes in a particular cluster which causes the cluster means to vary.

8.8.2 GEE1

As a popular example of a PAM, it has been described in Section 6.4.1. Briefly, GEE1 is a semi-parametric model since it does not specify the distribution of the cluster random effects. The GEE1 is a two-stage model because it uses two separate equations; one equation for modelling the mean parameters and the other to describe the correlation among the outcomes in a cluster. Liang and Zeger (1986) proposed a GEE1 that uses a working covariance structure with few nuisance parameters to model the true correlation of outcomes within clusters. GEE1 produces consistent parameter estimates regardless of whether the working correlation structure is correct or not but provided that the mean structure is correct. But, if the true correlation structure is wrongly specified, there is a possibility of some loss in efficiency.

8.8.3 QIF

QIF has been described in Section 6.4.3. It is a newer acclaimed alternative to GEE1 for estimating the parameters of a PAM. Qu, Lindsay and Bing (2000) proposed QIF to improve the efficiency of GEE1 when the correlation structure is misspecified, especially when the correlation among outcomes is substantial. First, this is because QIF avoids the direct use of a correlation matrix instead it uses basis matrices and some constants, secondly QIF penalises the contribution of each cluster; clusters with large variances are given less weight than the ones with small variances. QIF has other advantages over GEE1 such as robustness to outliers, produces a goodness-of-fit test statistic that follows a chi-square distribution like Akaike information criteria/Bayesian information criterion (AIC/BIC) and it does not suffer multiple roots problems as GEE1 (Qu, Lindsay and Bing, 2000; Odueyungbo *et al.*, 2008; Song *et al.*, 2009).

8.9 Performance measures

This Section presents the numerical measures used to quantify the performances of the methods. The performance measures are:

8.9.1 Convergence rate

This is the proportion of the modelling process that fails to produce results, that is, they did not converge to definite solutions. This might also highlight situations in which the DGM might not be plausible in practice.

$$Covergence\ rate\ (CR) = \frac{\#number\ not\ converging}{n_{sim}} \tag{8.4}$$

8.9.2 Bias

This is the deviation of the estimated intervention effect $\hat{\theta}$ from the true known intervention effect θ, on the average; $E[(\hat{\theta}) - \theta]$. The computational formula for the estimate of bias is given as

$$Bias = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \theta)$$

(8.5)

8.9.3 Empirical SE (ESE)

The empirical SE quantifies the uncertainty of the statistical methods. Methods with larger ESE have less precision or efficiency in the long run and vice versa; $\sqrt{Var(\hat{\theta})}$. The estimate formula is given as

$$ESE = \sqrt{\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \bar{\theta})^2}$$

(8.6)

8.9.4 Mean square error (MSE)

While the ESE focuses on the uncertainty of the estimator of θ, the MSE measures both the bias and precision of the estimator to provide a more robust evaluation. The sum of square bias and variance of $\hat{\theta}$ over the number of simulations is the MSE: $E\left[\left(\hat{\theta} - \theta\right)^2\right]$. The formular for MSE is given as

$$MSE = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \theta)^2$$

(8.7)

8.9.5 Coverage probability (CPr)

This is the proportion of times the nominal 95% CI of $\hat{\theta}$ from a method contains θ: $\Pr(\hat{\theta}_{ll} \leq \theta \leq \hat{\theta}_{ul})$. The computational formular is given as

$$CPr = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1(\hat{\theta}_{ll,i} \leq \theta \leq \hat{\theta}_{ul,i})$$

(8.8)

8.9.6 Type I error rate/Power

Type I error rate is the proportion of the P-values of the estimate of the intervention effect $\hat{\theta}$ that is ≤ 5% level of significance under the null hypothesis. An appropriate method chosen to analyse outcome data from cRCTs should have the capability to control the Type I error rate. Power on the other hand is the proportion of the P-values of the intervention effect that is ≤ 5% level of significance under the alternative hypothesis: $\Pr(p_i \leq \alpha)$. The computational formular is given as

$$Type\ I/Power = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1(p_i \leq \alpha) \tag{8.9}$$

8.10 Software

The statistical platform to be used for this simulation study is R version 4.2.1, which would be used to prepare the simulation script to be run on a high-performance computing (HPC) hub because of the high amount of computer resources needed. The following packages (versions) were installed to be used: afex (1.2.0), dplyr (1.0.10), tidyverse (1.3.2), rsimsum (0.11.3), qif (1.5), geepack (1.3.9), lme4 (1.1-31), broom (1.0.1), purr (0.3.5), pacman (0.5.1), pupillometryR (0.0.4), CaTools (1.18.2), bitops (1.0-7), cowplot (1.1.1), readr (2.1.3), lavaan (0.6-12), blandr (0.5.1), ggplot2 (3.4.0).

8.11 Summary

This chapter presents the plan of a simulation study to assess the comparative performance of the three statistical methods – GzLMM, GEE1, and QIF. These methods have been evaluated using real-world continuous and binary outcomes data in Chapter 7. Some patterns were observed in the properties of the methods, for example, QIF was observed to produce differing results to the other three methods especially when the cRCT has small or moderate numbers of clusters. But it was not feasible to conclude the superiority of the methods using only real-world data, because a reference truth is not known. Also, the real-world data mostly present limited scenarios and hence would not be adequate for a fair comparison. This necessitated the conduct of this current simulation study, the results from the simulation study are presented in the next Chapter.

Chapter 9 presents the results from the analysis of the simulated continuous outcome cRCT data using GzLMM, GEE1, and QIF. In a simulation study, the true parameters are known, and several scenarios are generated to facilitate fair comparison among the methods of interest. The simulation study of Chapter 9 is adequate to reach definite conclusions on the superiority of the methods in different scenarios peculiar to primary care and community based cRCTs. Chapter 10 discusses the results, limitations, strengths, contributions, recommendations, and conclusions of this thesis. Issues raised by this current research are also stated for future explorations.

Chapter 9

Comparison of statistical methods for analysing continuous outcome data from cRCTs: a simulation study

9.1 Introduction

This Chapter attempts to address one of the objectives of this research stated in Chapter 5 – the evaluation of the performance of the selected statistical methods for analysing outcome data from cRCTs, based on their long-run statistical properties. This simulation study was planned and documented in Chapter 8. This current Chapter presents the results of implementing the simulation protocol of Chapter 8. Although the methods have been compared in Chapter 7 with continuous and binary real-word outcome data from four cRCTs, this simulation study is necessary for investigating the long-run properties of the analytical methods to ascertain their superiority in different scenarios of cRCT designs in which the truth is known (Burton *et al.*, 2006; Morris, White and Crowther, 2019). The findings, limitations, and conclusions of this study are further discussed in Chapter 10.

9.2 Chapter aim

The main aim of this chapter is to present the results of the simulation study detailed in Chapter 8, which was conducted to evaluate the performance of the selected methods, which are:

1. Generalized Linear mixed model (with identity link function) with coefficients estimated by maximum likelihood estimation (labelled as GzLMM).
2. Generalized Linear model with coefficients estimated by first-order generalized estimating equations (labelled as GEE1).
3. Generalized Linear model with coefficients estimated by quadratic inference function (labelled as QIF).

The performance measures used to assess the performance of the statistical methods (more fully described in Chapter 8) are convergence rate, bias, empirical standard error, mean square error, coverage probability, Type I error, and power.

9.3 Results from the simulation

The R codes for conducting the series of simulations and the analysis of the datasets are provided (**Appendix 5**). This chapter presents the results of a series of simulations from 120 data-generating mechanisms (DGMs) with 4000 simulations per DGM per method (i.e., 120 x 4000 x 3 = 1,440,000 simulated estimates per population parameter). Preliminary checks of all the 1,440,000 simulated intervention effect estimates $\hat{\theta}_i$ from the 480,000 datasets showed that they are well-behaved, the distribution is symmetrical and there are no significant outlying values (**Appendix 6**), a sample is shown in **Figure 9.1**.

9.3.1 Convergence rate

Each of the 120 DGMs/scenarios was run 4000 times for each of the three methods, totalling 1,440,000 simulated estimates of the intervention effect (and its associated parameters). The number of simulations that failed to converge to results was 36 which is 0.0025% ($36/1440000 \times 100$). Non-convergence occurred only when GzLMM was used for analysis. The highest proportion of non-convergence was 2 out of 4000 simulations, for each of 18th, 30th, 47th, and 105th DGMs. These DGMs/scenarios have a mix of small to moderate number of clusters ($N: 10-50$), small and moderate ICC ($\rho: 0.001, 0.05$), and all effect sizes considered ($\theta: 0, 0.2, 0.3$).

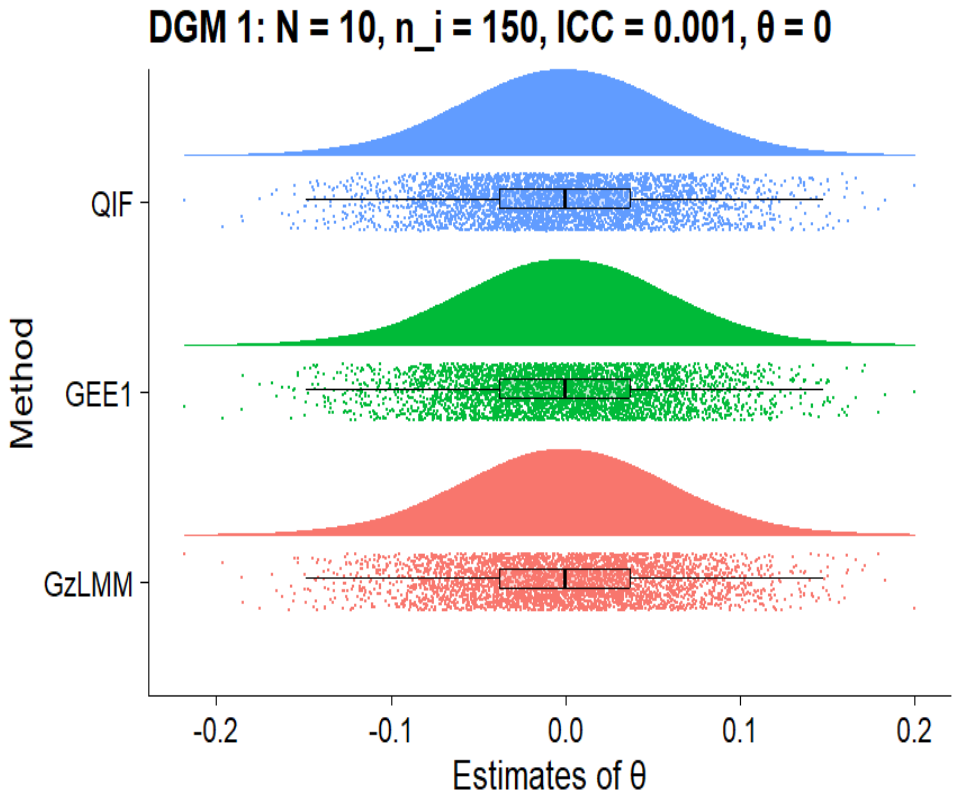


Figure 9.1 A raincloud plot showing the distribution (top density plot), the specific points (scatter plot) and the basic summary statistics (superimposed boxplot) of the simulated 4000 point estimates of the true intervention effect θ from each method for a single scenario, where N is the total number of clusters, n_i is the average cluster size and ICC is the intracluster correlation coefficient. The electronic version of the figure is in colour.

9.3.2 Bias

Bias was defined as the average difference between the true treatment effect and the treatment effect observed in the simulations per scenario per method. There was no substantial or meaningful bias from the methods across the 120 different scenarios, the absolute observed biases were small and close to zero (**Appendix 7**). The absolute maximum biases were $|0.0058|$, $|0.0035|$, and $|0.0051|$ for $\theta = 0$, 0.2, and 0.3 respectively. There were two scenarios (DGM, 7 and 26) where the confidence intervals (CIs) of the estimated bias from the 3 different statistical methods excluded 0 (**Appendix 7**). This was the scenario with the direct combinations of $\theta = \{0,0\}$, $N = \{10,50\}$, and $ICC = \{0.05,\ 0.01\}$ respectively. This shows that a high ICC could potentially lead to bias in the estimates of the treatment effect regardless of the statistical method being used. The bias in the estimate of the intervention effect from the methods was approximately the same and close to zero across most scenarios, and slightly higher when $\theta = 0.3$ compared to when $\theta = 0$ or 0.2 (**Figure 9.2**).

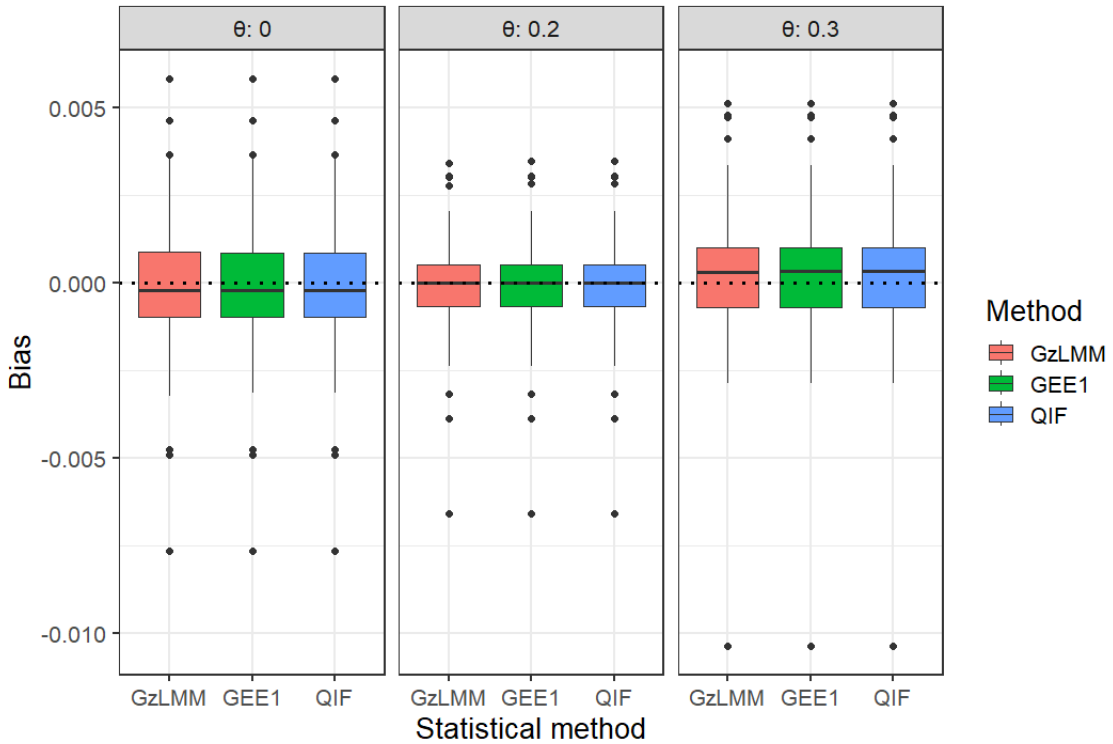


Figure 9.2 Box and whisker plots of the summary statistics of estimated bias of the estimates of the intervention effect from GzLMM, GEE1 and QIF by the true treatment effect θ ($\theta = 0$, 0.2, or 0.3).

Figure 9.3 shows that the estimated bias remained almost the same across the methods regardless of the number of clusters randomised and the size of the ICC, indicating that the relative consistency of the statistical methods was not affected by the number of clusters N and the level of clustering (ICC). It appears that as the ICC increases the biases of the methods deviate more from zero, but still within the acceptable range.

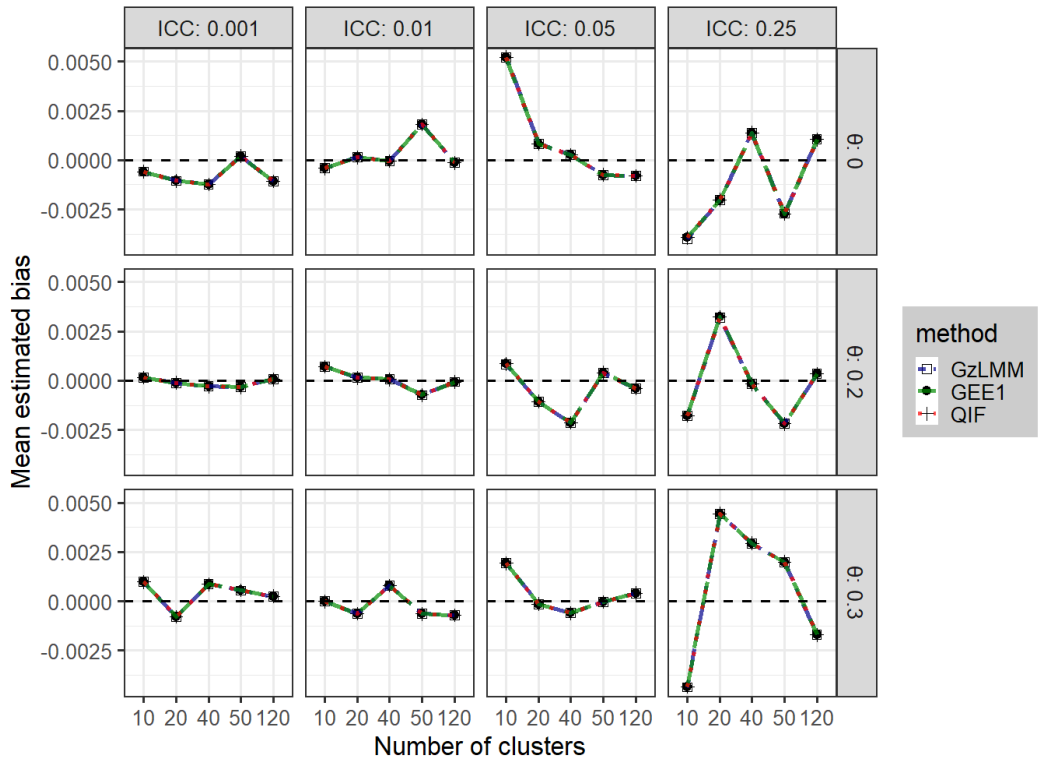


Figure 9.3 Mean estimated bias across all scenarios for the three statistical methods (GzLMM, GEE1, QIF) by three levels of theta (θ: 0, 0.2, and 0.3), four levels of ICC (ρ = 0.001, 0.01, 0.05 and 0.25), and several numbers of clusters (10, 20, 40, 50, and 120). Each data point is based on the estimates from the 480,000 simulated datasets. The electronic version is in colour.

9.3.3 Empirical standard error (ESE)

The ESE has been defined in detail in Section 8.9.3. Empirical SE is the standard deviation of the estimated treatment effects. It is a measure of the uncertainty of the estimates of the treatment effect obtained from the methods, in other words it measures the precision of the methods. Methods with larger ESE have less precision or efficiency in the long run and vice versa. The ESEs of the methods are similar in all 120 different scenarios (**Appendix 6**). Boxplots showing the ESEs across all scenarios for each method by the three different true effect estimates (0, 0.2 and 0.3) are

presented in **Figure 9.4**, they are almost the same regardless of the magnitude of the effect size (θ) and the median ESEs are < 0.1. Consequently, the relative efficiency of any pair of the methods is approximately one.

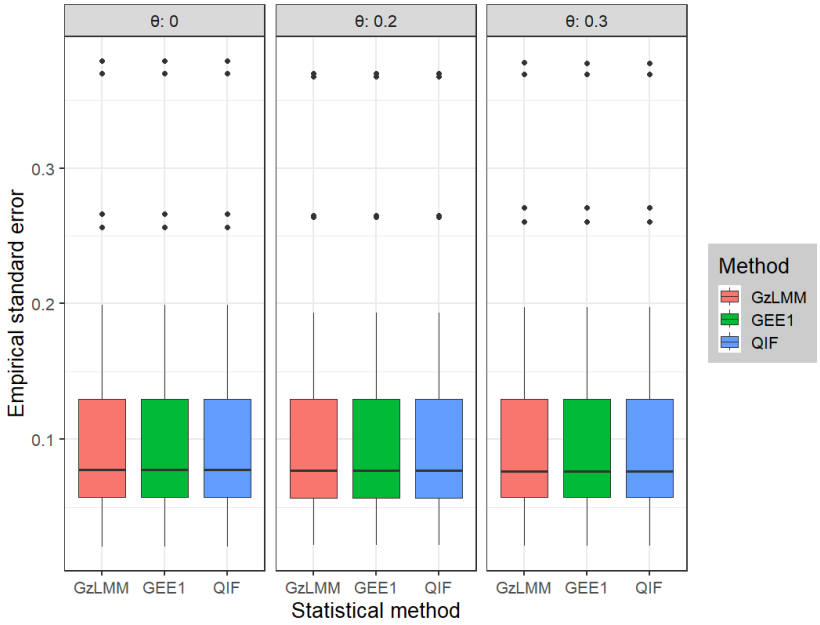


Figure 9.4 Box and whisker plots of the summary statistics of empirical standard error of the estimates of the intervention effect from GzLMM, GEE1 and QIF by the true treatment effect θ (θ = 0, 0.2, and 0.3).

The mean estimated ESE for each method partitioned by the different levels of ICC and θ is presented in **Figure 9.5**. A high ICC value of 0.25 was associated with high ESE estimates for the three statistical methods and decreases as N increases and this was similar for the different θs. **Figure 9.5** shows that the relative ESEs of the statistical methods were not affected by the number of clusters randomised or the level of ICC used in a trial or the effect size. The maximum ESE occurred when the number of clusters was small (ten clusters) and the ICC was high (0.25), of which the ESE was approximately <0.4 for each method.

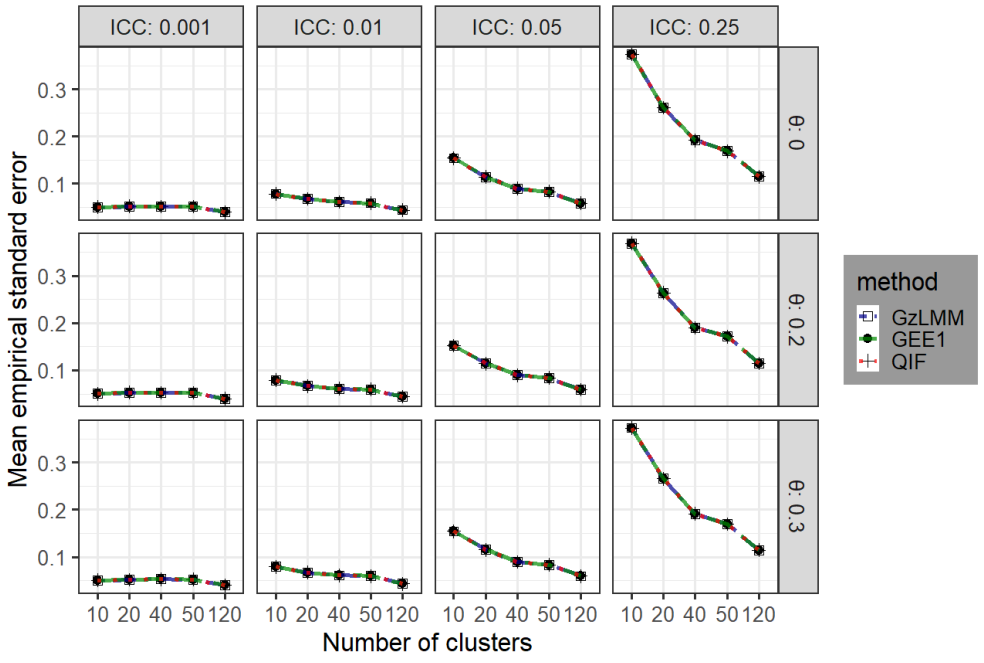


Figure 9.5 Mean empirical SEs across all scenarios for each method (GzLMM, GEE1, QIF) by three levels of ICC (0.001, 0.01, 0.05 and 0.25), three levels of true treatment effect θ ($\theta =$ 0, 0.2, and 0.3), and several numbers of clusters (10, 20, 40, 50, and 120). Each data point is based on the estimates from the 480,000 simulated datasets. The electronic version is in colour.

9.3.4 Mean square error (MSE)

The MSE is the mean sum of squares bias and variance of $\hat{\theta}$. It combines both the bias and precision of $\hat{\theta}$ from a method to assess its accuracy. This is a very useful measure because it adjusts for bias in an estimator with a spuriously low variance (Morris, White and Crowther, 2019). **Appendix 9** shows that the MSE was the same across all methods in each of the 120 cRCT scenarios, this was not unexpected since the methods have been established to be unbiased in Section 9.3.2 and they all had approximately the same ESEs (**Figure 9.5**). **Figure 9.6** shows that the MSE was mostly affected by the ICC, and moderately by the number of clusters. When the ICC was 0.001 or 0.01 or 0.05 the MSE from each of the methods was close to zero regardless of the number of clusters randomised (small is better). However, for higher ICC (0.25) the MSE for each of the methods was at the highest when the clusters were few and decayed as the number of clusters increased. The MSE as a basis for comparison among different methods is affected by the sample size, especially when the method(s) is biased (Morris, White and Crowther, 2019).

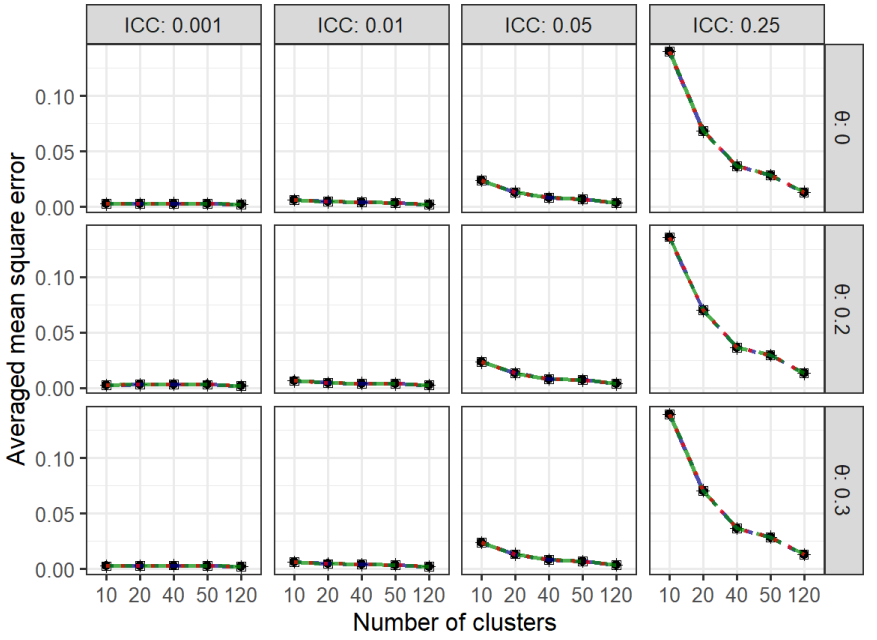


Figure 9.6 Average mean square error across all scenarios for each method (GzLMM, GEE1, QIF) by four levels of ICC (0.001, 0.01, 0.05 and 0.25), three levels of true treatment effect θ (θ = 0, 0.2, and 0.3), and several numbers of clusters (10, 20, 40, 50, and 120). Note each data point is based on the estimates from the 480,000 simulated datasets. The electronic version is in colour.

9.3.5 Coverage probability

The coverage probability is the proportion of the nominal 95% CI of the intervention effect estimates $\hat{\theta}$ that contains the true intervention effect θ. The coverage probabilities for each of the methods for each of the 120 scenarios are shown in **Appendix 10**. The results for only when the true treatment effect θ is equal to 0 or 0.2 are presented in **Table 9.2**, and for the remaining treatment effect, that is 0.3 is presented in **Table S9.1** (**Appendix 11**). **Figure 9.7** shows that low ICC resulted in better coverage probabilities for the GzLMM regardless of the number of clusters and the true effect, compared to GEE1 and QIF. GzLMM could be producing better coverage probabilities when the ICC is low, because of its ability to truncate negative ICC to zero. Therefore, when the ICC in the dataset is negative the design effect will be larger for the other methods (GEE1 and QIF) that allow negative ICC compared to GzLMM that does not. Larger design effect implies larger sample size, and consequently narrower CI and increased under coverage probabilities.

Additionally, under-coverage of the 95% CI of the estimate of the treatment effect occurred when the number of clusters was <50 regardless of the degree of clustering (ICC) and true treatment effect assumed, for GEE1 and QIF. GzLMM did not exhibit under-coverage when the ICC was 0.001. The coverage probability from the three methods was the same when the ICC was moderate to high (0.01 – 0.25) and improved as the number of clusters increased. Over-coverage was not observed since the coverage probabilities from the methods all lie below the upper limit of the CI of the coverage probability of the 95% CI of the estimate of the treatment (upper black dashed lines) in **Figure 9.7**.

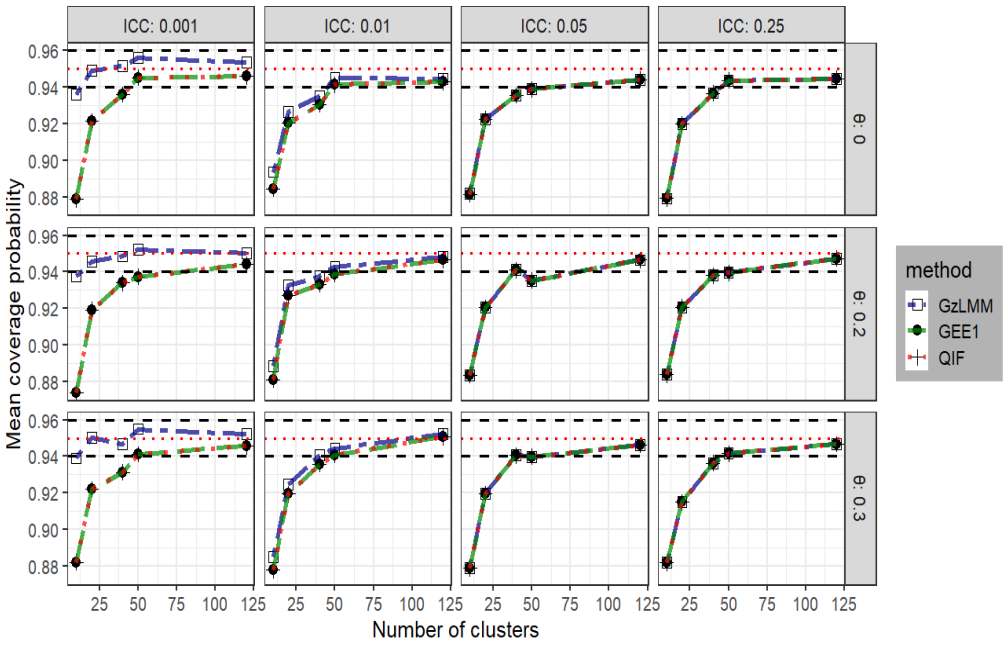


Figure 9.7 Empirical mean coverage probability of the estimated treatment effect $\hat{\theta}$ from each method (GzLMM, GEE1, QIF), across all scenarios by four levels of ICC (0.001, 0.01, 0.05 and 0.25), three levels of true treatment effect θ ($\theta = 0$, 0.2, and 0.3), and several numbers of clusters (10, 20, 40, 50, and 120). Note each data point is based on the estimates from the 480,000 simulated datasets. The red dotted line corresponds to the nominal coverage of 0.95 with its CI (upper and lower black dashed lines). The electronic version is in colour.

Figure 9.8 presents the coverage probability of the estimated treatment effect $\hat{\theta}$ from GzLMM, GEE1, and QIF by different cluster sizes, effect sizes, levels of ICCs, and numbers of clusters. It shows that increasing the number of clusters as opposed to increasing the cluster sizes had a bigger impact on the estimated coverage from each method. The GzLMM has a better coverage probability for a low degree of clustering/ICC, but this coincided with the other two methods as the ICC increases. In general, all the methods showed improvements in their coverage probabilities as the number of clusters increased, regardless of the ICC, the effect size, or the cluster sizes.

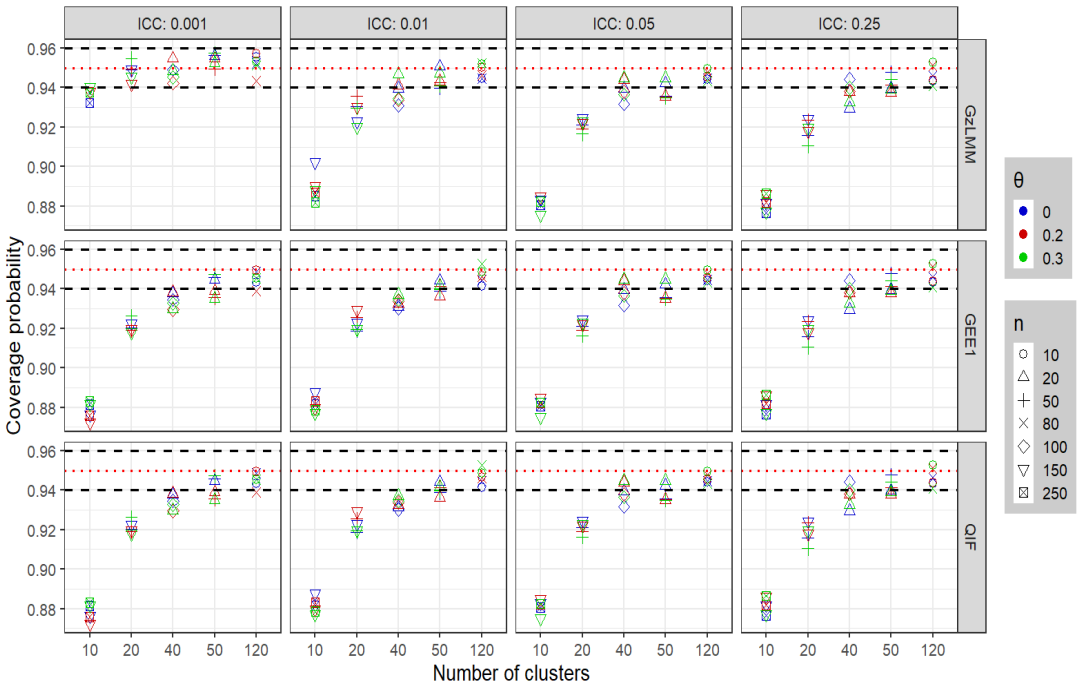


Figure 9.8 Coverage probability of the 95% confidence interval of $\hat{\theta}$ from the three statistical models (GzLMM, GEE1, QIF), for three levels of ICC (0.001, 0.01, 0.05 and 0.25) and several numbers of clusters (10, 20, 40, 50, and 120) and cluster sizes (10, 20, 50,100,150, and 250) for a variety of true treatment effects (θ = 0, 0.2, and 0.3). Note each data point is based on the estimates from 480,000 simulated datasets. The red dotted line corresponds to the nominal coverage of 0.95 with its CI (upper and lower black dashed lines). The electronic version is in colour.

9.3.6 Type I error rate

Plots of the Type I error rates for each scenario under the null hypothesis of no true treatment effect (i.e., $\theta = 0$), are presented in **Appendix 12**. The empirical mean Type I error rates by ICC and number of clusters (N) for each of the methods across all scenarios are presented in **Figure 9.9**. Similarly, the numerical results of the empirical Type I error rates are presented in **Table 9.1**. **Figure 9.9** shows that the methods have similar capabilities of controlling the Type I error rate, except when the ICC = 0.001 or 0.01 indicating lower level of clustering in the outcome data. When ICC was 0.001 GzLMM was able to maintain the nominal 5% Type I error rate when the number of clusters was > 20, compared to GEE1 and QIF. For ICC = 0.01 GzLMM had lower Type I error rate (but higher that the nominal 0.05) compared to GEE1 and QIF. In general, increasing the number of clusters causes the Type I error rate to decrease. **Figure 9.10** shows that increasing the number of clusters as opposed to increasing the cluster sizes has more impact on maintaining the nominal Type I error rate.

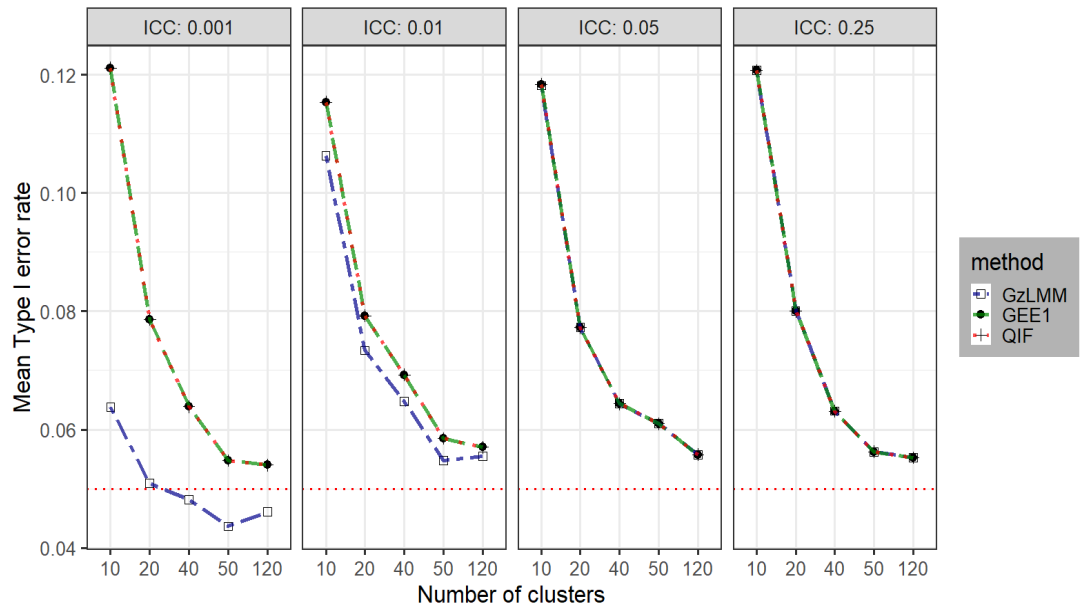


Figure 9.9 Mean Type I error rate for the three statistical methods (GzLMM, GEE1, QIF) for four levels of ICC (0.001, 0.01, 0.05 and 0.25) and various numbers of clusters (10, 20, 40, 50, and 120) under the null hypothesis of no true treatment effect $\theta = 0$. Note each data point is based on the estimates from 480,000 simulated datasets. The red dotted horizontal line corresponds to the nominal Type I error rate of 0.05. The electronic version is in colour.

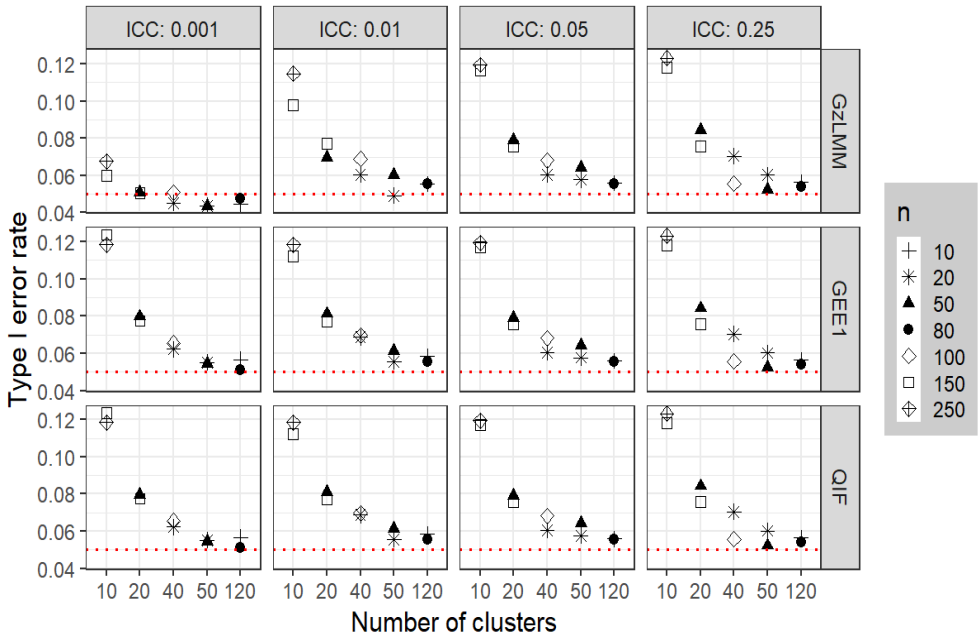


Figure 9.10 Type I error rate for the three statistical models (GzLMM, GEE1, QIF) for three levels of ICC (0.001, 0.01, 0.05 and 0.25), several cluster sizes (10, 20, 50, 80, 100, 150, and 250) and various numbers of clusters (10, 20, 40, 50, and 120) under the null hypothesis of no true treatment effect $\theta = 0$. Note each data point is based on the estimates from 480,000 simulated datasets. The red dotted horizontal line corresponds to the nominal Type I error rate of 0.05.

9.3.7 Power

The empirical power is the ability of the method to detect the least minimal effect of interest from the intervention administered when truly the intervention is effective, in other words, the proportion of the P-values of the intervention effect estimate that is $<$ 5% level of significance under the alternative hypothesis (i.e., H_1: $\theta = 0.2$ and 0.3). **Appendix 73** presents the power of methods for each of the 120 scenarios based on 4,000 estimates from each of the three methods – GzLMM, GEE1, and QIF. The results for the statistical power for $\theta = 0.2$ only are presented in **Table 9.1**, for $\theta = 0.3$, the results are presented in **Table S9.1** (**Appendix 11**). The median power of the GzLMM is 85% (IQR: 47% -100%), and that of GEE1 and QIF are the same, 86% (IQR: 47% - 100%) across all scenarios studied.

Figure 9.11 shows that for lower ICC (0.001 or 0.01) the power of all three methods was ≥80% and the number of clusters had no definite impact on power in this scenario. For moderate ICC (0.05), the power of the methods was the same and increased as the number of clusters increased. It appears that higher the effect sizes the higher the power. For effect size $\theta = 0.2$, all three methods had low powers (<80%) (black dashed line), however, when the number of clusters was >80 and the ICC was 0.05 the power became >80%. For effect size $\theta = 0.3$ and ICC = 0.05 all three methods had appropriate nominal powers (≥80%) when the number of clusters were ≥ 25. In general, for high clustering ICC = 0.25 all three methods had very low powers but increased as the number of clusters increased.

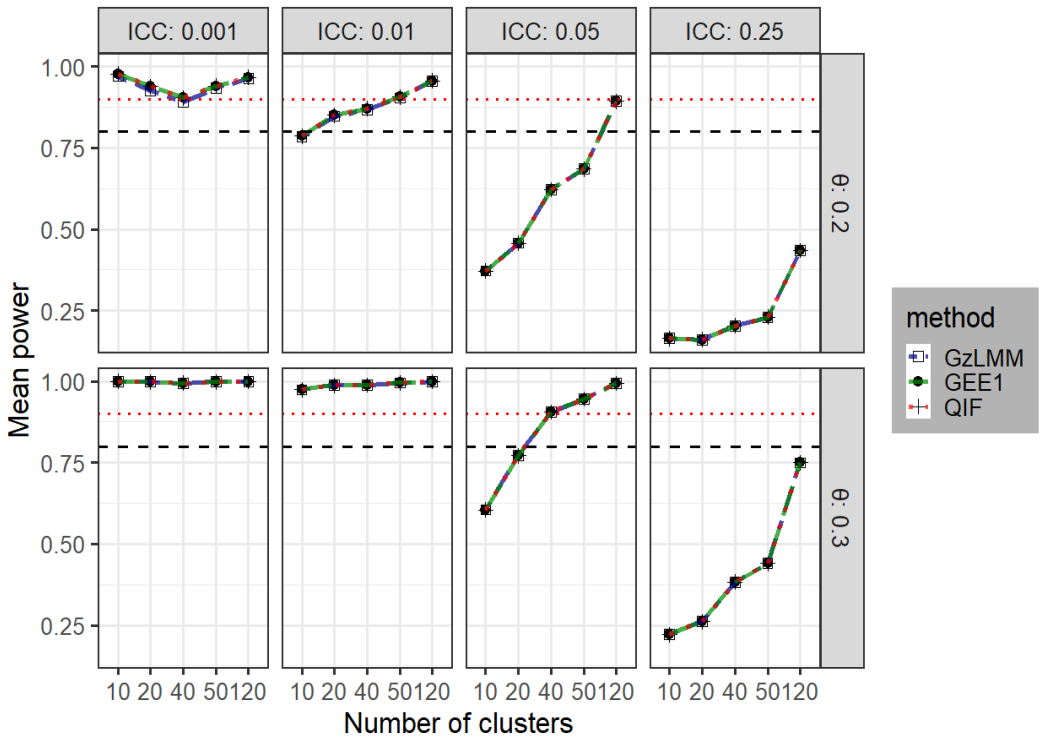


Figure 9.11 Mean statistical power for the three statistical methods (GzLMM, GEE1, QIF) for three levels of ICC (0.001, 0.01, 0.05, and 0.25) and several numbers of clusters (10, 20, 40, 50, and 120) under the alternative hypothesis of the existence of true treatment effect θ = 0.2 and 0.3. Note each data point is based on the estimates from 120,000 simulated datasets. The red dotted line is the nominal 90% power while the black dashed horizontal line is the nominal 80%. The electronic version is in colour.

Figure 9.12 shows the power of each method for different degrees of clustering, numbers of clusters, effect sizes, and cluster sizes. It shows that a low level of clustering and bigger effect sizes has more impact on the statistical power compared to cluster sizes. In summary, the power of the three methods is adequate when the ICC is low (0.001) regardless of the number of clusters randomised and the effect size assumed. For moderate clustering (ICC = 0.05) the power of the methods is adequate when the effect size was = 0.3 and the number of clusters ≥ 40. For higher clustering (ICC = 0.25) all the methods had inadequate power to detect a significant treatment effect and increasing the number of clusters or effect size or cluster sizes do not have any reasonable impact.

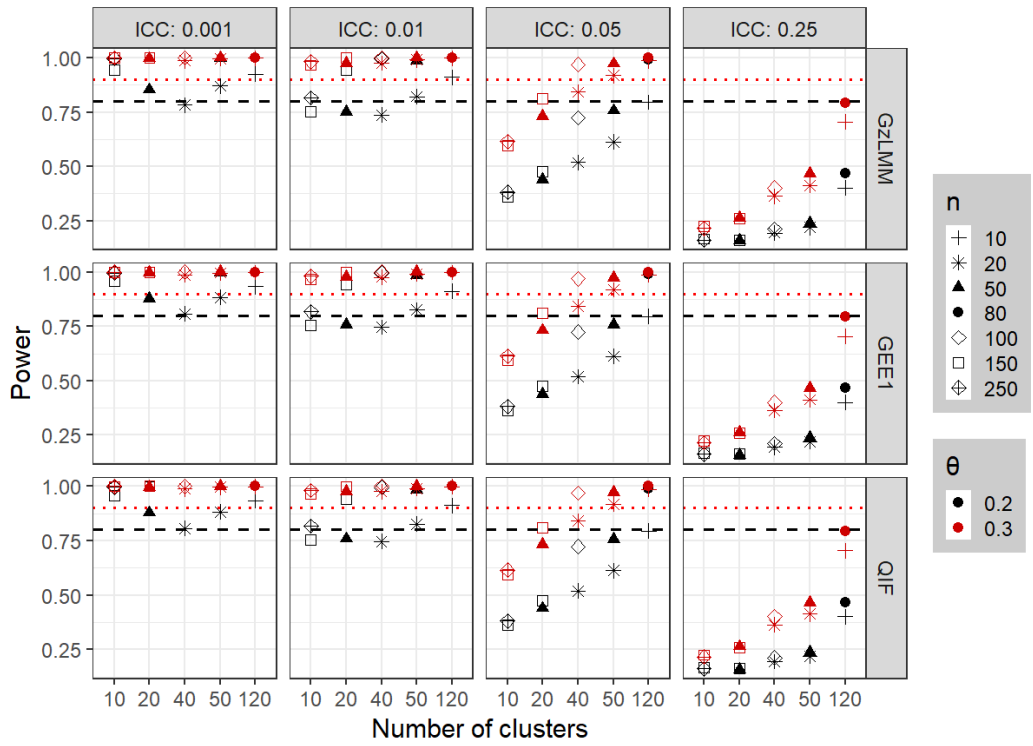


Figure 9.12 Statistical power for the three statistical methods (GzLMM, GEE1, QIF) for three levels of ICC (0.001, 0.01, 0.05 and 0.25), several cluster sizes (10, 20, 50, 80, 100, 150 and 250) and several numbers of clusters (10, 20, 40, 50, and 120) under the alternative hypothesis of the existence of treatment effect (θ = 0.2 and 0.3). Note each data point is based on the estimates from 4,000 simulated datasets. The red dotted line is the nominal 90% power while the black dashed line is the nominal 80%. The electronic version is in colour.

Table 9.1 Empirical Type I error rate and Power for each method for several cRCT scenarios specified by the combinations of N, n_i, ICC, and $\theta = 0$, and 0.2.

Parameters			$\theta = 0$			$\theta = 0.2$		
			Type I error rate (MCSE)			Power (MCSE)		
N	n_i	ICC	GzLMM	GEE1	QIF	GzLMM	GEE1	QIF
10	150	0.001	0.06 (0.0037)	0.12 (0.0052)	0.12 (0.0052)	0.94 (0.0036)	0.96 (0.0031)	0.96 (0.0031)
		0.01	0.10 (0.0047)	0.11 (0.0050)	0.11 (0.0050)	0.75 (0.0068)	0.76 (0.0068)	0.76 (0.0068)
		0.05	0.12 (0.0051)	0.12 (0.0051)	0.12 (0.0051)	0.36 (0.0076)	0.36 (0.0076)	0.36 (0.0076)
		0.25	0.12 (0.0051)	0.12 (0.0051)	0.12 (0.0051)	0.17 (0.0059)	0.17 (0.0059)	0.17 (0.0059)
	250	0.001	0.07 (0.0040)	0.12 (0.0051)	0.12 (0.0051)	1.00 (0.0010)	1.00 (0.0010)	1.00 (0.0010)
		0.01	0.12 (0.0050)	0.12 (0.0051)	0.12 (0.0051)	0.82 (0.0061)	0.82 (0.0061)	0.82 (0.0061)
		0.05	0.12 (0.0051)	0.12 (0.0051)	0.12 (0.0051)	0.38 (0.0077)	0.38 (0.0077)	0.38 (0.0077)
		0.25	0.12 (0.0052)	0.12 (0.0052)	0.12 (0.0052)	0.16 (0.0058)	0.16 (0.0058)	0.16 (0.0058)
20	50	0.001	0.05 (0.0035)	0.08 (0.0043)	0.08 (0.0043)	0.85 (0.0056)	0.88 (0.0052)	0.88 (0.0052)
		0.01	0.07 (0.0040)	0.08 (0.0043)	0.08 (0.0043)	0.75 (0.0068)	0.76 (0.0068)	0.76 (0.0068)
		0.05	0.08 (0.0043)	0.08 (0.0043)	0.08 (0.0043)	0.44 (0.0078)	0.44 (0.0078)	0.44 (0.0078)
		0.25	0.08 (0.0044)	0.08 (0.0044)	0.08 (0.0044)	0.16 (0.0058)	0.16 (0.0058)	0.16 (0.0058)
	150	0.001	0.05 (0.0035)	0.08 (0.0042)	0.08 (0.0042)	1.00 (0.0004)	1.00 (0.0004)	1.00 (0.0004)
		0.01	0.08 (0.0042)	0.08 (0.0042)	0.08 (0.0042)	0.94 (0.0037)	0.94 (0.0037)	0.94 (0.0037)
		0.05	0.08 (0.0042)	0.08 (0.0042)	0.08 (0.0042)	0.48 (0.0079)	0.48 (0.0079)	0.48 (0.0079)
		0.25	0.08 (0.0042)	0.08 (0.0042)	0.08 (0.0042)	0.16 (0.0058)	0.16 (0.0058)	0.16 (0.0058)
40	20	0.001	0.05 (0.0033)	0.06 (0.0038)	0.06 (0.0038)	0.78 (0.0065)	0.81 (0.0062)	0.81 (0.0062)
		0.01	0.06 (0.0040)	0.07 (0.0040)	0.07 (0.0040)	0.74 (0.0070)	0.75 (0.0069)	0.75 (0.0069)
		0.05	0.06 (0.0038)	0.06 (0.0038)	0.06 (0.0038)	0.52 (0.0079)	0.52 (0.0079)	0.52 (0.0079)
		0.25	0.07 (0.0040)	0.07 (0.0040)	0.07 (0.0040)	0.20 (0.0063)	0.20 (0.0063)	0.20 (0.0063)
	100	0.001	0.05 (0.0035)	0.07 (0.0039)	0.07 (0.0039)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.01	0.07 (0.0040)	0.07 (0.0040)	0.07 (0.0040)	1.00 (0.0011)	1.00 (0.0011)	1.00 (0.0011)
		0.05	0.07 (0.0040)	0.07 (0.0040)	0.07 (0.0040)	0.72 (0.0071)	0.72 (0.0071)	0.72 (0.0071)
		0.25	0.06 (0.0036)	0.06 (0.0036)	0.06 (0.0036)	0.21 (0.0065)	0.21 (0.0065)	0.21 (0.0065)
50	20	0.001	0.04 (0.0032)	0.06 (0.0036)	0.06 (0.0036)	0.87 (0.0053)	0.88 (0.0051)	0.88 (0.0051)
		0.01	0.05 (0.0034)	0.06 (0.0036)	0.06 (0.0036)	0.82 (0.0061)	0.83 (0.0060)	0.83 (0.0060)
		0.05	0.06 (0.0037)	0.06 (0.0037)	0.06 (0.0037)	0.61 (0.0077)	0.61 (0.0077)	0.61 (0.0077)
		0.25	0.06 (0.0038)	0.06 (0.0038)	0.06 (0.0038)	0.22 (0.0066)	0.22 (0.0066)	0.22 (0.0066)
		0.001	0.04 (0.0032)	0.05 (0.0036)	0.05 (0.0036)	1.00 (0.0007)	1.00 (0.0007)	1.00 (0.0007)
	50	0.01	0.06 (0.0038)	0.06 (0.0038)	0.06 (0.0038)	0.99 (0.0019)	0.99 (0.0019)	0.99 (0.0019)
		0.05	0.06 (0.0039)	0.06 (0.0039)	0.06 (0.0039)	0.76 (0.0068)	0.76 (0.0068)	0.76 (0.0068)
		0.25	0.05 (0.0035)	0.05 (0.0035)	0.05 (0.0035)	0.24 (0.0067)	0.24 (0.0067)	0.24 (0.0067)
		0.001	0.05 (0.0033)	0.06 (0.0037)	0.06 (0.0037)	0.93 (0.0042)	0.93 (0.0039)	0.93 (0.0039)
120	10	0.01	0.06 (0.0036)	0.06 (0.0037)	0.06 (0.0037)	0.91 (0.0045)	0.91 (0.0045)	0.91 (0.0045)
		0.05	0.06 (0.0036)	0.06 (0.0036)	0.06 (0.0036)	0.80 (0.0064)	0.80 (0.0064)	0.80 (0.0064)
		0.25	0.06 (0.0037)	0.06 (0.0037)	0.06 (0.0037)	0.40 (0.0078)	0.40 (0.0078)	0.40 (0.0078)
		0.001	0.05 (0.0034)	0.05 (0.0035)	0.05 (0.0035)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
	80	0.01	0.06 (0.0036)	0.06 (0.0036)	0.06 (0.0036)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.05	0.06 (0.0036)	0.06 (0.0036)	0.06 (0.0036)	0.99 (0.0015)	0.99 (0.0015)	0.99 (0.0015)
		0.25	0.05 (0.0036)	0.05 (0.0036)	0.05 (0.0036)	0.47 (0.0079)	0.47 (0.0079)	0.47 (0.0079)

Note: N is the number of clusters; n_i is the i^{th} cluster size; ICC is the intracluster correlation coefficient; θ is the true treatment effect; MCSE is the Monte Carlo standard error. Values shaded blue are equal or within the range of the expected nominal value (2 decimal places), while the ones shaded orange are greater than the expected nominal value (2 decimal places), and the ones shaded green are less than the expected nominal value. Note each cell in the table is based on 4,000 simulated datasets, except cells corresponding to GzLMM where some simulations failed to converge. The electronic version is in colour.

Table 9.2 Empirical coverage probability for each method for several cRCT scenarios specified by the combinations of N, n_i, ICC, and $\theta = 0$ and 0.2

Parameters			$\theta = 0$			$\theta = 0.2$		
			Coverage probability (MCSE)			Coverage probability (MCSE)		
N	n_i	ICC	GzLMM	GEE1	QIF	GzLMM	GEE1	QIF
10	150	0.001	0.94 (0.0038)	0.88 (0.0052)	0.88 (0.0052)	0.94 (0.0038)	0.87 (0.0053)	0.87 (0.0053)
		0.01	0.90 (0.0047)	0.89 (0.0050)	0.89 (0.0050)	0.89 (0.0050)	0.88 (0.0051)	0.88 (0.0051)
		0.05	0.88 (0.0051)	0.88 (0.0051)	0.88 (0.0051)	0.89 (0.0050)	0.89 (0.0050)	0.89 (0.0050)
		0.25	0.88 (0.0051)	0.88 (0.0051)	0.88 (0.0051)	0.89 (0.0050)	0.87 (0.0050)	0.87 (0.0050)
	250	0.001	0.93 (0.0040)	0.88 (0.0051)	0.88 (0.0051)	0.94 (0.0038)	0.88 (0.0052)	0.88 (0.0052)
		0.01	0.89 (0.0050)	0.88 (0.0051)	0.88 (0.0051)	0.89 (0.0050)	0.88 (0.0051)	0.88 (0.0051)
		0.05	0.88 (0.0051)	0.88 (0.0051)	0.88 (0.0051)	0.88 (0.0051)	0.88 (0.0051)	0.88 (0.0051)
		0.25	0.88 (0.0052)	0.88 (0.0052)	0.88 (0.0052)	0.88 (0.0051)	0.88 (0.0051)	0.88 (0.0051)
20	50	0.001	0.95 (0.0035)	0.92 (0.0043)	0.92 (0.0043)	0.95 (0.0035)	0.92 (0.0043)	0.92 (0.0043)
		0.01	0.93 (0.0035)	0.92 (0.0043)	0.92 (0.0043)	0.94 (0.0039)	0.93 (0.0042)	0.93 (0.0042)
		0.05	0.92 (0.0043)	0.92 (0.0043)	0.92 (0.0043)	0.92 (0.0043)	0.92 (0.0043)	0.92 (0.0043)
		0.25	0.92 (0.0043)	0.92 (0.0043)	0.92 (0.0043)	0.92 (0.0042)	0.92 (0.0042)	0.92 (0.0042)
	150	0.001	0.95 (0.0035)	0.92 (0.0042)	0.92 (0.0042)	0.94 (0.0037)	0.92 (0.0043)	0.92 (0.0043)
		0.01	0.92 (0.0042)	0.92 (0.0042)	0.92 (0.0042)	0.93 (0.0040)	0.93 (0.0041)	0.93 (0.0041)
		0.05	0.93 (0.0042)	0.93 (0.0042)	0.93 (0.0042)	0.92 (0.0042)	0.92 (0.0042)	0.92 (0.0042)
		0.25	0.92 (0.0042)	0.92 (0.0042)	0.92 (0.0042)	0.92 (0.0043)	0.92 (0.0043)	0.92 (0.0043)
40	20	0.001	0.96 (0.0033)	0.94 (0.0038)	0.94 (0.0038)	0.96 (0.0034)	0.94 (0.0034)	0.94 (0.0034)
		0.01	0.94 (0.0038)	0.93 (0.0040)	0.93 (0.0040)	0.94 (0.0037)	0.93 (0.0040)	0.93 (0.0040)
		0.05	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0036)	0.94 (0.0036)	0.94 (0.0036)
		0.25	0.93 (0.0041)	0.93 (0.0041)	0.93 (0.0041)	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)
	100	0.001	0.95 (0.0035)	0.94 (0.0039)	0.94 (0.0039)	0.94 (0.0037)	0.93 (0.0041)	0.93 (0.0041)
		0.01	0.93 (0.0040)	0.93 (0.0040)	0.93 (0.0040)	0.93 (0.0039)	0.93 (0.0039)	0.93 (0.0039)
		0.05	0.93 (0.0040)	0.93 (0.0040)	0.93 (0.0040)	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)
		0.25	0.94 (0.0036)	0.94 (0.0036)	0.94 (0.0036)	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)
50	20	0.001	0.96 (0.0032)	0.95 (0.0036)	0.95 (0.0036)	0.96 (0.0033)	0.94 (0.0038)	0.94 (0.0038)
		0.01	0.95 (0.0034)	0.94 (0.0036)	0.94 (0.0036)	0.94 (0.0037)	0.94 (0.0039)	0.94 (0.0039)
		0.05	0.94 (0.0037)	0.94 (0.0037)	0.94 (0.0037)	0.94 (0.0039)	0.94 (0.0039)	0.94 (0.0039)
		0.25	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)
		0.001	0.96 (0.0032)	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0035)	0.94 (0.0039)	0.94 (0.0039)
	50	0.01	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0037)	0.94 (0.0037)	0.94 (0.0037)
		0.05	0.94 (0.0039)	0.94 (0.0039)	0.94 (0.0039)	0.94 (0.0039)	0.94 (0.0039)	0.94 (0.0039)
		0.25	0.95 (0.0035)	0.95 (0.0035)	0.95 (0.0035)	0.94 (0.0037)	0.94 (0.0037)	0.94 (0.0037)
		0.001	0.96 (0.0033)	0.94 (0.0037)	0.94 (0.0037)	0.96 (0.0032)	0.95 (0.0035)	0.95 (0.0035)
120	10	0.01	0.95 (0.0036)	0.94 (0.0037)	0.94 (0.0037)	0.95 (0.0034)	0.95 (0.0036)	0.95 (0.0036)
		0.05	0.94 (0.0036)	0.94 (0.0036)	0.94 (0.0036)	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0036)
		0.25	0.94 (0.0037)	0.94 (0.0037)	0.94 (0.0037)	0.94 (0.0036)	0.94 (0.0036)	0.94 (0.0036)
		0.001	0.95 (0.0034)	0.95 (0.0035)	0.95 (0.0035)	0.94 (0.0037)	0.94 (0.0038)	0.94 (0.0038)
	80	0.01	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0036)
		0.05	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0035)	0.95 (0.0035)	0.95 (0.0035)
		0.25	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0034)	0.95 (0.0034)	0.95 (0.0034)

N is the number of clusters; n_i is the i^{th} cluster size; ICC is the intracluster correlation coefficient; θ is the true treatment effect; MCSE is the Monte Carlo Standard Error. Values shaded blue are equal or within the range of the expected nominal value (2 decimal places), and the ones shaded green are less than the expected nominal value. Note each cell in the table is based on 4,000 simulated datasets, except cells corresponding to GzLMM where some simulations failed to converge. The electronic version is in colour.

9.4 Discussion

This study was conducted to comprehensively evaluate the statistical properties of the analytical methods for analysing continuous outcome data from cRCTs. These methods: GzLMM (with identity link function and parameters estimated by MLE), GEE1, and QIF, were selected from the results of methodological and practice reviews conducted in Chapters 3 and 4, respectively. In Chapter 3, I found that GzLMM and GEE1 were the most studied methods (which may be indicative of higher interest on them), while QIF was the most studied newer method and a plausible alternative to GEE1. However, this claim about QIF has not been comprehensively investigated in the context of cRCTs. Furthermore, In Chapter 4, GzLMM and GEE1 were identified as the most used statistical methods in analysing outcome data from cRCTs. However, QIF was never considered in practice in cRCTs. I theorised this could be because the advantages of QIF over GEE1 has not been comprehensively investigated in the context of cRCT, which is the main purpose of this current Chapter. To the best of my knowledge, this is the first simulation study comparing these three methods.

This study was designed to mimic cRCTs with primary care and community settings; which are often simple designs (O'Cathain *et al.*, 2002; Morrell *et al.*, 2009; Speed *et al.*, 2010; Julious *et al.*, 2016; Surr *et al.*, 2020; Wyld *et al.*, 2021). Hence, most of the parameters for the simulation study were obtained from a review of the cRCTs published in the NIHR Journals Library where publicly funded cRCTs with primary care and community settings are often reported.

The results of this current study showed that the estimates of the intervention effect from all three methods were close to the specified true intervention effect θ, hence, their biases were close to zero. This result is consistent with that of Leyrat et al. (2018), where the results showed that the relative biases from the methods were equal, therefore bias was not an issue, and the focus was shifted to other performance measures. Their study assessed the performance of cluster-level analysis, mixed model (with parameters estimated by REML), and GEE1 in conjunction with small sample corrections for studies with few clusters (≤ 40). This current study did not assess any cluster-level analysis nor adjusted for small numbers of clusters but included QIF as one of the methods. This study simulated designs like primary care and community cluster trials with the number of clusters ranging from 10 to 120. The finding regarding bias was similar for ESE and MSE in this current study. Although, a high degree of clustering and few clusters randomised do

increase their estimates, and they show a decreasing trend as the number of clusters increases (**Figure 9.5** and **Figure 9.6**).

The significant differences in the performances of the methods occurred for coverage probability and Type I error rate. For the 95% coverage of the CI of the estimates of the intervention effect, GzLMM performed better when the degree of clustering was very low (ICC = 0.001) regardless of the number of clusters randomised which ideally should be reasonable (N > 40) (Leyrat *et al.*, 2018). GEE1 and QIF had improved coverage only when the number of clusters was moderate $N \geq 50$. For moderate and high degrees of clustering all three methods became equivalent, in this situation, they had better coverage when the number of clusters $N \geq 50$. The results of this study are similar to those of Yu, Li and Turner (2020) that compared GEE1 and QIF, their coverage probabilities were the same in all scenarios where an exchangeable correlation structure was assumed. Similarly, the coverage probability between GzLMM and GEE1 was equivalent for scenarios with low variance inflation factor in the simulation study of Ma et al. (2013).

For Type I error rate, only GzLMM was able to maintain the nominal 5% when the clustering was very low ICC = 0.001 and the number of clusters was > 20. For moderate and high clustering ICC = 0.05 and 0.25, respectively, the Type I error rate of the three methods were the same and well above the nominal 5%. The high Type I error rate decayed as the number of clusters N increases, but still exceeded the nominal 5%. Like the results of coverage for GEE1 and QIF, their capability to control the Type I error rate was equivalent in all scenarios (**Figure 9.9**). The recommendation is to use a small sample correction both for cluster-specific models like GzLMM and population average models like GEE1 when the number of clusters is <40(Leyrat *et al.*, 2018; Thompson *et al.*, 2022). The only study that directly compared GEE1 to QIF concerning power found that QIF has more power compared to GEE1 across all scenarios, however, this study was not comprehensive; the study used some parameters with single fixed value each to form its DGMs, such as 100 clusters only (number of clusters) with 25 subjects per cluster only (cluster size). Some other similar studies mainly focussed on the impact of small number of clusters on the efficiency of the two methods (i.e., GEE1 and QIF) (Westgate, 2012; Westgate and Braun, 2012, 2013).

The results for power showed that the statistical power of the three methods is the same for the different degrees of clustering, and the true intervention effects considered. The statistical power of the three methods was above 95% when clustering was low (ICC = 0.001 or 0.01). For moderate clustering (ICC = 0.05) the power of the three methods was $\geq$ 80% when the number of clusters

N >20, and high true treatment effect (θ = 0.3), which is impressive since it is typical of primary care and community cluster trials to have ICC within this range (Adams *et al.*, 2004). For a high degree of clustering (ICC = 0.25), all the statistical methods had power < 40%, fortunately, high degree of clustering is not common in cRCTs with primary care or community settings. It is apparent from **Figure 9.12** that effect size does have an impact on the power of the methods whereas that was not the case for cluster sizes.

In summary, the three analytical methods, GzLMM, GEE1, and QIF were equivalent with regards to bias, ESE, MSE, and power within the parameters of the simulations. Some differences occurred for coverage probability and Type I error rate. GzLMM (with identity link and parameters estimated by MLE) performed better w.r.t coverage probability and Type I error rate, when the degree of clustering is low (ICC = 0.001) compared to GEE1 and QIF within the parameters of the simulations. Approximately, GEE1 and QIF performance was the same in all scenarios w.r.t to all the performance measures used to evaluate their statistical properties, and within the parameters of the simulations. These findings are generalisable to most cRCTs with primary and community settings conducted in the UK based on the findings of the practice review of publicly funded cluster trials reported in the NHIR Journal Library (Chapter 4).

9.5 Summary

In this Chapter, the results obtained from implementing the simulation study detailed in Chapter 8 are presented. This simulation study was informed by the review of methods available in the literature for analysing cRCTs in Chapter 3 and the review of the methods used in practice in Chapter 4. Results from the analysis of motivating datasets from cRCTs in Chapter 7 showed that GEE1 and GEE2 produced equivalent estimates for the treatment effect, and its SE, 95% CI, and P-value. However, their estimated ICCs were different, hence, GEE2 was not included as a method in this simulation study for further investigation. Studies have shown that the extra complexity of GEE2 is not necessarily relevant if the mean parameters are of primary interest (Liang, Zeger and Qaqish, 1992; Balemi and Lee, 2009).

This chapter attempts to elucidate the capabilities of the three methods by evaluating their performance based on their statistical properties through a comprehensive computer simulation study. In summary, the findings of this study showed that for a continuous outcome the performance of GzLMM (with identity link and parameters estimated using MLE), GEE1, and

QIF are similar based on most of the performance measures calculated, except for coverage probability and Type I error rate within the parameters of the simulations. GzLMM had better coverage probability when clustering was low (ICC = 0.001) compared to GEE1 and QIF, whereas that of GEE1 and QIF were equivalent across all scenarios. Similarly, GzLMM had better control of the 5% nominal Type I error rate compared to GEE1 and QIF for low clustering (ICC = 0.001). However, for moderate and high degrees of clustering (ICC = 0.05 and 0.25) the ability of the three different methods to maintain the Type I error rate was equivalent. These results are generalisable to most cluster randomised controlled trials with primary care and community settings conducted in the UK. Chapter 10 discusses these findings explicitly with limitations that could affect the conclusion reached in this study.

Chapter 10

Discussion, Conclusions, and Future Studies

10.1 Introduction

Since the acclaimed promising alternatives (GEE2 and QIF) to GEE1 were proposed, their comparative advantages to commonly used GEE1 are still unclear in the cRCT literature. Studies investigating the performances of GEE2 and QIF against GEE1 are sparse. It is also worth noting that no study has investigated the performance of QIF as a population average model against any cluster-specific model, for instance, the GzLMM with parameters estimated by MLE/REML. The primary aim of this research was to carry out a comprehensive investigation on the statistical performance of the selected methods (GzLMM, GEE1, and QIF) within the context of cRCT design. Chapter 2 provided basic background on the major concepts used in this research, such as types of cRCT designs, reasons why a researcher would choose a cRCT over an RCT design, and descriptions of classical statistical methods for analysing outcome data from cRCTs with examples of their application to real-world data. Chapter 3 went beyond the classical methods described in Chapter 2, by reviewing the literature for available and appropriate methods for analysing outcome data from cRCT designs. Hence, Chapter 3 covered one of the objectives of this study – to identify the available and appropriate methods in the cRCT literature.

In Chapter 3, five NIHR Journals Library were audited to review the statistical methods that are used in practice to analyse outcome data from cRCTs published in the UK (also one of the thesis objectives). Results from the methodological review in Chapter 3 showed that QIF and GEE2 are the most potential alternatives to GEE1, while the results in Chapter 4 showed that GzLMM and GEE1 are the commonly used methods in practice, however, none of the acclaimed alternatives, QIF and GEE2, were used in practice (Chapter 4). The gaps in knowledge identified from the two reviews conducted in Chapters 3 and 4 led to the conceptualisation of the key ideas for this research in Chapter 5. The research questions, aim, and objectives were presented in Chapter 5. Chapter 6 describes the selected methods – GzLMM with parameters estimated by MLE/REML, GEE1, GEE2, and QIF. This was one of the main objectives of this research. The methods were applied to four cRCT example data in Chapter 7, to assess their performance in

real-world cRCT settings with finite sample sizes. This also addressed one of the major objectives of this research.

Chapter 8 presented the details of a planned simulation study to investigate the superiority of the methods in different cRCT scenarios with primary and community settings. The simulation study involved generating pseudo-random continuous outcome data which was then analysed using the three selected methods – GzLMM (with identity link function and parameters estimated by MLE), GEE1, and QIF. The results in Chapter 7 showed that GEE1 and GEE2 produced the same estimates of the regression parameters, the only difference was their parameter estimates for the ICC. The ICC (an association parameter) was not of primary interest in this study, hence, GEE2 was dropped from further investigation. Simulation studies are computer experiments that provide several scenarios that depict real-world clinical trials (Burton *et al.*, 2006; Morris, White and Crowther, 2019). The simulation study of Chapter 9 generated several scenarios depicting cRCT with primary care and community settings (Ukoumunne *et al.*, 1999; Eldridge *et al.*, 2001; Adams *et al.*, 2004; Walters, Morrell and Slade, 2011).

This present Chapter pulls together the findings from previous Chapters 3, 4, 7, and 9. It harnesses the findings from the methodological scoping review of the available methods for analysing cRCTs (Chapter 3), the practice review (Chapter 4), and results from applying the four selected methods to real-world (Chapter 7) and simulated datasets (Chapters 8 and 9) to provide some practical recommendations on the design and analysis of cRCTs. This Chapter begins by stating its primary aim in Section 10.2, the comprehensive summary of the reviews, and real-world and simulated data analyses are summarised in Section 10.3, with the limitations of the study presented in Section 10.4. The implications and recommendations deduced from the findings of this research are given in Section 10.5. The definite conclusions reached are spelled out in Section 10.6, and this research concludes by describing areas that should be explored in future research in Section 10.7. Finally, a recap of issues discussed in this Chapter is presented in Section 10.8.

10.2 Chapter aim

This Chapter aims to pull together all the findings from previous chapters to deduce the implications to practice, make possible recommendations while considering the limitations of the research, and reach definite conclusions about the performance, and hence the superiority of the statistical methods evaluated.

10.3 Summary of thesis

Chapter 2 presented an overview of cluster randomised controlled trials with a focus on primary care and community settings. Concepts covering the merits and demerits of cRCTs compared to RCTs, different ways of randomising participants to treatment arms giving rise to types of cRCT designs, ways of achieving balance in conjunction with the randomisation scheme, ICC, different approaches of analysing outcome data from cRCTs and the consequences of ignoring the correlation of outcomes in a cluster (AKA clustering) were described. When designing a cRCT one of the important issues to consider is the proper accounting of clustering through the incorporation of the ICC in the sample size calculation, this helps with the improvement of the efficiency of the cRCT design (Hayes and Moulton, 2009). Due to the randomisation at the cluster level, cRCTs have two sources of variability - between-cluster variance σ_b^2 and individual participants variance σ_w^2. Therefore, the total variance which is a parameter in the sample size calculation formular should account for both σ_b^2 and σ_w^2 in cRCTs (i.e., $\sigma^2 = \sigma_b^2 + \sigma_w^2$), whereas in RCT, $\sigma^2 = \sigma_w^2$ is adequate. Any attempt to ignore σ_b^2 in the sample size calculation would cause the study to be underpowered, and consequently produce less precise parameter estimates.

One of the important issues in cRCTs at the analysis phase is accounting for clustering using appropriate methods. Recent reviews found that clustering is still being ignored and the consequences are false low SEs, false small P-values, and narrower confidence intervals (Hayes and Moulton, 2009; Campbell and Walters, 2014a; Leyrat et al., 2018; Thompson et al., 2022). Recommended approaches for analysing cRCTs outcome data are broadly categorised into two: cluster-level and individual-level analyses. The cluster-level analysis approach was briefly described in Section 2.8.2, but this thesis focused on the individual-level analysis approach described in Section 2.8.3. Under this analytical approach, there are two broad categories based on how clustering is addressed, they are cluster-specific/conditional and population average/marginal regression approaches. The most common models under this modelling approaches are the GzLMM and GEE1, respectively (Twardella, Bruckner and Blettner, 2005; Offorha, Walters and Jacques, 2022), and they are briefly described therein.

Chapter 3 focussed on discovering more analytical methods that are available and appropriate for analysing outcome data from cRCTs in the literature beyond the classical methods described in Chapter 2. This was achieved by conducting a methodological scoping review involving a

systematic search and auditing of five journal databases, namely: MEDLINE, EMBASE, PsycINFO (via Ovid), CINAHL (via EBSCO), and SCOPUS. Also, attempts were made to locate grey literature, pearl growing was also used for a comprehensive search result. The scope of the review covered mostly methodological papers that proposed, refined, or compared statistical methods suitable for analysing cluster trials. Here, cluster trials are trials with clustering in all treatment arms (Hayes and Moulton, 2009; Campbell et al., 2012; Campbell and Walters, 2014a). Of the 1573 papers identified, only 1073 were remaining after deduplication. And further exclusion was done after the title and abstract of the remaining papers were screened to identify papers reporting cRCT(s), which resulted in 116 papers being shortlisted for further full-text assessment, of which 55 articles were finally approved for information synthesis. Most of the papers focussed on comparing already existing methods (62%, 34/55), followed by 25% proposing new methods, and lastly 7% refined existing methods. The review identified 27 unique statistical methods for analysing outcome data from cRCTs, and they were studied 112 times.

The results of Chapter 3 showed that GEE1 was the most studied method (21%, 23/112), followed by GzLMM with parameters estimated by MLE (16%, 18/112), Bayesian methods (12%, 13/112), REML (11%, 12/112), and t-test (6%, 7/112). Other than these common methods, QIF was the most studied newer method (5%, 5/112). Statistical methods for analysing binary outcome data were the most studied, followed by those for continuous and counts outcome data respectively (**Table 3.2**). Most of the trials recorded some missing data but did not explicitly state how it was addressed (83%, 46/55). It is plausible that they did not use the recommended missing data analytical methods. Likewise, some studies based their inference on complete cases analysed compared to the other few that used recommended methods for analysing missing data like multiple imputations, 4/9 vs 5/9. Inadequate handling of missing data remains an issue in cRCTs (Twardella, Bruckner and Blettner, 2005; Offorha, Walters and Jacques, 2022), which has attracted research in this regard (Ma *et al.*, 2013; Díaz-Ordaz *et al.*, 2014; Caille, Leyrat and Giraudeau, 2016; Prague *et al.*, 2016).

Most of the papers that proposed a new method did compare them to already existing ones, and QIF was the most studied in this regard (**Table 3.1**). Qu, Lindsay and Bing (2000) proposed QIF as an alternative to GEE1 and have shown it to have some advantages over GEE1 in the context of longitudinal studies. Studies have attempted to investigate if this advantageous claim of QIF over GEE1 holds in the context of cRCT, but there is sparse literature in this regard (Song *et al.*, 2009; Westgate, 2012; Westgate and Braun, 2012, 2013; Yu, Li and Turner, 2020). Three out of

these five studies identified in the review of Chapter 3 focussed only on the relative efficiency of the methods (Westgate, 2012; Westgate and Braun, 2012, 2013). The most recent study that used several performance measures to compare the two methods was not comprehensive. The study only investigated scenarios where the parameters were fixed (i.e., had single level), for example, the number of clusters was N = 100, and cluster size was $n_i = 25$ (Yu, Li and Turner, 2020). It is not uncommon to have few to moderate clusters in cRCTs, for example two of the studies used in the empirical analysis of Chapter 7, had 10 and 43 clusters respectively (O'Cathain *et al.*, 2002; Wyld *et al.*, 2021). Thus, the claim that QIF is an alternative to GEE1 has not been subjected to a comprehensive assessment, especially in cRCT settings.

Chapter 4 could be considered a complement to Chapter 3; it was conducted to investigate the statistical methods used in practice to analyse primary outcome data from cRCTs. Specifically, information on the design, conduct, and analysis were extracted from each report identified. To identify the reports on cRCTs (with clustering in treatment arms), each of the five online NIHR Journals Library was searched systematically from January 1997 to 15th July 2021. The search identified 1,942 reports. After each title and abstract of the 1,942 reports were screened, 118 reports containing cRCTs were identified. Further full-text screening resulted in 79 reports containing 86 cluster randomised controlled trials, this was because some reports contained more than one independent trial. It is worth noting that information from Chapters 3 and 4 was fundamental to the simulation study of Chapters 8 and 9.

Of the 86 trials included, the majority were parallel group cRCTs with two treatment arms. These trials were mainly conducted within primary care settings (29%, 25/86), hospitals (5%, 4/86), and communities (3%, 3/86). One hundred primary outcomes were analysed in the 86 trials, and most of the trials measured continuous (65%, 65/100) or binary (28%) outcome data. The most used analytical method was GzLMM (80%, 80/100), regression with robust SEs (7%, 7/100), and GEE1 (6%, 6/100). It is worth noting that few researchers ignored clustering and used standard statistical methods (5%). Almost all the analyses were individual-level analysis carried out using individual-level participant outcome data, except for two trials that used cluster-level approach. These results are consistent with that of a similar systematic review that focussed more on the methods used to handle missing outcome data and less on the methods used to estimate the intervention effect (Twardella, Bruckner and Blettner, 2005). The results also showed that no trial used any of the newer analytical methods identified in Chapter 3, for example, QIF.

Based on the gaps in knowledge that were identified in Chapters 3 and 4, this current study's research questions, aim and objectives were conceptualised in Chapter 5. Recall that Chapter 3 showed that the common methods included, GzLMM (with MLE/REML), GEE1, and QIF (a newer method). However, in Chapter 4, none of the trials used QIF for primary outcome analysis or any other newer methods. This may not be unconnected to the fact that the performance of QIF has not been comprehensively evaluated against those common methods of Chapter 4 to warrant its routine application in cRCTs. This gave rise to the primary aim of this thesis which was "to comprehensively evaluate the statistical performance of the identified emerging methods against already existing methods based on their long-run statistical properties". Based on this, two commonly used methods were selected, GzLMM and GEE1, to be compared against two newer alternatives – QIF and GEE2.

These four selected methods were described in detail in Chapter 6. Briefly, GEE1 is one of the commonly used population average models in cRCTs and it uses a separate estimating equation different from the mean model to allow for the association among outcomes in a cluster without explaining its origin. GEE1 as an estimator is consistent even when the assumed correlation structure is wrong, but there is some efficiency penalty inherent especially when the number of clusters is small, true correlation is substantial, the assumed correlation is not close to the true, and the cluster size is informative. Due to this shortcoming, GEE2 was proposed to explicitly model the mean and association parameters simultaneously using different estimating equations, this strategy could minimise the impact of the misspecification of the correlation structure. The merit of GEE2 over GEE1 is greatest when the association among outcomes is also of primary interest. Another acclaimed improvement over GEE1 is the QIF. Instead of using a working correlation matrix, the QIF uses basis matrices and some constants in its covariance matrix formulation, hence, it avoids the problems of misspecification of the correlation structure.

The fourth method described in Chapter 6 is the GzLMM with its parameters estimated by MLE or REML. The GzLMM typifies a cluster-specific model in which a single equation models both the fixed effects of the covariates and the random effects of the clusters. In general, parameter estimates from a GzLMM are interpreted differently from those of the population average models. A commonly used estimator of the coefficients of a GzLMM is MLE. The MLE partially differentiates the maximum joint log-likelihoods of the responses, and it does have some flaws when there are few clusters in a trial, such as producing falsely low SEs, inflated Type I error rate, and abnormally small P-values. An alternative estimator for the GzLMM when MLE is not optimal

is REML. It circumvents the problems of using the MLE when the number of clusters is small in a trial (Lam and Ip, 2003; Zhang, 2015; Leyrat *et al.*, 2018).

The modelling approaches described in full in Chapter 6 and briefly above were applied to both continuous and binary outcome data from four example cRCT datasets in Chapter 7. The primary aim of Chapter 7 was to provide practical guidance on the application of the four methods and make comments on their general behaviour across the four different case studies. The number of clusters ranged from 10 to 100 with individual participants ranging from 784 to 9,207. In most cases, the parameter estimates, for the treatment effect, from the four methods were equivalent. However, for case studies with small to moderate clusters, the results obtained from QIF were different from the other three methods. The parameter estimates from GEE1 and GEE2 were the same, except for estimates of the ICC. In Chapter 7, small sample size corrections were used in conjunction with GzLMM and GEE1, the corrections for GEE2 and QIF are not readily available in standard statistical packages. The results from the small sample correction analysis of the Informed Choice trial showed that the differences lie in the P-values and 95% CIs of the treatment effect estimates, for both the continuous and binary outcomes. The corrected P-values are bigger, and the CIs are wider. Since using only example data to reach definite conclusions on the superiority of the methods in different scenarios would be against scientific principles, a simulation study was planned in Chapter 8.

Chapter 8 presented the plan for a simulation study. A total of 120 data generating mechanisms were generated by combinations of 5 numbers of clusters, 6 cluster sizes, 4 ICCs, 3 effect sizes, and 3 statistical methods. Each DGM/scenario was repeated 4000 times resulting in 480,000 datasets of 1,440,000 estimates of the treatment effect with its associated SE, CI, and P-value. The performance measures calculated include bias, empirical SE, mean square error, coverage, Type I error rate, and power. The statistical methods applied to the continuous simulated outcome data and their performance investigated were GzLMM (with identity link function and parameters estimated by MLE), GEE1, and QIF. The two population average models, GEE1 and QIF, were based on exchangeable correlation structures and robust SEs. The results from the simulation study are presented in Chapter 9.

10.4 Discussion

In this research, four different approaches for analysing cRCTs with clustering in the treatment arms were identified from reviews of Chapters 3 and 4 and they were described in Chapters 2 and 6. The four analytical methods (GzLMM, GEE1, GEE2, and QIF) were applied to the four case studies with different features, to demonstrate their implementation and evaluate their use in practice. To the best of my knowledge, this study is the first to compare these four methods in the context of cRCTs. The initial intention was to use a free and open-source software package to conduct the analysis of the example data in Chapter 7, such as R statistical software, but I resorted to using R to fit only GEE1 and GEE2. I used the SAS macro "*QIF"* to fit QIF models because its sister version in R could not fit the QIF models to datasets of trials with a cluster size of one (i.e., where only one outcome was observed in a cluster). PoNDER and Bridging the Age Gap trials both had clusters with one observed outcome only. We communicated this to one of the authors of both software packages, Peter X.K. Song, through email correspondence and Song promised to investigate it.

The case studies considered have small estimates for the ICC which is expected from trials similar in design to primary care (Adams *et al.*, 2004) and community cluster trials (Ukoumunne *et al.*, 1999). All had an ICC less than 0.05 and three studies had ICC estimates that were less than 0.02. This indicated that there was little clustering of outcomes as would be expected from cRCTs with primary care and community settings. Three studies produced negative estimates for the ICC when GEE1 and QIF were used. Theoretically, the ICC is bounded between 0 and 1. But in practice, negative ICCs can be realised from real-world data with finite samples. GzLMM truncates the ICC estimate to zero instead of producing a negative value, somewhat fitting a GzLM (Nelder and Wedderburn, 1972), but that is not the same for the other three population average models – GEE1, GEE2, and QIF (Eldridge, Ukoumunne and Carlin, 2009).

Upon reading the documentation of the functions for fitting the PAMs, *geeglm (for GEE1), geese (for GEE2)* functions in *R* and the *qif* macro in SAS it was difficult to ascertain which of the estimators; either equation

(2.1) or (2.7) was employed in computing their ICC estimates. However, it is more likely that the PAMs used (2.7) or a method like it, which could explain why negative ICC estimates were obtained. From a sample survey perspective, sampling error due to finite sample cluster size

compared to the population cluster size which is assumed to be infinite could be one of the causes of the negative ICC estimates (Eldridge, Ukoumunne and Carlin, 2009). Another reason is when there are large discrepancies in the allotment of trial resources within the clusters, this would cause large variations in the observed outcomes (Campbell and Walters, 2014a).

Results showed that estimates for the intervention effect, and its associated SE, P-value, and 95% CI were the same for GEE1 and GEE2 methods in almost all cases, they only differ in their estimates for the ICC. Effectively both methods are fitting the same models regardless of whether the correlation parameter is estimated or considered to be a nuisance within the method formulations. In GEE2 the ICC parameter is explicitly modelled which could be recourse to producing a more consistent ICC estimate compared to GEE1 (Yan and Fine, 2004; Crespi, Wong and Mishra, 2009). The four case studies covered the key features of some settings of the cRCT design, hence the need for a simulation study to cover more scenarios of cRCTs. Simulation studies present situations where the truth is known and used as a reference to assess the estimates. The impact of these key features of the four case studies on the estimates from the four statistical methods is evident in the results obtained.

For example, the PoNDER trial was conducted in a primary care setting and hence had a large sample size (100 clusters with an average cluster size of 26 participants). The unadjusted and adjusted estimates of the intervention effect from the four different methods are slightly different for the continuous outcome but were the same for the binary outcome. The odds ratios obtained from the adjusted analysis possibly show the noncollapsible feature of the logistic regression model (with a logit link) – where including a baseline covariate changes the size of the estimate of the intervention effect if the covariate is related to the outcome, even if the covariate is not related to the treatment conditions (Westgate and Braun, 2012; Daniel, Zhang and Farewell, 2021).

On the aspect of hypothesis testing, the conclusions were the same using any of the four statistical methods which are consistent with the original analysis by Morrell et al. (2009); a significant benefit of training health visitors to adequately manage women with postnatal depressive symptoms (i.e., favouring the intervention arm). The ICC estimates were small, for the binary outcome the ICC estimates from the population average models were even negative. This result is consistent with the findings of Adams et al. (2004), they reanalyse thirty-one cRCTs conducted within primary care settings and provided ICC estimates for several common variables, their median unadjusted ICC was 0.01 while the adjusted was 0.005. The results from PoNDER trial also conform to that

of previous simulation studies, these studies found that both cluster-specific models (GzLMM) and population average models (e.g., GEE1) gave similar results for cRCTs that have many clusters and small ICC with binary (Heo and Leon, 2005) or continuous outcomes (Leyrat *et al.*, 2018). Hence, for large trials with low correlation among outcomes, any of the four models – GzLMM, GEE1, GEE2, and QIF could be used. Therefore, the choice of which method should be used should be based on the aim of the research.

Bridging the Age Gap trial had a moderate sample size (43 clusters with an average cluster size of 18 participants), and small ICC estimates. When the ICC estimate was negative, the estimates of the intervention effect were also negative across the population average models. A negative ICC is difficult to interpret since by the definition of the ICC it should be constrained between 0 and 1. This implies that negative ICC indicates no clustering among outcomes, however, methods for analysing clustered outcomes should still be used since it is difficult to know whether a negative ICC would be obtained when the data has not been analysed. It is a bad practice to look into the data before deciding on the analytical method to use, it is recommended that the chosen analytical approach should be determined and stated in the protocol (Campbell and Walters, 2014a). Across the four statistical models, the unadjusted intervention effect estimates were unstable ranging from -0.28 to 0.12 but became stable (mean difference = 1.78) after the baseline outcome covariate was adjusted for, except for QIF (mean difference = 1.46) which also had the smallest SE estimates. This elucidates the importance of accounting for relevant prognostic factors in clinical trials, especially the baseline outcome values (Samsa and Neely, 2018).

QIF appears to be slightly more precise than the other methods (i.e., had smaller SEs). However, this result should be interpreted with caution since the estimate of its intervention effect could be biased – methods that are biased toward the null hypothesis often tend to have smaller SEs (Morris, White and Crowther, 2019). A Study has confirmed the possibility of QIF producing a biased estimate of SE for trials with small to moderate numbers of clusters (Westgate and Braun, 2013). Similarly, studies have found that the GzLMM with parameters estimated by REML performed better than GEE1 in maintaining the nominal Type I error rate and power, for continuous (Leyrat *et al.*, 2018) and binary outcomes (Thompson *et al.*, 2022) when the number of clusters is small or moderate. However, all four statistical methods resulted in the same inference and are consistent with that of the original analysis which was "no significant difference in the Global QoL between the control and the intervention arms" (Wyld *et al.*, 2021). Nonetheless, an inflated Type I error rate should be avoided, methods that tend to inflate the Type I error rate for

a particular scenario should not be used, because they would lead to more false positive results (i.e., a significant result even when the null hypothesis of no effect is true) than the usually assumed nominal rates (1% or 5%).

The Informed Choice trial had a few clusters (10 clusters) with a large average cluster size (median cluster size = 145). The original study was based on a cross-sectional repeated measurement approach, so the estimate for the treatment effect was the interaction term between the treatment group (*group*) and the time of measurement (*time*). For demonstration, we used only the "after intervention" postnatal sample. Both cluster and individual-level covariates were accounted for in the adjusted analyses. Three of the methods produced approximately the same estimates which differed from that of QIF, for both continuous and binary outcomes. The most obvious difference lies in the P-values and CIs. For the continuous outcome, the adjusted P-value of GEE1 (and the other population average models) was significant whereas that of the GzLMM was not (**Table 7.5**). This could indicate that the small number of clusters had more impact on the population average models than on the cluster-specific model.

For binary outcome, the unadjusted and adjusted P-values of QIF were significant but that of the other three methods were not. The QIF CI estimates were also narrower than that of the other models. This is indicative of a possible inflated test size, and bias in the estimated intervention effect which is consistent with the findings of previous studies by Westgate (2012), and Westgate and Braun (2013). The impact of the interplay between the small number of clusters, covariates, and cluster size imbalance on QIF and GEE1 has been studied. It was found that QIF was severely affected compared to the GEE1 (Westgate and Braun, 2012). A correction was proposed to improve the empirically estimated covariance matrix that causes the QIF to be poorly behaved (Westgate and Braun, 2013). Also, GzLMM was found to perform better than GEE1 in maintaining the nominal Type I error and power in trials with few clusters (from 30 to 40 clusters) for both continuous (Leyrat *et al.*, 2018) and binary outcomes (Thompson *et al.*, 2022). Our current results confirmed these previous findings; however, it is more likely that the differing performance of the QIF estimator is due to the small number of clusters rather than covariate imbalance. Given these findings, it is likely that the QIF is severely affected by few to moderate numbers of clusters, followed by GEE1 then GzLMM. Although, no simulation study has been carried out to compare these three methods in this regard, to reach a definite conclusion. Small sample corrections were applied to GzLMM and GEE1. There are recommended corrections for GEE2 (Zhang *et al.*, 2023)

and QIF (Westgate, 2012), however, they are not readily available or easy to implement in standard statistical packages, respectively.

The corrections resulted in bigger P-values and wider CIs for both GzLMM and GEE1. Hence, small sample adjustments should be made when a trial has few to moderate clusters to avoid spurious results (Hemming and Taljaard, 2023). These findings are consistent with the findings of previous studies (McNeish and Stapleton, 2016; Leyrat *et al.*, 2018; Thompson *et al.*, 2022). Small sample size corrections should be used in conjunction with the chosen analytical method for trials with few clusters, although, these corrections do negatively affect the power of the study differently across the different scenarios. Therefore, the optimal analytical method will be the one that can maintain the nominal Type I error rate and power, especially when applied in conjunction with a small sample correction, in a particular scenario based on the study features like the number of clusters, degree of clustering, and average cluster size (Leyrat *et al.*, 2018).

Lastly, the NOSH trial measured only binary primary outcomes and had large sample sizes (92 clusters with an average cluster size of 100 participants). The parameter estimates from the four statistical methods are approximately the same, hence, their performance was equivalent. The key feature of the NOSH trial which is different from other case studies is that only cluster-level covariates were adjusted for. The performance of the four methods performed was equivalent in this case.

In Chapter 9, the simulation study results showed that only GzLMM failed to converge to solutions with a small non-convergence rate of 0.0025%. The three statistical methods produced an unbiased estimate of the intervention effect in most cases, neither the number of clusters nor the degree of correlation affected this result. Similarly, the empirical SEs and power of the three methods were the same. The differences in the performance of the methods became obvious for coverage probability and Type I error rate. For coverage, low clustering (ICC = 0.001) resulted in better coverage for GzLMM regardless of the number of clusters, unlike GEE1 and QIF which showed better coverage only when the number of clusters was big ($\geq$ 50). When the ICC became $\geq$ 0.01 all the three methods had equivalent coverage probability. The average cluster size had no impact on the 95% coverage probability of the intervention effect estimate from the three methods. All three methods had inflated Type I error rates in most scenarios, but only GzLMM was able to control the Type I error rate when the correlation was low (i.e., ICC = 0.001) and the

number of clusters was > 20. An inflated Type I error rate should be avoided at all costs in a clinical trial because it leads to increased chances of false positive results (Leyrat *et al.*, 2018).

The power to detect a statistically significant intervention effect, under the various alternative hypotheses (i.e., $\theta = 0.2$ or 0.3) was the same for all methods, regardless of the number of clusters and the level of clustering (**Figure 9.11**). When clustering was low (ICC=0.001), the methods had powers equal to or above the nominal 90% regardless of the number of clusters in the trial. When clustering increased to ICC = 0.01, the power of all the methods was still above the nominal 90% for higher effect size ($\theta = 0.3$), and for $\theta = 0.2$ the powers were within the acceptable range ≥ 80 provided that the number of clusters was not less than 12. This power started deteriorating as clustering increased and became worse (i.e., <50%) when clustering was high (ICC = 0.25) and effect size was lower ($\theta = 0.2$) regardless of the number of clusters. Although in most primary care or community based clinical trials clustering is often low or moderate – it is rare to observe high clustering (Adams et al., 2004; Twardella, Bruckner and Blettner, 2005; Offorha, Walters and Jacques, 2022).

Let's recall the research questions stipulated in this thesis, and briefly explain how they have been answered.

- What are the appropriate, and available statistical methods, in the literature for analysing outcome data from cRCTs?
    - This research question was addressed using a scoping review to synthesise evidence from the literature of cRCTs in Chapter 3. In the scoping review, systematic searching approach was applied to the electronic bibliographic databases of MEDLINE, EMBASE, PsycINFO, CINAHL, and SCOPUS. Twenty-seven unique methods that are appropriate for analysing outcome data from cRCTs were identified.
- What statistical methods are used in practice for analysing outcome data from cRCTs?
    - To answer this question, I search the online table of content of the NIHR Journal Library in a chronological order from 1st January 1997 to 15th July 2021, to identify reports containing cRCTs in Chapter 4. The results showed that 80% of the primary outcomes were analysed using GzLMM, 7% used RMRSE, and 6% used GEE1.
- What criteria should be used in deciding the appropriateness of the identified methods?

- The SISAQOL review (Coens *et al.*, 2020) was adopted as basis to answer this research question. The SISAQOL review is peculiar to statistical methods for analysing patient-reported outcome (PRO). Hence, it was adapted for identifying "appropriate" statistical method for analysing cluster RCTs. The criteria are, the ability of the methods to handle clustered data (correlated outcome data within a cluster), the ability to do a comparative test (statistical significance), the ability to produce interpretable treatment effects (and associated uncertainty), the ability to adjust for relevant covariates, the ability to handle missing data with the least restrictions.

- How well do the selected methods perform, when compared?
    - To answer this question, four methods were selected based on the findings of the reviews conducted in Chapters 3 and 4. The four methods – GzLMM, GEE1, GEE2, and QIF were technically described in Chapter 6. Practical guidance on how to implement the methods was provided in Chapter 7, by analysing four real-world datasets from cRCTs. GEE2 was dropped from further analysis, because it produces same parameter estimates as GEE1, except for ICC. To reach definite conclusions on the performance of the methods, a simulation study was conducted in Chapters 8 and 9. The three methods performed equivalently with respect to bias, empirical standard error, mean square error, and power. GzLMM performed better compared to GEE1 and QIF, with respect to coverage probability and Type I error rate, especially when the correlation is small.

The limitations of this simulation study are extensively discussed in Section 10.6 below. In brief, the simulation study was based on unadjusted model/analysis where no baseline covariate was adjusted for, except the treatment arm indicator variable. Adjustments for covariates that are correlated with the outcome may improve the efficiency of the parameter estimates (Samsa and Neely, 2018). There was no allowance for missing data, all analyses were based on a complete case sample. Ideally, in practice, missing individual outcome data or cluster dropout due to loss of follow-up or other reasons are not uncommon. Hence, missing data should be taken into account to minimise if not eliminate spurious findings (Ma *et al.*, 2013; Díaz-Ordaz *et al.*, 2014; Hossain *et al.*, 2017). Also, the simulation study assumed a common ICC across all clusters and treatment arms, this may not be true in practice.

10.5 Strengths and contributions of this study

This study contributes reasonably to the literature on cRCTs with clustering in treatment arms, especially in the aspect of design and analysis of cRCTs with a focus on primary care and community based trials. This study employed robust scientific approaches to achieve its aim and objectives. The primary purpose of this study was to evaluate the statistical methods for analysing outcome data from cRCTs, which was initiated by a methodological scoping review of cRCT literature in Chapter 3 to identify available and appropriate methods for analysing outcome data from cRCTs. The scoping review used robust methods to ensure that it is of high quality by developing a protocol, using a search strategy with the input of an Information Scientist, searching five popular online databases and grey literature, adopting the PRISMA-ScR guidelines, validating the search strategy, piloting the data collection form, submitting it as part of an article to a peer review journal.

The second review, a practice review conducted in Chapter 4 followed a similar process as the one mentioned above but was restricted to the NIHR Journals Library, however, some aspects of the reports have been published as result articles in other journals that are independent of the NIHR Journals Library. The search strategy involved a chronological search of the online table of contents of the five NIHR Journals Library, which was done to minimise the chances of missing reports of cRCT.

The analysis of real-world data in Chapter 7 was rich because it involved the analysis of four case studies with continuous and binary outcomes. The case studies covered a range of common scenarios in cRCTs, such as trials with few or many clusters, moderate or big cluster sizes, continuous or binary outcomes, and individual and/or cluster levels covariates. One of the strengths of the simulation study of Chapters 8 and 9 is that it was designed following the advice of Morris, White and Crowther (2019) and Burton et al. (2006). The levels of the parameters used were sourced from both previous and more recent studies. For example, the levels assumed for the ICC were obtained from a previous study (Ukoumunne *et al.*, 1999) and recent reviews (Adams et al., 2004; Twardella, Bruckner and Blettner, 2005; Offorha, Walters and Jacques, 2022) to cover a range of scenarios that are peculiar to most cRCT with primary care and community settings. The parameters varied are the number of independent clusters, cluster size, ICC, and effect size. The simulation study would have an impact in practice due to its strength of being generalisable to commonly used cRCT designs. The combination of the levels of the parameters generated

several realistic scenarios, and the methods used for analysis are among the most used methods that are necessary for practical recommendations for cRCTs analysis. The R codes and SAS syntax used to analyse the real-world data sets in Chapter 7 and pseudo simulated data in Chapter 9 are provided for transparency and reproducibility.

There is a growing interest in the comparative capability of GEE1 and QIF, few studies have assessed this in the context of cRCTs but were not comprehensive with regards to common cRCT designs (Yu, Li and Turner, 2020) and inclusion of more relevant performance measures (Westgate, 2012; Westgate and Braun, 2012, 2013). This study contributes to this regard, the claim that QIF is a promising alternative to GEE1 was comprehensively investigated using both example data analysis and a simulation study (for continuous outcome). Also, to the best of my knowledge, this research is the first to compare a cluster-specific model (GzLMM) against QIF. The results from this research were adequate to reach some definite conclusions in Section 10.8 below. Some of the findings from this study have been disseminated to the wider research community. Two manuscripts based on Chapters 3, 4, and 7 were submitted to Trials and BMC Medical Research Methodology peer-reviewed Journals. The practice review of Chapter 4 has been published (Offorha, Walters and Jacques, 2022) and the methodological review (Chapter 3) in conjunction with the analysis of case studies (Chapter 7) has been published in BMC Medical Research Methodology journal (Offorha, Walters and Jacques, 2023). Additionally, posters based on the two manuscripts have been presented at the 6th International Clinical Trials Methodology Conference (ICTMC) in 2022.

10.6 Thesis limitations

This study employed a formal and structured search of relevant literature to capture most of the relevant work conducted. However, this was not an exhaustive review of all work in this area. For instance, the systematic search strategy used in identifying relevant papers in Chapter 3 could have missed some relevant papers. This also applies to the practice review of Chapter 4 where the outline of each Journal of the NIHR Journals Library was searched chronologically. These limitations are not severe since the number of papers reviewed in both studies was reasonable and rich. Also, it is possible that using only reports from the NIHR Journals Library could cause publication bias. To reduce the chances of this occurring, all identified articles were included rather than a sample, and some aspects of most of the reports have been published in other journals indexed in secular databases, as results articles.

The results obtained and inferences made in Chapter 7 apply to trials with similar properties to the case studies. For example, in the four case studies used for the example data analysis, only binary and continuous endpoints were measured and analysed, small and negative ICCs were observed, and the clusters randomised and analysed were few or moderate in some cases. The analysis of these case studies used complete cases only, hence any data collected on patients for whom the outcome of interest was not recorded was ignored. Likely, this data limitation (i.e., missing data) might not result in adverse consequences since the proportions of missing data were small. Although, the other data limitation (i.e., a small number of clusters) might be. The small ICCs observed in the analysis does not constitute a severe limitation since they are reasonable values and are not uncommon in cRCTs with primary care and community settings (Adams et al., 2004; Offorha, Walters and Jacques, 2022).

The simulation study on the other hand made some assumptions that may not be the case, in reality. For instance, it was assumed that the numbers of clusters randomised and analysed were equal across treatment arms. Although in an ideal cRCT, the number of clusters should be the same between the treatment arms, usually they are not in practice. The plausible reasons for the varying numbers of clusters are the inability to recruit as planned and the loss to follow-up (Hayes and Moulton, 2009). The plan is to explore this limitation in future studies to assess the impact of having different numbers of clusters across the treatment arms.

Similarly, the cluster sizes randomised and analysed were assumed to be equal across clusters, which is more likely not to be the case in practice. It is not uncommon to observe cluster size imbalance in a trial because of differing recruitments across clusters, and loss to follow-up of individual participants. For example, in **Table 4.3** of the practice review, the average number of participants randomised was slightly different from the average number analysed (Section 4.4.4). If the variability in the cluster sizes is substantial the impact of the statistical methods would be severe, this has been demonstrated between GEE1 vs. QIF, where QIF was found to be more affected than GEE1(Westgate and Braun, 2012). Whereas unadjusted GzLMM was found to be severely affected by covariate imbalance, for the adjusted GzLMM the impact was negligible (Moerbeek and Van Schie, 2016).

The simulation study of Chapters 8 and 9, did not allow for baseline covariates in the models. There are several known benefits from adjusting for baseline covariates in an adjusted analysis, such as protection against imbalance in baseline participant prognostic covariates between the

groups (Kahan *et al.*, 2014), increased power and precision for linear models with a continuous response variable, especially when there is some association between the covariate(s) and outcome (Hauck, Anderson and Marcus, 1998; Kahan *et al.*, 2014; Samsa and Neely, 2018), to obtain an estimate of the intervention effect that has a closer individual-level interpretation, and to account for special features of the study design like stratification and subgroup consideration (Campbell *et al.*, 2012). Using simulation, adjusting for prognostic and non-prognostic covariates led to increased and reduced power, respectively (Kahan *et al.*, 2014). For nonlinear models with binary or categorical or count response variables, adjustment for relevant prognostic covariates would also lead to an increase in power, but a reduction in precision (i.e., larger SE) (Kahan *et al.*, 2014), which is offset by an increased estimated intervention effect (Hauck, Anderson and Marcus, 1998). In general, for a balanced trial with a continuous outcome, the unadjusted and adjusted analysis would produce equivalent estimates, but the adjusted analysis will be more precise, especially when the covariates are strongly correlated with the outcome (Samsa and Neely, 2018).

Furthermore, the simulation study of Chapters 8 and 9 was limited to only continuous outcomes. Continuous and binary outcomes are the most common types of outcomes in cRCT (Offorha, Walters and Jacques, 2022), but this study did not investigate the statistical methods against binary, count, time-to-event, and ordinal outcome types. When the response value is continuous, GzLMM reduces to LMM (i.e., equation (2.11)), it is well known that in general the interpretation of the parameter estimates of GzLMM (a cluster-specific model) is different from GEE1/QIF (a population average model), especially for non-linear models – models where the outcome type is either binary or discrete or time-to-event (Hubbard *et al.*, 2010). To be precise, the interpretation of the result of the treatment effect is mostly affected by the chosen estimand framework stipulated at the beginning of the trial. The stipulated estimand could be 1) the participant-average treatment effect (the average treatment effect across participants) or 2) the cluster-average treatment effect (the average treatment effect across clusters) (Kahan *et al.*, 2014). Decisions about how to analyse cluster randomised trials can unintentionally result in answering different questions about the interventions (i.e., target estimands). The 'participant-average treatment effect' answers the question 'How effective is the intervention for the average participant?' whereas the 'cluster-average treatment effect' answers the question 'How effective is the intervention for the average cluster?'. Therefore, efforts should be geared towards ensuring that the selected analytical approach in conjunction with the chosen estimator, targets the estimand that answers the stipulated research question (Kahan *et al.*, 2014). It is worth noting that GEE1 and QIF were developed for only data

types that are compatible with the linear exponential family distribution, of which time-to-event is not. Similarly, the simulation study assumed that no missing outcome data and no cluster drop out which may not be the case in practice. Missing outcome data are likely to have an impact on the methods differently, for instance, GEE1 was found to perform better than GzLMM in most scenarios with missing outcomes (Ma *et al.*, 2013).

Also, the simulation study assumed that the cluster random effects are Normally distributed, the simulated data and methods of analysis (especially the GzLMM with identity link) were based on this assumption (Chapters 8 and 9). However, cluster-level means can follow other types of distributions that are non-Normal (Zhang *et al.*, 2008; McCulloch and Neuhaus, 2011). The MLE as a common estimator of the GzLMM is known to be consistent and efficient provided that the distributional assumptions made are correct. One such assumption is that the cluster random effects are Normally distributed. Previous studies overstated the impact of misspecifying the distribution of the cluster random effects on the MLE (Agresti, Caffo and Ohman-Strickland, 2004; Litière, Alonso and Molenberghs, 2008). However, a recent study has shown that the MLE is quite robust to the impact of the misspecification of the cluster random effects in most situations considered previously, even when the cluster size is informative (Neuhaus and McCulloch, 2011). Similarly, GEE1 and QIF are not significantly affected since they are semi-parametric models, except of course when the correlation between outcomes is substantial (Liang and Zeger, 1986; Prentice, 1988; Prentice and Zhao, 1991; Qu, Lindsay and Bing, 2000; Yan and Fine, 2004; Crespi, Wong and Mishra, 2009). For the population average models, an exchangeable correlation structure was used. This means that the correlation between pairs of outcomes across clusters is equal. The exchangeable correlation structure is the most assumed in cRCTs, but in practice, the fixed exchangeable correlation structure may not be true in some cases. QIF performs slightly better than GEE1 in terms of precision when the correlation across the treatment arms is different, but this gain depreciates as the number of clusters reduces. In general, the relative performance of GEE1 and QIF in terms of precision is affected more by the number of clusters than the varying degrees of correlation across treatment arms (Westgate and Braun, 2012).

Additionally, another limitation of the thesis is that the results strictly speaking only apply when the outcome data is missing completely at random. With missing completely at random, all the methods are adequate, especially when the amount of missing data is small (say, <10%).

Lastly, the simulation study was based on a completely randomised parallel arm cRCT design. This is a type of cRCT design in which each cluster is randomised to one of the distinct treatment arms, meaning that each cluster has an equal probability of being randomised to a treatment arm. Other cRCT designs include factorial, crossover, and stepped wedge (Section 2.4). The compared methods are likely to behave differently when assessed using the above-stated cRCT designs. However, the scope of this thesis did not cover these other designs because of the limited time available.

Given these limitations, caution must be taken when generalising the results of this thesis to other studies and other types of cRCT designs. The combinations of the values of the parameter used in the simulation study were selected to represent the typical values observed in most cRCTs with primary care and community settings, but this is by no means exhaustive. There are many more possible combinations of the values of the parameters that were not considered in this thesis, for example, the ICC could assume any value between 0 and 1 but was categorised into four levels in the simulation study (ρ: 0.001, 0.01, 0.05, 0.25). Without categorising the ICC, it would be difficult or almost impossible to use it for the simulation study. Hence, extrapolating beyond the levels of the combination of the values of the parameters used in this work is not recommended. In summary, the conclusions reached in this study only applied to studies with similar features and should be done with caution.

10.7 Implications and Recommendations

In this Section, the impact of this research findings on the theories, policies, and practice via the design, analysis, and reporting of cRCTs are presented first followed by several recommendations at the end.

10.7.1 Design of cRCTs

Sample size calculation is an important aspect of the designing phase of a cRCT, and the multilevel nature of a cRCT design must be accounted for if an adequate sample size is desired (Campbell and Walters, 2014a). The ICC ρ is required to calculate the design effect multiplier factor used in inflating the calculated sample size for RCTs. Some of the approaches to obtaining the prior estimate of the ICC during the design phase are conducting an internal pilot study thereby estimating ICC from the data or using the observed ICC from previous studies similar to the

current one being designed. This research provides the median observed ICC as 0.02 (IQR: 0.001, 0.06) for primary care and community based cRCTs, this could serve as a starting point for researchers when planning a cRCT (see, **Table 4.4**).

However, the general use of the ICC suggested by this research should be cautioned. In practice, other factors are considered when estimating the ICC making its generalisability impossible, such as covariates adjusted for, type of data, and technique used to achieve balance (Ukoumunne *et al.*, 1999). And most importantly the uncertainty in a single ICC estimate should be considered in calculating the optimal sample size for a trial (Lewis and Julious, 2021; Sarkodie, Wason and Grayling, 2023). Hence, the ICC suggested by this research could serve as an initial value for planning a cRCT, especially for a simple cRCT.

10.7.2 Analysis of cRCTs

Analysis of outcome data from cRCTs is one of the major issues that have attracted a lot of attention over the years. As stated in Chapter 2 (Section 2.7), the appropriate sample size formula for cRCT should incorporate its multilevel nature due to randomisation at the cluster level (Campbell and Walters, 2014a). An appropriate analytical approach must also account for the multilevel nature of the outcome data as discussed in Section 2.8. The two broad approaches for analysing a cRCT are cluster-level (Section 2.8.2) and individual-participant level analyses (Section 2.8.3). This research focussed on individual-level analysis using individual-level continuous and binary outcome data. The models used under the individual-level analysis approach target two kinds of inferences 1) cluster specific and 2) population average, any chosen individual-level analytical method should match any of these two desired inferences mentioned above and should follow the research question.

In Chapter 2, an example for each of the analytical approaches at the individual level was discussed in detail, known as the classical methods for analysing outcome data from cRCTs. The methodological review of available and appropriate methods for analysing cRCTs in Chapter 3 identified 27 unique analytical methods beyond the classical ones. An instant decision could be made about the appropriate analytical method to use that matches the interest of the research using **Table 3.3** (see, Section 3.5.3). It has been decades since the first CONSORT 2010 statement: extension to cluster randomised trials (CONSORT-cRT) was stipulated to guide a researcher in planning, conducting, analysing, and reporting high-quality cRCTs, Item 12a concerns the

appropriate statistical method that should be used – one that takes clustering into account (Campbell and Walters, 2014a). The results of the practice review of Chapter 2 indicated that this is not the case in some trials. Some trials ignored clustering which could lead to spurious findings especially when the correlation in the data is substantial (Leyrat *et al.*, 2018; Thompson *et al.*, 2022). Again, this finding resulted in a series of checklists for determining if a method is appropriate for analysing outcome data from cRCTs (**Appendix 6.1**).

The review of Chapter 3 identified two emerging methods and that of Chapter 4 identified two most used classical methods, and the guidelines for assessing if a method is appropriate (see, Section 5.3) confirmed them to be appropriate methods. One of the two emerging methods called QIF is growing in popularity as an alternative to GEE1. Several comprehensive studies have confirmed this in the context of longitudinal design, but it has enjoyed little to no usage in cRCT. I theorised that it could be because its capabilities have not been comprehensively put to the test against GEE1 in the context of cRCT design to warrant its routine application. Results from Chapters 7 and 9 suggest that QIF has no special advantage over GEE1 in the context of cRCTs designs like simple primary care or community based trials.

10.7.3 Reporting of cRCTs

The dissemination phase of the findings of a trial is very important and should be considered throughout the trial. Publishing a trial report would necessitate transparency and reproducibility. The report of a trial should contain the necessary information to enable readers to understand the methodology and interpret the findings with ease. For cRCTs, it is highly recommended that researchers should endeavour to follow the reporting guidelines of CONSORT-cRCT as strictly as possible (Campbell et al., 2012; Offorha, Walters and Jacques, 2022). For instance, Item 17a of CONSORT-cRCT states "**results at cluster or individual participant level as applicable and a coefficient of intracluster correlation (ICC or k) for each primary outcome**". It recommended that the ICC or k should be reported for each primary outcome alongside the point estimate of the intervention effect and its CI. The results of **Table 4.4** showed that 42% (about 4 in 10) observed primary ICC was not reported. This finding suggests that the reporting quality of publicly funded cRCTs conducted in the UK is suboptimal. This result also depicts a high rate of non-adherence to Item 17a of CONSORT-cRCT, the success of a guideline is measured by the rate of its adherence (Ivers *et al.*, 2011).

10.7.4 Recommendations

The recommendations on the specific method to use in a particular scenario are summarised in **Table 10.1** and **Table 10.2** for coverage probability and Type I error rate respectively.

Design of cRCTs

1. Appropriate methods that account for clustering in sample size calculation must be used to obtain adequate samples to power a cRCT.
2. Varying values of the ICC should be used to re-estimate the sample size to assess the sensitivity of the chosen ICC value.

Analysis of cRCTs

1. Appropriate analytical methods that account for clustering in the outcome data from cRCTs should be used for analysis and reported accordingly, where not possible the reason should be clearly stated, and sensitivity analysis should be carried out.
2. Researchers should consider other factors such as type of outcome, expected size of the intervention treatment, number of clusters randomised, and primary research aim when selecting the analytical method for cRCT.
3. Emerging methods should be comprehensibly evaluated against established methods to assess their theoretical capabilities before recommending their routine usage (if necessary).
4. For cRCTs with few or moderate clusters randomised (< 50), small sample correction in conjunction with the chosen analytical method should be used to avoid spurious findings.
5. For cRCTs with continuous outcomes and expected low degree of correlation among outcomes (ICC ≤ 0.01) like in primary care settings, to adequately control the Type I error rate, GzLMM (with identity link and parameters estimated by MLE) is recommended as against GEE1 or QIF, but if $N < 50$, GzLMM with parameters estimated by REML in conjunction with appropriate small sample correction like "Satterthwaite" should be used. This is consistent with the recommendation of (Leyrat *et al.*, 2018). Furthermore, the analytical approach should be chosen with consideration for the scientific question of interest.

Reporting of cRCTs

1. Researchers should report each observed primary ICC and its precision which are important for planning future trials.
2. The method used to account for clustering should be reported, and any other sensitivity analysis method.
3. The precision of the observed ICC should be added to Item 17a of CONSORT-cRCT.
4. A comprehensive evaluation of adherence to the reporting guidelines of CONSORT-cRCT should be conducted.

Table 10.1 Recommended method to achieve approximately 95% coverage probability for the confidence interval of the treatment effect*§ in different scenarios.

ICC		Analytical strategy		
	N	**GzLMM♀**	**GEE1**	**QIF**
Lower ICC (e.g., ≤ 0.001)	< 50	✓	×	×
	50 – 120	✓	✓	✓
Low ICC ($0.001 < \rho \leq 0.01$)	< 50	×	×	×
	50 – 120	×	×	×
Moderate to high ICC (e.g., >0.01)	< 50	×	×	×
	50 – 120	✓	✓	✓

ICC = Intracluster correlation coefficient, N is the number of clusters, GzLMM = Generalized linear mixed model, GEE1 = First-order generalized estimating equations, GEE1 = First-order generalized estimating equations, QIF = Quadratic inference function.

*: "Recommended" indicates the method that achieved approximately the 95% coverage probability of the intervention effect estimate. Other factors should be considered in choosing the appropriate method such as the target of inference and adjusted analysis.

§: The 95% coverage probability with associated CI is given as 94% and 96%.

♀: for $N < 50$, REML with appropriate small sample correction like Satterthwaite should be used.

Table 10.2 Recommended method to maintain 5% nominal Type I error rate[§] in different scenarios.

		Analytical strategy		
ICC	***N***	**GzLMM**	**GEE1**	**QIF**
Lower ICC (e.g., ≤ 0.001)	≤ 20	×	×	×
	> 20	✓	×	×
Low ICC (0.001< ρ ≤ 0.01)	≤ 20	×	×	×
	> 20	×	×	×
Moderate to high ICC (e.g., ≥ 0.01)	≤ 20	×	×	×
	> 20	×	×	×

ICC = Intracluster correlation coefficient, N is the number of clusters, GzLMM = Generalized linear mixed model, GEE1 = First-order generalized estimating equations, GEE1 = First-order generalized estimating equations, QIF = Quadratic inference function.

§: "Recommended" indicates the method that controlled the Type I error rate below the nominal and achieved the highest power. Other factors should be considered in choosing the appropriate method such as the target of inference and adjusted analysis.

10.8 Conclusions

In this research, I have employed rigorous approaches to achieving the aim and objectives set out. Recall, that the primary aim of this research was to evaluate the statistical performance of the statistical methods for analysing outcome data from cRCTs, and these methods were identified following the reviews conducted in Chapters 3 and 4. The methods include two existing (GzLMM and GEE1) and two emerging methods (GEE2 and QIF), it has always been a practice by methodologists to understand the behaviours of different methods in different scenarios to recommend the appropriate one to use in specific situations or in general. I have used both real-world outcome data from four cRCTs and simulated continuous outcome data to address the primary research questions concerning the relative performance of the analytical methods for cRCTs. In no order the following conclusions are deduced from the findings in this research, they are:

1. Empirical results from the real-world data analysis showed that in most cases where clustering is low the methods are equivalent, in terms of the point estimate of treatment effect, its SE, CI, and P-value, and when they differ it is mostly QIF that differs. For

instance, a few or moderate numbers of clusters have severe adverse impact on the estimators of the population average models which is more severe on QIF compared to the others – GEE1 and GEE2. Hence, trials with clusters ≤ 40 must apply small sample size correction if QIF is to be used.

2. Also, GEE1 and GEE2 are equivalent but differ only in their parameter estimate for the ICC. If the association parameter or correlation heterogeneity is not of primary interest in research, GEE1 is recommended because it is easier to implement and interpret, otherwise, GEE2 should be used. Furthermore, it is common knowledge that in most cRCTs with primary care and community settings, the mean parameter (intervention effect) is of primary interest and GEE1 is the most used analytical method. In general, I recommend that current practice should be maintained.
3. The above conclusions informed on the basis of the results from the analysis of the four real datasets are generalisable to studies with similar characteristics as the four analysed cRCTs.
4. From the simulation study results with continuous outcome data obtained, GzLMM (with identity link and parameters estimates by MLE) outperformed GEE1 and QIF in terms of the 95% coverage of the confidence interval of the treatment effect estimate and Type I error rate for a situation where clustering is low but suffer from the inability to converge to a solution in a few cases. For the estimators of the population average models – GEE1 and QIF, results showed that they are equivalent. Hence, QIF has no advantage over GEE1 in simple cRCTs with primary care and community settings and this opposes the claim in the literature of longitudinal study. It is worth noting that the conclusions above (for the simulation study) are within the parameters of the simulations.
5. Previous and recent reviews have identified that GzLMM and GEE1 are the most used methods in cRCTs. In general, results from this research have shown them to be stable and performed equivalently to the two emerging methods – GEE2 and QIF within the parameters of the real and simulated datasets used in this research. Therefore, I recommend that current practice should be maintained – GzLMM and GEE1 should be used for analysing outcome data from cRCTs with consideration about the research question and intended inference.

10.9 Issues for future research

Although this study used robust methods to achieve its aim and objectives, there are potential issues to research further in the future. For instance, the simulation study conducted to evaluate the properties of the statistical methods was based on continuous outcome data. Although, studies have shown that two out of the three compared methods, GzLMM and GEE1, behave almost equivalently in similar simulation studies with continuous (Leyrat *et al.*, 2018) or binary (Thompson *et al.*, 2022) outcomes. Nonetheless, further investigation is required within the context of cRCTs to evaluate the relative performance of QIF against the other two methods when binary outcome is measured. Note that few studies comparing GEE1 and QIF with binary outcomes exist in the literature, their primary focus was on the investigation of the impact of small number of clusters, and the relative precision of the methods (Westgate, 2012; Westgate and Braun, 2012, 2013).

The parameters used in the simulation study were assumed to be fixed across clusters or treatment arms, depicting an ideal trial (Chapter 7). Future studies are required to evaluate the performance of the methods when these parameters are allowed to vary across clusters or treatment arms. Common causes of this variability in levels of parameters are attributable to staggering recruitment, loss to follow-up, and shared factors within clusters by the participants. Westgate (2012), Westgate and Braun (2012) and Westgate and Braun (2013) in their studies varied the ICC ρ and numbers of clusters N across treatment arms and cluster sizes n_i across clusters with a primary focus on the efficiency of GEE1 vs QIF. Future work could extend this assessment to other performance measures like bias, Type I error rate, power, and coverage probability of the confidence interval of the treatment effect estimate.

The review in Chapter 4 showed that missing data are not adequately handled in most cases, of the 92% of the trials that reported observing missing outcome data, only 8% used recommended analytical methods to handle the missing data. Missing data are inevitable in cRCTs and occur through three mechanisms which are missing completely at random (MCR), missing at random (MAR), and missing not at random (MNAR), most methods are only compatible with MCR and MNAR. Missing data induces variability in the levels of factors that affect the performance of statistical methods, especially their efficiency, multiple imputation is a common method for resolving this problem. It would be reasonable to investigate the impact of missing data on the relative performance of the methods.

One of the main findings of this research was that the quality of reporting cRCTs was suboptimal with regards to the observed ICC (Offorha, Walters and Jacques, 2022), it has been over a decade since a comprehensive study was conducted to assess the adherence to CONSORT extension to cluster randomised trials (Ivers *et al.*, 2011). Future studies could focus on updating the results of (Ivers *et al.*, 2011) to determine if adherence to CONSORT guidelines and the reporting quality of cRCTs have improved after about 12 years.

10.10 Summary

This chapter consolidated the findings of previous chapters to reach definite conclusions and make recommendations. The reviews of Chapters 3 and 4 enabled the identification of potential methods to evaluate, and research gaps to address. Two classical methods, GzLMM and GEE1, were identified as the most used methods in Chapters 3 (and 4) and described in Chapters 2 and 6 while two emerging methods, GEE2 and QIF, were identified in Chapter 3 and described in Chapter 6.

Given the gaps in knowledge identified in Chapters 3 and 4, a series of research questions, aim and objectives were generated in Chapter 5. To achieve the primary aim set out in this research, the results from analysing four cRCT real-world datasets are presented in Chapter 7 while that of simulated datasets is presented in Chapters 8 and 9. This Chapter summarised the main findings in previous Chapters in Section 10.3, and these findings were compared to those of other studies in Section 10.4, the strengths and contributions of this research to the cRCT literature are stated in Section 10.5, regardless of the limitations faced in this research (Section 10.6) some definite implications, recommendations and conclusions were reached (Section 10.7 and 10.8).

The scope of this research is limited to completely randomised parallel group randomised cRCTs and the simulation study focused on continuous outcome data, due to these limitations, other issues that could be investigated further are explained in Section 10.8. In summary, for continuous outcome data, when the number of clusters $N > 40$ and the correlation among outcomes is low (ICC ≤ 0.001, GzLMM (with identity link and parameters estimated by MLE) is recommended when $N \leq 40$, GzLMM (with identity link and parameters estimated by REML) with appropriate small sample correction like "Satterthwaite" is recommended for low to moderate clustering (i.e., ICC ≤ 0.05, and most especially if the primary scientific question is targeting the intervention effect at the individua-level. If the targeted inference is towards the population from where the

trial clusters were sampled, then GEE1 with appropriate small sample size correction like "Fay and Graubard" is recommended.

Appendices

Appendix 1

A review of statistical methods used in practice for analysing cRCTs (Chapter 4)

Table S4.1 Data collection tool

Item	Description
Auditor	Who reviewed and extracted the data
URL	URL for NIHR journal page
Year	NIHR Journal Year (on the website)
Month_Pub	NIHR Journal Month published (on the website)
Volume	NIHR Journal Volume
Issue	NIHR Journal Issue
Journal	NIHR Journal Name
Study_Name	Acronym or short title
Lead_Author	Surname and initials
Pub_Year	Year published (may not be the same as the YEAR) (on PDF)
ISRCTN	Trial registration number
Trial Design	Trial Design Type
Arms	Number of treatment arms
Clinical area	Illness/disease area being trialled
Setting	(1) Hospital (2) General Practice (3) Mixed (4) Community (5) Others
CTU	Clinical Trial Unit Involvement in the Trial, Yes/No
Original target	Original Sample Size Target
Revised target	If Original Target was Revised, Y/N
Final target	The Sample Size Target if Original Target was Revised
Extension to recruitment	Extension of Recruitment Period, Y/N
Intervention type	(1) Drugs (2) Therapy (3) Surgery (4) Complex Intervention (5) Others
Control type	Placebo or Active
Active specify	Active Control Type
Patient blinded	Yes or No
Pilot	Yes or No
Geographical region	(1) Multiple Regions (2) Regional
Recruitment centre type	
No_FUs	Number of Follow-ups
Primary_FU	Length of Primary Follow-up (in months)
Time of longest primary follow-up (in months)	The Numeric Value
Final_FU	Final Follow-Up (in months)
Final_FU_Months	The Numeric Value
Screened	The Number of Screened Participants
Eligible	The Number of Eligible Participants
Declined_Consent	The Number of Participants who Declined to Participate
Subjects_Recruited	The Number of Participants Recruited and Randomised

Item	Description
Level of clustering	Numeric value
Unit of randomisation	
Clusters_Randomised	The Number of Clusters Randomised
Target no .of clusters	The Planned Number of Clusters
Max_time (months)	
Original cluster target met	Yes or No
Original cluster target 80% met	Yes or No
Trial hypothesis	Superiority or Non-inferiority or Equivalence
Type of trial	Feasibility or Pilot or Main cRCT
Cluster or Stepped wedge	
Type of follow-up	Closed Cohort, Cross-Sectional, Repeated Cross-Sectional, and Open Cohort
For Stepped Wedge cRCTs	
Number of Sequence/Steps	
Type of stepped wedge RCT	
No. of clusters randomised to each sequence	
No. of periods	
Duration of time between each sequence/step	
Sample size	
Primary outcome (Name)	
Data type of primary outcome	
Primary endpoint/timepoint	As clearly stated in the report or used in sample size calculation
Assumptions for samples size: fixed or variable cluster sizes	Equal or Unequal
Planned Number of clusters or subjects per cluster	
Planned variability in cluster sizes	Yes or No
Planned coefficient of variation	
planned ICC	
Planned Design effect	
Target effect size	Numeric value
Target effect size(type)	
Target power	
Target significance level	
Did sample size allow for attrition?	
If so what level(s) of attrition	Yes or No
Target no .of clusters	
Target no .of participants	
Level of statistical analysis for the primary endpoint	Individual-level or Cluster-level
Description of analysis of primary outcome	Brief Excerpt from the Original Published Report
Method of adjusting for clustering	
Specific statistical model	
If random effect model used; random component	The Random component of a GzLMM model
If marginal (GEE1) model is used; the correlation structure	Correlation structure of the working correlation
If marginal (GEE1) model is used; robust standard errors?	Yes or No
cluster accounted	Yes or No
Observed ICC	A numeric value for each primary outcome
ICC derivation	Covariate(s) adjusted or unadjusted
No. of clusters in analysis	In the primary analysis

Item	Description
No. of subjects in analysis	In the primary analysis
Observed effect and 95% CI	
Does the 95% CI for the observed effect include the target effect in the original sample size calculation	Yes or No
Missing data acknowledged	Yes or No
Number of Missing data	
Imputated missing data	Yes or No
Type of imputation of missing data	
No. of imputations	Number of Imputations
Interim analysis	Yes or No
Met original target	Yes or No
Met final target	Yes or No
Met 80% of the original target	Yes or No
Met 80% of the final target	Yes or No
Revised target	Yes or No
Revised target downward	Yes or No
Revised target upward	Yes or No
Revised target achieved at 100%	Yes or No
Revised target achieved at 80%	Yes or No

Appendix 2

Details of the cRCT reports retrieved from the NIHR Journals Library (Chapter 4)

Table S1 List of all included reports in the practice review

Study_name	URL	Year	Month	Volume	Issue	Journal	Lead_Author	Pub_Year	ISRCTN
the EPIC cluster RCT	Dementia Care Mapping to reduce agitation in care home residents with dementia: the EPIC cluster RCT	2020	MAR	24	16	HTA	Surr CA	2020	82288852
cluster RCT using electronic health records and cohort study	Electronically delivered interventions to reduce antibiotic prescribing for respiratory infections in primary care: cluster RCT using electronic health records and cohort study	2019	MAR	23	11	HTA	Gulliford MC	2019	95232781
Psychological treatment for insomnia in the regulation of long-term hypnotic drug use	Psychological treatment for insomnia in the regulation of long-term hypnotic drug use	2004	FEB	8	8	HTA	Morgan K	2004	NR
Improving the referral process for familial breast cancer genetic counselling:	Improving the referral process for familial breast cancer genetic counselling: findings of three randomised controlled trials of two interventions	2005	FEB	9	3	HTA	Wilson BJ	2005	NR
The PoNDER trial	Psychological interventions for postnatal depression: cluster randomised trial and economic evaluation. The PONDER trial	2009	JUN	13	30	HTA	Morell CJ	2009	92195776
The Screening for Haemoglobinopathies in First Trimester (SHIFT) trial	Antenatal screening for haemoglobinopathies in primary care: a cohort study and cluster randomised trial to inform a simulation model. The Screening for Haemoglobinopathies in First Trimester (SHIFT) trial	2010	APR	14	20	HTA	Dormandy E	2010	677850
the DEPICTED study	Development and evaluation by a cluster randomised trial of a psychosocial intervention in children and teenagers experiencing diabetes: the DEPICTED study	2011	AUG	15	29	HTA	Gregory JW	2011	61568050
the MINT study	Managing Injuries of the Neck Trial (MINT): a randomised controlled trial of treatments for whiplash injuries	2012	DEC	16	49	HTA	Lamb SE	2012	33302125
OPERA study	Exercise for depression in care home residents: a randomised controlled trial with cost-effectiveness analysis (OPERA)	2013	MAY	17	18	HTA	Underwood M	2013	43769277
A cluster randomised controlled trial of a manualised cognitive "behavioural anger management intervention delivered by	A cluster randomised controlled trial of a manualised cognitive behavioural anger management intervention delivered by supervised lay therapists to people with intellectual disabilities	2013	MAY	17	21	HTA	Willner P	2013	37509773

Study_name	URL	Year	Month	Volume	Issue	Journal	Lead_Author	Pub_Year	ISRCTN
supervised lay therapists to people with intellectual disabilities									
the TRACS trial	A cluster randomised controlled trial and economic evaluation of a structured training programme for caregivers of inpatients after stroke: the TRACS trial	2013	OCT	17	46	HTA	Forster A	2013	49208824
the clinical effectiveness and cost-effectiveness of classroom-based CBT in reducing symptoms of depression in high-risk adolescents	A cluster randomised controlled trial to determine the clinical effectiveness and cost-effectiveness of classroom-based cognitive-behavioural therapy (CBT) in reducing symptoms of depression in high-risk adolescents	2013	OCT	17	47	HTA	Stallard P	2013	19083628
the CASCADE study	Structured, intensive education maximising engagement, motivation and long-term change for children and young people with diabetes: a cluster randomised controlled trial with integral process and economic evaluation the CASCADE study	2014	MAR	18	20	HTA	Christie D	2014	52537669
Multicentre cluster randomised trial comparing a community group exercise programme and home-based exercise with usual care for people aged 65 years and over in primary care	Multi-centre cluster randomised trial comparing a community group exercise programme with home based exercise with usual care for people aged 65 and over in primary care	2014	AUG	18	49	HTA	Iliffe S	2014	43453770
the ESTEEM trial	The clinical effectiveness and cost-effectiveness of telephone triage for managing same-day consultation requests in general practice: a cluster randomised controlled trial comparing general practitioner-led and nurse-led management systems with usual care (the ESTEEM trial)	2015	FEB	19	13	HTA	Campbell JL	2015	20687662
the CADET study	Clinical effectiveness and cost-effectiveness of collaborative care for depression in UK primary care (CADET): a cluster randomised controlled trial	2016	FEB	20	14	HTA	Richards DA	2016	32829227
OTCH	An Occupational Therapy intervention for residents with stroke-related disabilities in UK Care Homes (OTCH): cluster randomised controlled trial with economic evaluation	2016	FEB	20	15	HTA	Sackley CM	2016	757750
STRATEGIC	A cluster randomised trial of strategies to increase cervical screening uptake at first invitation (STRATEGIC)	2016	SEP	20	68	HTA	Kitchener HC	2016	52303479
FiNRncial incentives to improve adherence to antipsychotic mainteNRnce medication in non-adherent patients:	Financial incentives to improve adherence to antipsychotic maintenance medication in non-adherent patients: a cluster randomised controlled trial	2016	SEP	20	70	HTA	Priebe S	2016	77769281
PLEASANT	PLEASANT: Preventing and Lessening Exacerbations of Asthma in School-age children Associated with a New Term a cluster randomised controlled trial and economic evaluation	2016	DEC	20	93	HTA	Julious SA	2016	3000938
PARAMEDIC	Prehospital randomised assessment of a mechanical compression device in out-of-hospital cardiac arrest (PARAMEDIC): a pragmatic, cluster randomised trial and economic evaluation	2017	MAR	21	11	HTA	Gates S	2017	8233942
SAFER 2	Support and Assessment for Fall Emergency Referrals (SAFER) 2: a cluster randomised trial and systematic review of clinical effectiveness and cost-effectiveness of new	2017	MAR	21	13	HTA	Snooks HA	2017	60481756

Study_name	URL	Year	Month	Volume	Issue	Journal	Lead_Author	Pub_Year	ISRCTN
	protocols for emergency ambulance paramedics to assess older people following a fall with referral to community-based care when appropriate								
the REPOSE trial	A cluster randomised trial, cost-effectiveness analysis and psychosocial evaluation of insulin pump therapy compared with multiple injections during flexible intensive insulin therapy for type 1 diabetes: the REPOSE Trial	2017	APR	21	20	HTA	Heller S	2017	61215213
the WAVES study	The West Midlands ActiVe lifestyle and healthy Eating in School children (WAVES) study: a cluster randomised controlled trial testing the clinical effectiveness and cost-effectiveness of a multifaceted obesity prevention intervention programme targeted at children aged 6-7 years	2018	FEB	22	8	HTA	Adab P	2018	97000586
the EpAID study	Training nurses in a competency framework to support adults with epilepsy and intellectual disability: the EpAID cluster RCT	2018	FEB	22	10	HTA	Ring H	2018	96895428
Positive behaviour support training for staff for treating challenging behaviour in people with intellectual disabilities	Positive behaviour support training for staff for treating challenging behaviour in people with intellectual disabilities: a cluster RCT	2018	MAR	22	15	HTA	Hassiotis A	2018	NA
the PACE-UP study	A pedometer-based walking intervention in 45- to 75-year-olds, with and without practice nurse support: the PACE-UP three-arm cluster RCT	2018	JUN	22	37	HTA	Harris T	2018	98538934
the IQuaD study	Improving the Quality of Dentistry (IQuaD): A cluster factorial randomised controlled trial comparing the effectiveness and cost-benefit of oral hygiene advice and/or periodontal instrumentation with routine care for the prevention and management of periodontal disease in dentate adults attending dental primary care	2018	JUL	22	38	HTA	Ramsay CR	2018	56465715
the SCIN cluster RCT	A behaviour change package to prevent hand dermatitis in nurses working in health care: the SCIN cluster RCT	2019	OCT	23	58	HTA	Madan I	2019	53303171
Does the Royal Horticultural Society Campaign for School Gardening increase intake of fruit and vegetables in children?	Does the Royal Horticultural Society Campaign for School Gardening increase intake of fruit and vegetables in children? Results from two randomised controlled trials.	2014	AUG	2	4	PHR	Christian M	2014	11396528
the FRIENDS study	A cluster randomised controlled trial comparing the effectiveness and cost-effectiveness of a school-based cognitive behavioural therapy programme (FRIENDS)in the reduction of anxiety and improvement in mood in children aged 9/10 years	2015	NOV	3	14	PHR	Stallard P	2015	23563048
Bristol Girls Dance Project: a cluster randomised controlled trial of an after-school dance programme to increase physical activity among 11- to 12-year-old girls	Bristol Girls Dance Project: a cluster randomised controlled trial of an after-school dance programme to increase physical activity among 11- to 12-year-old girls	2016	MAY	4	6	PHR	Jago R	2016	52882523
Active for Life Year 5:	Active for Life Year 5: a cluster randomised controlled trial of a primary school-based intervention to increase levels of physical activity, decrease sedentary behaviour and improve diet	2016	JUN	4	7	PHR	Lawlor DA	2016	50133740

Study_name	URL	Year	Month	Volume	Issue	Journal	Lead_Author	Pub_Year	ISRCTN
the STAMPP study	Steps Towards Alcohol Misuse Prevention Programme (STAMPP): a school- and community-based cluster randomised controlled trial	2017	APR	5	2	PHR	Sumnall H	2017	47028486
the HeLP study	Cluster randomised controlled trial and economic and process evaluation to determine the effectiveness and cost effectiveness of a novel intervention [Healthy Lifestyles Programme (HeLP)] to prevent obesity in school children	2018	JAN	6	1	PHR	Wyatt K	2018	15811706
A cluster randomised controlled trial and evaluation and cost-effectiveness aNRlysis of the Roots of Empathy schools-based programme for improving social and emotioNRl well-being outcomes among 8- to 9-year-olds in Northern Ireland	A cluster randomised controlled trial and evaluation and cost-effectiveness analysis of the Roots of Empathy schools-based programme for improving social and emotional well-being outcomes among 8- to 9-year-olds in Northern Ireland	2018	MAR	6	4	PHR	Connolly P	2018	7540423
The PATHS curriculum for promoting social and emotioNRl well-being among children aged 7â€“9 years	The PATHS curriculum for promoting social and emotional well-being among children aged 7 9 years: a cluster RCT	2018	AUG	6	10	PHR	Humphrey N	2018	85087674
A school-based intervention (â€˜Girls Activeâ€™) to increase physical activity levels among 11- to 14-year-old girls	A school-based intervention ('Girls Active') to increase physical activity levels among 11- to 14-year-old girls: cluster RCT	2019	FEB	7	5	PHR	Harrington DM	2019	10688342
the STARS cluster RCT	Training teachers in classroom management to improve mental health in primary school children: the STARS cluster RCT	2019	MAR	7	6	PHR	Ford T	2019	84130388
the Travel to Work cluster RCT	A workplace-based intervention to increase levels of daily physical activity: the Travel to Work cluster RCT	2019	MAY	7	11	PHR	Audrey S	2019	15009100
A loyalty scheme to encourage physical activity in office workers: a cluster RCT	A loyalty scheme to encourage physical activity in office workers: a cluster RCT	2019	AUG	7	15	PHR	Hunter RF	2019	17975376
the INCLUSIVE cluster RCT	Modifying the secondary school environment to reduce bullying and aggression: the INCLUSIVE cluster RCT	2019	OCT	7	18	PHR	Bonell C	2019	10751359
the PRIMASTIC stepped-wedge trial	Predictive risk stratification model: a randomised stepped-wedge trial in primary care (PRISMATIC)	2018	JAN	6	1	HSDR	Snooks H	2018	55538212
A randomised controlled trial to evaluate the impact of a human rights based approach to dementia care in inpatient ward and care home settings	A randomised controlled trial to evaluate the impact of a human rights based approach to dementia care in inpatient ward and care home settings.	2018	MAR	6	13	HSDR	Kinderman P	2018	94553028
A patient-centred intervention to improve the maNRgement of	A patient-centred intervention to improve the management of multimorbidity in general practice: the 3D RCT	2019	FEB	7	5	HSDR	Salisbury C	2019	6180958

Study_name	URL	Year	Month	Volume	Issue	Journal	Lead_Author	Pub_Year	ISRCTN
multimorbidity in general practice: the 3D RCT									
the POPPI feasibility study and cluster RCT	A nurse-led, preventive, psychological intervention to reduce PTSD symptom severity in critically ill patients: the POPPI feasibility study and cluster RCT	2019	AUG	7	30	HSDR	Mouncey PR	2019	53448131
the EPOCH stepped-wedge cluster RCT	A national quality improvement programme to improve survival after emergency abdominal surgery: the EPOCH stepped-wedge cluster RCT	2019	SEP	7	32	HSDR	Peden CJ	2019	80682973
the DAFNE study	Improving management of Type 1 diabetes in the UK: the Dose Adjustment for Normal Eating (DAFNE) programme as a research test-bed. A mixed method analysis of the barriers and facilitators to successful diabetes self management, a health economic analysis, a cluster RCT of different models of delivery of an educational intervention and the potential of insulin pumps and additional educator input to improve outcomes	2014	DEC	2	5	PGAR	Heller S	2014	61215213
the LoTS care research programme	Development and evaluation of interventions and tools to improve patient and carer centred outcomes in longer-term stroke care (LoTS care) and exploration of adjustment post stroke	2014	DEC	2	6	PGAR	Forster A	2014	67932305
A community-based primary prevention programme for type 2 diabetes mellitus integrating identification and lifestyle intervention for prevention	A community-based primary prevention programme for type 2 diabetes mellitus integrating identification and lifestyle intervention for prevention: a cluster randomised controlled trial	2017	JAN	5	2	PGAR	Davies MJ	2017	80605705
Testing innovative strategies to reduce the social gradient in the uptake of bowel cancer screening	Testing innovative strategies to reduce the social gradient in the uptake of bowel cancer screening: a programme of four qualitatively enhanced randomised controlled trials	2017	APR	5	8	PGAR	Raine R	2017	74121020
A randomised controlled trial, cost-effectiveness and process evaluation of the implementation of self-maNRgement for chronic gastrointestiNRl disorders in primary care, and linked projects on identification and risk assessment	A randomised controlled trial, cost-effectiveness and process evaluation of the implementation of self-management for chronic gastrointestinal disorders in primary care, and linked projects on identification and risk assessment	2018	MAR	6	1	PGAR	Thompson DG	2018	90940049
Optimal primary care maNRgement of clinical osteoarthritis and joint pain in older people	Optimal primary care management of clinical osteoarthritis and joint pain in older people: a mixed-methods programme of systematic reviews, observational and qualitative studies, and randomised controlled trials	2018	JUL	6	4	PGAR	Hay E	2018	40721988
the PERCEIVE programme	Patient involvement in improving the evidence base on mental health inpatient care: the PERCEIVE programme	2018	DEC	6	7	PGAR	Wykes T	2018	6545047
the Primrose research programme including cluster RCT	Primary care management of cardiovascular risk for people with severe mental illnesses: the Primrose research programme including cluster RCT	2019	APR	7	2	PGAR	Osborn D	2019	13762819
the EQUIP research programme including a cluster RCT	Training to enhance user and carer involvement in mental health-care planning: the EQUIP research programme including a cluster RCT	2019	OCT	7	9	PGAR	Lovell K	2019	16488358

Study_name	URL	Year	Month	Volume	Issue	Journal	Lead_Author	Pub_Year	ISRCTN
a research programme including the IMPaCT RCT	A health promotion intervention to improve lifestyle choices and health outcomes in people with psychosis: a research programme including the IMPaCT RCT	2020	JAN	8	1	PGAR	Gaughran F	2020	58667926
the EVIDEM study	Changing practice in dementia care in the community: developing and testing evidence-based interventions, from timely diagnosis to end of life (EVIDEM)	2015	APR	3	3	PGAR	Ilife S	2015	1423159
Improving patient safety through the involvement of patients:	Improving patient safety through the involvement of patients: development and evaluation of novel interventions to engage patients in preventing patient safety incidents and protecting them against unintended harm	2016	OCT	4	15	PGAR	Wright J	2016	7689702
Effective patientâ€“clinician interaction to improve treatment outcomes for patients with psychosis	Effective patient clinician interaction to improve treatment outcomes for patients with psychosis: a mixed-methods design	2017	FEB	5	6	PGAR	Priebe S	2017	34757603
the REAL study	The Rehabilitation Effectiveness for Activities for Life (REAL) study: a national programme of research into NHS inpatient mental health rehabilitation services across England	2017	MAR	5	7	PGAR	Killaspy H	2017	25898179
Keeping Children Safe	Keeping Children Safe: a multicentre programme of research to increase the evidence base for preventing unintentional injuries in the home in the under-fives	2017	JUL	5	14	PGAR	Kendrick D	2017	65067450
Challenge Demcare	Challenge Demcare: Management of challenging behaviour in dementia at home and in care homes	2017	AUG	5	15	PGAR	Moniz-Cook E	2017	2553381
Treatment of anorexia nervosa	Treatment of anorexia nervosa: a multimethod investigation translating experimental neuroscience into clinical practice	2017	AUG	5	16	PGAR	Schmidt U	2017	42594993
Crisis resolution teams for people experiencing mental health crises	Crisis resolution teams for people experiencing mental health crises: the CORE mixed-methods research programme including two RCTs	2019	APR	7	1	PGAR	Lloyd-Evans B	2019	47185233
Redesigning postnatal care	Redesigning postnatal care: a randomised controlled trial of protocol-based midwifery-led care focussed on individual women's physical and psychological health needs	2003	NOV	7	37	HTA	MacArthur C	2003	NA
Work package 5: randomised controlled trial and field testing of the WHELD programme in care homes	Improving mental health and reducing antipsychotic use in people with dementia in care homes: the WHELD research programme including two RCTs	2020	JUL	8	6	PGAR	Ballard C	2020	62237498
Management and control of tuberculosis control in socially complex groups	Management and control of tuberculosis control in socially complex groups: a research programme including three RCTs	2020	OCT	8	9	PGAR	Story A	2020	17270334
A randomised controlled trial of prehospital intravenous fluid replacement therapy in serious trauma	A randomised controlled trial of prehospital intravenous fluid replacement therapy in serious trauma	2000	NOV	4	31	HTA	Turner J	2000	NR
A randomised controlled trial to assess the impact of a package comprising a patient-orientated, evidence-based self-help	A randomised controlled trial to assess the impact of a package comprising a patient-orientated, evidence-based self-help guidebook and patient-centred consultations on disease management and satisfaction in inflammatory bowel disease	2003	OCT	7	28	HTA	Kennedy A	2003	NR

Study_name	URL	Year	Month	Volume	Issue	Journal	Lead_Author	Pub_Year	ISRCTN
guidebook and patient-centred consultations on disease management and satisfaction in inflammatory bowel disease									
The SAFE study	A randomised controlled trial and cost-effectiveness study of systematic screening (targeted and total population screening) versus routine practice for the detection of atrial fibrillation in people aged 65 and over: The SAFE study	2005	OCT	9	40	HTA	Hobbs R	2005	NR
the MAVARIC study	MAVARIC - a comparison of automation-assisted and manual cervical screening: a randomised controlled trial	2011	JAN	15	3	HTA	Kitchener HC	2011	66377374
(ADDITION-Europe) study	A randomised trial of the effect and cost-effectiveness of early intensive multifactorial therapy on 5-year cardiovascular outcomes in individuals with screen-detected type 2 diabetes: the Anglo Danish Dutch Study of Intensive Treatment in People with Screen-Detected Diabetes in Primary Care (ADDITION-Europe) study	2016	AUG	20	64	HTA	Simmons RK	2016	NR
the ReMemBrIn RCT	A group memory rehabilitation programme for people with traumatic brain injuries: the ReMemBrIn RCT	2019	APR	23	16	HTA	das Nair R	2019	65792154
The LEGS study	Understanding causes of and developing effective interventions for schizophrenia and other psychoses	2016	MAR	4	2	PGAR	Perez J	2016	70185866
the ASPIRE research programme	Developing and evaluating packages to support implementation of quality indicators in general practice: the ASPIRE research programme, including two cluster RCTs	2020	APR	8	4	PGAR	Foy R	2020	91989345
the Bridge-it RCT	Provision of the progestogen-only pill by community pharmacies as bridging contraception for women receiving emergency contraception: the Bridge-it RCT	2021	MAY	25	27	HTA	Cameron ST	2021	70616901
the PreFIT three-arm cluster RCT	Fall prevention interventions in primary care to reduce fractures and falls in people aged 70 years and over: the PreFIT three-arm cluster RCT	2021	MAY	25	34	HTA	Bruce J	2021	71002650
the GoActive cluster RCT	A school-based, peer-led programme to increase physical activity among 13- to 14-year-old adolescents: the GoActive cluster RCT	2021	MAY	9	6	PHR	Corder KL	2021	31583496

Appendix 3

A review of statistical methods used in practice for analysing cRCTs (Chapter 4)

Offorha et al. Trials (2022) 23:115
https://doi.org/10.1186/s13063-022-06025-1

Trials

REVIEW Open Access

Statistical analysis of publicly funded cluster randomised controlled trials: a review of the National Institute for Health Research Journals Library

Bright C. Offorha*, Stephen J. Walters and Richard M. Jacques

Abstract

Background: In cluster randomised controlled trials (cRCTs), groups of individuals (rather than individuals) are randomised to minimise the risk of contamination and/or efficiently use limited resources or solve logistic and administrative problems. A major concern in the primary analysis of cRCT is the use of appropriate statistical methods to account for correlation among outcomes from a particular group/cluster. This review aimed to investigate the statistical methods used in practice for analysing the primary outcomes in publicly funded cluster randomised controlled trials, adherence to the CONSORT (Consolidated Standards of Reporting Trials) reporting guidelines for cRCTs and the recruitment abilities of the cluster trials design.

Methods: We manually searched the United Kingdom's National Institute for Health Research (NIHR) online Journals Library, from 1 January 1997 to 15 July 2021 chronologically for reports of cRCTs. Information on the statistical methods used in the primary analyses was extracted. One reviewer conducted the search and extraction while the two other independent reviewers supervised and validated 25% of the total trials reviewed.

Results: A total of 1942 reports, published online in the NIHR Journals Library were screened for eligibility, 118 reports of cRCTs met the initial inclusion criteria, of these 79 reports containing the results of 86 trials with 100 primary outcomes analysed were finally included. Two primary outcomes were analysed at the cluster-level using a generalized linear model. At the individual-level, the generalized linear mixed model was the most used statistical method (80%, 80/100), followed by regression with robust standard errors (7%) then generalized estimating equations (6%). Ninety-five percent (95/100) of the primary outcomes in the trials were analysed with appropriate statistical methods that accounted for clustering while 5% were not. The mean observed intracluster correlation coefficient (ICC) was 0.06 (SD, 0.12; range, − 0.02 to 0.63), and the median value was 0.02 (IQR, 0.001–0.060), although 42% of the observed ICCs for the analysed primary outcomes were not reported.

* Correspondence: bcofforha1@sheffield.ac.uk
School of Health and Related Research, The University of Sheffield, Sheffield, UK

Conclusions: In practice, most of the publicly funded cluster trials adjusted for clustering using appropriate statistical method(s), with most of the primary analyses done at the individual level using generalized linear mixed models. However, the inadequate analysis and poor reporting of cluster trials published in the UK is still happening in recent times, despite the availability of the CONSORT reporting guidelines for cluster trials published over a decade ago.

Keywords: Intracluster correlation coefficient, Cluster randomised controlled trials, Clustering, CONSORT, Statistical methods, Recruitment

Background

Randomised controlled trials (RCTs) are the gold standard in medical and public health research when assessing the safety, clinical and cost-effectiveness of new drugs, new health technologies and new social interventions [1]. Conventionally, in RCTs, individuals are randomised to the experimental arms using either a randomisation or minimisation technique to ensure random allocation and balance in participants characteristics across the experimental arms.

Individually randomised controlled trials (iRCTs) are common, but in practice, this trial design may suffer from the potential contamination of outcomes from participants in the trial. Contamination could occur when participants in proximity are randomised to different experimental arms, there are chances that they will share their experiences of the trial which in turn may influence their outcomes. The cluster randomised controlled trial (cRCT) design can be used to minimise the risks posed by contamination [2, 3].Other rationales for using a cRCT design are maximisation of limited resources, problems with logistics, and administrative convenience [2].

A cRCT is potentially a more powerful design in handling the above-named issues, with groups of individuals (rather than individuals) randomly allocated to the experimental arms, resulting in outcome data that is clustered. Clustered data can also arise from repeated measurements over time on the same individuals in a longitudinal study. Going forward, for simplicity we have interchangeably used "cluster trials" to mean cRCTs.

In cluster trials, outcomes from a cluster/group tend to be more similar than outcomes from any other randomly selected cluster/group. This similarity (or correlation) of outcomes within a cluster is also known as the intracluster correlation. This correlation or non-independence of outcomes violates the assumptions of standard statistical methods used for assessing the effectiveness of an intervention to control, such as *t-test, F-test, chi-square test* or statistical regression methods used when researchers are also interested in adjusting for the effects of covariates and confounders, such as *linear regression, Poisson regression* and *logistic regression*. Standard statistical methods assume that the outcomes from participants in a trial are independent, most of the time this assumption does not hold in cluster trials. Ignoring the dependence among outcomes in the same cluster may lead to reduced standard errors which means—an increased value of the test statistic, smaller *P*-values and narrower confidence intervals which could increase the risk of false-positive results [1, 3, 4].Campbell and Walters [1] grouped the recognised statistical methods for analysing cluster trials into four broad approaches: (1) cluster-level analysis—using aggregate summary measures for each cluster, (2) individual-level analysis—using regression models with robust standard errors, (3) individual-level analysis—using generalized linear mixed models (random effects models), and (4) individual-level analysis—using a generalized linear model with generalized estimating equations (GEE) to estimate the model coefficients. These broad groupings relate to the way the statistical methods account for correlation among outcomes from the same cluster. The primary objective of this review is to investigate the use of these statistical methods in practice, with a focus on their prevalence.

The Consolidated Standards of Reporting Trials (CONSORT) statement was first published in 1996 to guide the reporting of iRCTs [5]. The extension of the CONSORT statement to cover cluster trials was first suggested in 2001 [6] and was then extended in 2004 [7], based on the revision of the CONSORT statement in 2001. There were still inadequacies in the reporting of iRCTs; hence, in 2010, the previous version of 2001 was updated [9]. The 2012 extension to cover cluster trials was based on this updated CONSORT 2010 statement [8]. These guidelines are meant to aid researchers in the planning, conducting, analysing and reporting of cluster trials to reduce the problems occurring from the poor reporting of cRCTs. Most of the information extracted from each trial reviewed in this study is based on this CONSORT statements extended for cluster trials.

Adherence to the CONSORT reporting guidelines for cluster trials and its impact on the quality of reporting cluster trials has attracted the interest of researchers since it was published [10–12]. The adherence to different aspects of the CONSORT statement for cluster trials is usually of interest to researchers, for example a review found that though some aspect of treatment compliance

by the participants in the studies are reported, but in general, comprehensive reporting of treatment compliance by participants is poor and inadequate [12]. Another review concluded that despite the availability of the CONSORT reporting guidelines for cluster trials, the reporting of all aspects of sample size calculation was inadequate [11]. Ivers et al. [10] went a step further and investigated adherence to all the new items included in the CONSORT extension for cluster trials; they found that improvement was only evident in few aspects, while in general, the adherence to the CONSORT statement extension for cluster trials was inadequate. The success of any guideline can be measured by the rate of its implementation in practice [13].

One of the justifications for conducting this study was to contribute to the debate in the literature on the adherence to the CONSORT reporting guidelines extension for cluster trials; our focus is on the aspect of the reporting quality of the intracluster correlation coefficient in the cluster trials reviewed. It is justifiable to investigate how well the extended CONSORT reporting guidelines for cluster trials is been implemented in practice, with the aim of recommending how to improve the quality of reporting cluster trials (if necessary).

Established in 2006, the National Institute for Health Research (NIHR) is now the largest funder of public health and social research in England. The NIHR publishes its commissioned research in the online open access NIHR Journals Library which consists of five journals: *Public Health Research* (PHR; https://www.journalslibrary.nihr.ac.uk/phr/#/), *Health Services and Delivery Research* (HSDR; https://www.journalslibrary.nihr.ac.uk/hsdr/#/), *Efficacy and Mechanism Evaluation* (EME; https://www.journalslibrary.nihr.ac.uk/eme/#/), *Programme Grants for Applied Research* (PGfAR; https://www.journalslibrary.nihr.ac.uk/pgfar/#/) and *Health Technology Assessment* (HTA; https://www.journalslibrary.nihr.ac.uk/hta/#/). In 2019/2020, the NIHR awarded over £250 million to fund 310 research projects. The NIHR Health Technology Assessment (HTA) programme received the highest amount of about £96.1 million [14].

This review aimed to investigate the prevalence and appropriateness of the statistical methods considered, in the planning and the analyses of cluster trials in practice for publicly funded trials, to evaluate the adherence by researchers to the reporting guidelines stipulated in the CONSORT 2010 statement for cluster trials and the recruitment abilities of cluster randomised controlled trials.

Methods

Search strategy

We manually searched through the online table-of-contents of each of the five NIHR journals, from 1 January 1997 to 15 July 2021 chronologically. The title and abstract of each report were screened to identify if a cluster randomised controlled trial was reported in it. If the title and abstract did not provide sufficient information to determine whether a cluster trial was reported, we had to read through the introduction and methodology chapters of the report to decide if the report should be included.

Trial identification

To identify reports to be included in this review, we followed the procedure described in the "Search strategy" subsection. Apart from the HTA Journal that published its first volume in 1997, the other four journals are recent editions to the NIHR Journals Library. The HSDR, PGfAR and PHR journals published their first volume in 2013 while EME published its first volume in 2014. A search through the NIHR HTA archive from 1 January 1997 to 15 July 2021 showed that the first report of a cluster randomised controlled trial was published in 2000 [15]. However, choosing 1997 as the starting point enabled us to assess the adherence to the CONSORT reporting guidelines before and after the publication of the CONSORT 2010 statement extension for cluster trials. Our interest was solely on trials in which groups of individuals was the unit of randomisation.

One researcher (BCO) conducted the search and extraction of the information while two other independent reviewers (SJW and RMJ) supervised and validated a sample (25%) of the total trials reviewed. If the inclusion of a report was in doubt, this was discussed by all three reviewers until a consensus was reached. The cRCT reports were obtained from the NIHR Journals Library website (https://www.journalslibrary.nihr.ac.uk/#/ date last accessed 9 August 2021) along with any previously published trial paper, protocol paper or trial protocol, where available. For trials that had a published International Standardised Randomised Controlled Trial Number (ISRCTN) number, this was used to check the ISRCTN register of clinical trials for any additional information, a trial website or any previously unobtainable trial reports (cf. http://www.isrctn.com/). The trial reports published in the NIHR Journals Library were was used as the main resource where there were discrepancies in reporting. January 1997 was chosen as a start date for the review as this was the date of publication of the first report in the NIHR Journals Library (in the NIHR HTA Journal).

Eligibility criteria

For a study to be eligible, it must be a cluster randomised controlled trial (involving the randomisation of groups of individuals) or stepped wedge cRCT published in any of the five online NIHR Journals library, from 1

January 1997 to 15 July 2021. Reports on all other study designs were excluded. Pilot and/or feasibility cRCTs were excluded as these have separate specific design and analysis issues including outcomes, sample size and statistical analysis and reporting. Full texts of identified reports were retrieved for further assessment.

Patient and public involvement

Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

Data extraction

Once the NIHR Journals Library reports on cluster trials have been selected for inclusion, necessary information was extracted, using a standardised and piloted data extraction form. When the information of interest was not found, this was indicated with "Not Reported (NR)"; NR indicates that the author(s) did not consider or make use of the method/item of interest or might have used or considered the method/item of interest but did not report it.

The relevant information was extracted and stored in an Excel spreadsheet for further analysis. The information obtained was informed by the review of Walters et al .[16] and the relevant components for cRCTs as stipulated in the CONSORT 2004 statement and its subsequent update. These are the details of the article, information on sample size calculation, recruitment, follow-up, details on clustering, allocation, design/type of trial, primary outcome, primary analysis and results. An additional file presents the list and description (where necessary) of all the items extracted (see Additional file 1). The extracted information was analysed and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [17] guidelines where applicable (see Additional file 2 for the populated PRISMA checklist). In this review, the main outcome was the statistical methods used for analysing the primary outcome(s) of the cRCTs.

Analysis

During the review, we identified that several of the individual reports in the NIHR Journals Library reported the results of two or more separate cRCTs [18–22], as well as the results for two or more primary outcomes per trial [21, 23–31].

Descriptive statistics using frequencies and percentages were generated for the levels of all the categorical characteristics of the trials reviewed, while mean, standard deviation, range, median and interquartile range were obtained for continuous outcomes. All analysis was done using an Excel spreadsheet (Microsoft® Excel for Mac, version 16.51) and R studio (Version 1.4.1717).

Results

Trial characteristics

Reports were extracted from the five online NIHR Journals Library published from 1 January 1997 to 15 July 2021. In total, 1942 reports were screened for eligibility, 118 cRCTs reports met the initial inclusion criteria and 3 of the reports were stepped wedge cRCTs [29, 32, 33]. Two reports were excluded because their trials were stopped due to poor recruitment, and only qualitative findings were thereby reported [34, 35. Thirty-seven other pilot/feasibility cRCTs were further excluded. Seventy-nine reports containing the results of 86 cluster trials were included. Five reports contained the results of multiple cluster trials (4 reports of 2 cRCTs each and 1 report included 4 cRCTs) 19–23. A total of 100 primary outcomes (11 trials in 10 reports had multiple primary outcomes) were assessed in this review. The search and selection processes are presented in Fig. 1. The list and URL of all included reports are available in a separate (Additional file 3).

Table 1 summarises the characteristics of the 86 trials included in this review. Most of the trials reviewed were conducted in different regions but solely within the United Kingdom (UK) except for Simmons et al. [35] which involved other European locations. The trials design used was mostly a parallel-group cluster trial that involved a direct comparison between intervention and control experimental arms (85%, 73/86), and this was mostly done using two experimental arms for comparison (80%) (Table 1).

Statistical methods used for analysing cluster trials

Of the 100 primary outcomes reported in the 86 trials, the data type of most of the primary outcomes was continuous (63%, 63/100), followed by binary outcomes (28%), and then counts (5%), time-to-event [33, 36] and percentage [37, 38] were the least (2%, respectively). In the description of the statistical analysis of the primary outcomes of the cRCTs, a variety of phrases were used to describe the multilevel regression methods used to account for clustering, such as *generalized linear mixed-model*, *two-level hierarchical model*, *mixed-effects*, *multilevel regression* and *two-level heteroscedastic linear regression model*; hence, we used a generic name "generalized linear mixed model (GLMM)" to cover all the multilevel regression methods.

Of the 100 analysed primary outcomes in the trials, 80% (80/100) used a GLMM to account for clustering, 7% used regression methods with robust standard errors and 6% used generalized estimating equations (GEE) to estimate the regression coefficients of the models. Most

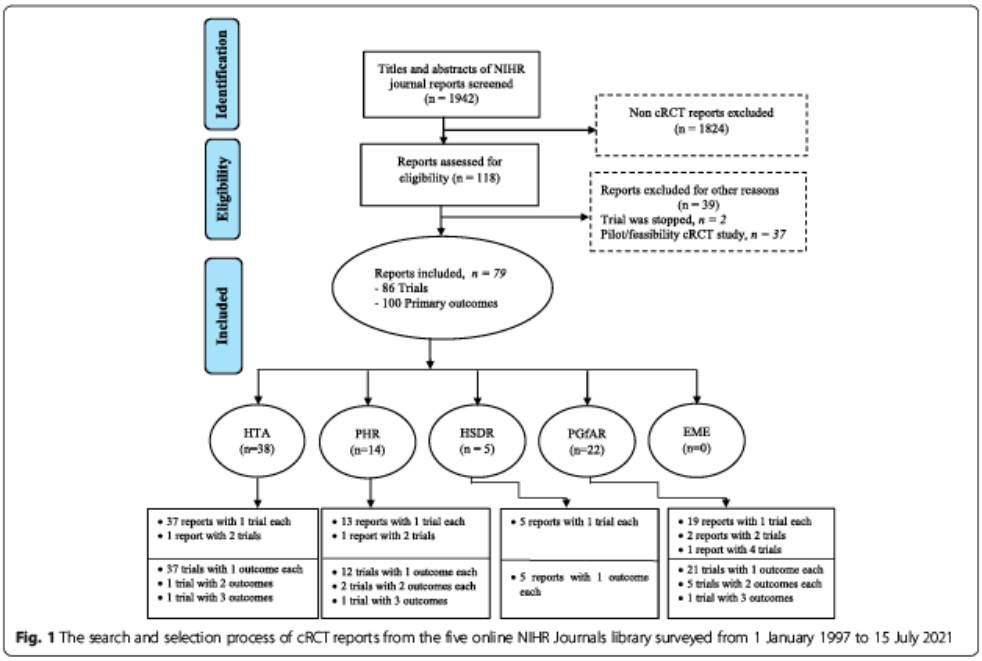


Fig. 1 The search and selection process of cRCT reports from the five online NIHR Journals library surveyed from 1 January 1997 to 15 July 2021

of these analyses were carried out using individual participant outcomes. Only 2 trials used aggregated cluster level outcomes as data points in their primary analyses [31, 38]. The different statistical methods used to account for the clustering of outcomes at the analysis phase are presented in Table 2.

Overall, 95% of the primary analyses used recognised statistical methods to adjust for clustering in their analyses, 5% did not; they ignored clustering and used standard statistical methods such as the *chi-square test, standard linear, logistic* and *Poisson regressions* [15, 29, 39–41]. Continuous outcomes were dichotomised in some studies to enable the use of logistic regression. The trial hypothesis was "superiority" in all the cluster trials except for Heller et al. [42] which was a non-inferiority trial. Table 2 also shows that most trials recruited and followed up the cohort of participants until the end of the trial; this often leads to missing data due to loss to follow-up (88%, 76/86).

Although 92% of the trials acknowledged the occurrence of missing data, most of them went ahead to analyse only complete cases (84%). Imputation of missing outcome data was done for just 16% of the trials reviewed [20, 28, 30, 43–53].

Planned recruitment targets of participants and clusters

Recruitment characteristics are summarised in Table 3, with 67% (58/86) of cRCTs achieving their planned final individual participant recruitment target and 87% of the trials achieving ≥ 80% of their final individual participant recruitment target; this indicates successful recruitment to final targeted sample size for most of the cluster trials. This also applies to the original cluster recruitment target, with 89% of the trials successfully recruiting (and randomising) ≥ 80% of their original targeted number of clusters.

Cluster and sample size characteristics

In Table 4, the cluster and sample size characteristics of the included trials are summarised and presented. The design effect if not reported was calculated using the formula, $1 + (m - 1) \times ICC$ or $1 + [(CV^2 + 1)m - 1] \times ICC$ for equal and unequal/varying cluster sizes respectively, where CV is the coefficient of variation, and m is the average cluster size. This is possible if the ICC and cluster size were reported. The median number of clusters randomised was 44 (IQR, 25–74), the minimum was 7 clusters randomised [42], and the maximum was 922 clusters randomised, in a trial of which households were the clusters [55]. A reasonable proportion of the

Table 1 Characteristics of cluster randomised controlled trials published in the NIHR Journals Library, from 1 January 1997 to 15 July 2021

Characteristic	*n*	%
NIHR journal the cRCT was reported in (*N* = 79[+])		
HTA	38	48
PHR	14	18
HSDR	5	6
PGAR	22	28
EME	0	0
Trial design (*N* = 86)		
Parallel	73	85
Factorial	7	8
Cross-over	2	2
Others*	4	5
Number of trial arms (*N* = 86)		
2	69	80
3	10	12
2 × 2	4	5
2 × 2 × 2	2	2
2 × 6	1	1
Clinical area (*N* = 86)		
Cancer/oncology	8	9
Mental health (including neurosciences/psychiatry/psychology/dementia)	21	25
Orthopaedics/rheumatology/musculoskeletal (including back pain)	2	2
Obstetrics and gynaecology	2	2
Primary care	6	7
Cardiovascular	1	1
Gastrointestinal	2	2
Respiratory	1	1
Stroke	4	5
Diabetes	6	7
Dermatology (including ulcers)	1	1
Others[†]	32	37
Setting (*N* = 86)		
Hospital	4	5
General practice	25	29
Mixed	3	3
Community	3	3
Others[‡]	51	59
Level of clustering (*N* = 86)		
2	85	99
3	1	1
Trial registration (*N* = 86)		
ISRCTN	78	91
NTC	2	2
Not reported	6	7

Table 1 Characteristics of cluster randomised controlled trials published in the NIHR Journals Library, from 1 January 1997 to 15 July 2021 *(Continued)*

Characteristic	*n*	%
Type of intervention (*N* = 86)		
Therapy	8	9
Behaviour change technique	4	5
Complex intervention	17	20
Education	12	14
Exercise	3	3
Information and communication technology	3	3
Medical device	2	2
Screening	2	2
Training	17	20
Others[§]	18	21
Type of control (*N* = 86)		
Active	86	100
Are patient blinded (*N* = 86)		
Yes	8	9
No	78	91
Any form of a pilot study[a] (*N* = 86)		
Yes	72	84
No	14	16
Geographical region (*N* = 86)		
Multiple regions	54	63
Regional	32	37

[a]These are internal pilot studies carried out within the main trials; they are different from the external pilot/feasibility studies mentioned initially in text
[*]79, the total number of journal reports included, which reported the results of 86 cRCTs (79 reports included the results of 86 cRCTs)
[*]Partial factorial and step-wedged trials
[†]Insomnia, paediatrics, youth bullying and other aggressive behaviours, traumatic brain injury, autism spectrum disorders, prehospital emergency care, obesity, epilepsy, oral health, end of life care, children fruit and vegetable intake, alcohol abuse, physical activity, psychosocial work environments, relationship and sexuality education, illicit drug use, smoking prevention, social and emotional wellbeing of children, dating and relationship violence, emergency admission to hospital, care for older people, multimorbidity, abdominal surgery, care of people with long-term conditions, care planning in secondary care mental health services and psychosis, eating disorder, injuries in under-fives children, patient involvement in safety, psychosis, care planning in secondary care mental health services
[‡]Care homes, nursing homes, clinics, NHS trust, residential services, stroke rehab unit, children centre, paediatrics diabetes clinic, schools, ambulances services, dental practice, stroke services
[§]Telephone triage, strategies to increase screening, financial incentive, invitation letter, leaflet, behavioural approaches, questionnaire, redesigned care model, health promotion, operational protocol, implementation package, time

randomised clusters were retained throughout the follow-up period, with a median of 43 clusters (IQR, 25–69) included in the analysis which is quite close to the number of clusters randomised. Also, for the number of subjects recruited/randomised, the median was 1184 (IQR, 597–3653), while the median number of subjects included in the analyses was 870 (IQR, 441–2356).

In the planning stage, 38% (33/86) of the planned ICCs used in the sample size calculations fell in the 0.03–0.05 range. The median planned ICC in the sample size calculation was 0.05 (IQR, 0.026–0.07). The observed ICCs from the analysed primary outcomes in the trials has a median value of approximately 0.02 (IQR, 0.001–0.060) with most of the reported ICCs occurring in the – 0.02 to 0.02 range (Table 4). After excluding two trials that were analysed at the cluster-level, we found that 42% (42/100) of the observed ICC from the primary analyses of the primary outcomes were not reported. Thirty-one percent of the observed ICC was not reported before the publication of CONSORT 2010 statement compared to 44% after its publication (Table 4). One study carried out a pair matched randomisation using a minimisation technique; however, they analysed their primary outcomes at the individual-level [28]. Pair matching of clusters reduces the population heterogeneity at the cluster level which could result in a negligible ICC from the analysed primary outcome and also improve the statistical efficiency of the trial [8, 10].

Not reporting the observed ICC for the analysed primary outcomes contradicts the CONSORT 2010

Table 2 Characteristics of the determinants of (and) the statistical methods used for analysing the primary outcomes in cluster trials

Characteristics	*n*	%
Type of follow-up RCT (N = 86)		
Closed cohort follow-up	76	88
Open cohort follow-up	4	5
Cross-sectional	4	5
Repeated cross-sectional	2	2
Data type of primary outcome (N = 100)		
Continuous	63	63
Binary	28	28
Counts	5	5
Time to event	2	2
Percentage	2	2
Method of adjusting for clustering (N = 100)		
Cluster-level analysis:		
Standard generalized linear model	2	2
Individual-level analysis:		
Generalized linear mixed model	80	80
Robust standard errors	7	7
Generalized estimating equations	6	6
Clustering not accounted for:		
Statistical hypothesis test statistic—chi-square	1	1
Standard generalized linear model	4	4
Specific statistical model (N = 100)		
Linear regression	57	57
Logistic regression	25	25
Analysis of covariance	6	6
Relative sensitivity	1	1
Negative binomial regression	2	2
Analysis of proportions	1	1
Cox Proportional Hazards model	2	2
Poisson regression	4	4
Weibull regression model	1	1
Chi-square test	1	1
Random component of GLMM (N = 80)		
Random intercept	76	95
Shared frailty	1	1
Random intercept and slope (repeated measures)	3	4
Correlation structure in GEE (N = 6)		
Exchangeable correlation	5	83
Correlation structure not reported	1	17

N = total number of trials; *n* = counts observed in each level of a category; RCT = randomised controlled trial; GLMM = generalized linear mixed model; GEE = generalized estimating equations. Not reported means that the information of interest was not considered and/or provided in the trial

reporting guidelines for cluster trials, which recommends that authors should report "a coefficient of intracluster correlation (ICC or k) for each primary outcome". The minimum observed ICC value appears to be an outlier (– 0.02) and was found in Heller et al. [42].

Figure 2 shows the trend and comparison of the practice of not reporting the observed ICCs for the analysed primary outcomes, before and after CONSORT 2010 statement. No observable trend appears to be present in Fig. 2. Before the publication of CONSORT 2010 guidelines for cRCT, the years that trials were carried out, 2003, 2005 and 2011 also recorded non-reporting of the observed ICCs for the analysed primary outcomes (20%, 100% and 50%, respectively). However, after the publication of the CONSORT 2010 statement, almost in each year aside 2013, some of the observed ICCs for the analysed primary outcomes were not reported, ranging from 28 to 90%. From Table 5, a higher proportion still did not report their observed ICCs from analysed primary outcomes after the publication of the CONSORT 2010 statement compared to the proportion that did not before its publication (44% vs 31%).

Discussion

This review was carried out to investigate the statistical methods used for analysing cluster randomised controlled trials in practice; to this end, we surveyed publicly funded cluster trials funded by the National Institute for Health Research.

Most of the trials used appropriate/recognised statistical methods to adjust for clustering in the main analyses of the primary outcomes from the trials (95%, 95/100). Few of the outcomes (and trials) 5% ignored clustering and used standard statistical methods that assumed independence among outcomes from participants in a cluster. This approach is not reccommended as it could lead to smaller standard errors and consequently, an increased value of the test statistic, smaller *P*-values, narrower confidence interval and possibly increase the type I error rate compared with the statistical methods that allow for clustering. If this happens to be the case, misleading conclusions and decisions will be made; this could have detrimental effects on public health.

The generalized linear mixed model (GLMM) was the most popular choice in adjusting for clustering and was more popular than GEE (80% vs 6%). For the GLMMs with two levels of clustering (trial participants nested within clusters), the cluster unit is usually incorporated as a random intercept to account for clustering. Where the primary outcome was measured more than once or the level of clustering is more than two levels, statistical models with random intercept and the random slope were used. Four trials that used the GEE method assumed an exchangeable working correlation structure in

Table 3 Planned participants and clusters recruitment to targets in cluster trials

Characteristics	*n*	%	Mean (SD)	Median	Range	IQR
Original individual participant target sample size (*N* = 84[b])						
≤ 300	11	13	10,035 (31357)	1250	136–250,000	550–4466
301–600	11	13				
601–900	13	15				
901–1200	3	4				
1201–1500	11	13				
1501–1800	3	4				
> 1800	29	35				
Not reported	3	3				
Final individual participant target sample size (*N* = 84[b])						
≤ 300	11	13	9372 (30173)	1212	136–250,000	534–4258
301–600	11	13				
601–900	14	17				
901–1200	5	6				
1201–1500	11	13				
1501–1800	3	4				
> 1800	27	32				
Not reported	2	2				
Original individual participant target sample size met (*N* = 86)						
Yes	57	66				
No, but >= 80% met	14	16				
No and < 80% met	9	11				
Final individual participant recruitment target met (*N* = 86)						
Yes	58	67				
No, but 80% >= of target met	17	20				
No and < 80% of target met	6	7				
Not reported	5	6				
Revised original individual participant target sample size (*N* = 86)						
Yes, upward	13	15				
Yes, downward	9	10				
Yes, direction not reported	4	4				
No	61	71				
Original cluster recruitment target met (*N* = 86)						
Yes	68	79				
No, but >= 80% met	9	11				
No, and < 80% met	1	1				
Not reported	8	9				

[b]Two studies were excluded because the original and final targets were expressed in person-years of observation and not specific number of participants [41, 54]. *N* = total number of trials; *n* = counts observed in each level of a variable; SD = standard deviation; IQR = interquartile range. Not reported means that the information of interest was not considered and/or provided in the trial

the primary analysis [18, 56–58], while 1 trial did not report the correlation structure that was assumed [42].

Fiero et al. [59] conducted a systematic review that focused more on the handling of missing data than on the statistical methods used for analysing cluster trials and found similar results to ours. They found that most of the trials analysed their primary outcome using GLMMs, and the cluster unit were modelled as the random intercept to account for clustering. Also, they found that all 14 (100%) of the trials that used GEE to account for

Table 4 Cluster and sample size characteristics of the trials included in the review

Characteristics	*n*	%	Mean (SD)	Median	Range	IQR
No. of clusters randomised (*N* = 86)						
4–10	2	2	77 (121)	44	7–922	25–74
11–20	11	13				
21–50	40	47				
51–100	21	24				
101–200	5	6				
> 200	7	8				
No. of clusters analysed (*N* = 86)						
0–10	2	2	76 (118)	43	7–864	25–69
11–20	12	14				
21–50	40	47				
51–100	21	24				
101–200	4	5				
> 200	7	8				
No. of subjects recruited (*N* = 84[b])						
≤ 300	7	8	15,348 (48315)	1184	141–265,434	597–3653
301–600	14	17				
601–900	11	13				
901–1200	9	11				
1201–1500	9	11				
1501–1800	3	4				
> 1800	29	34				
Not reported	2	2				
No. of subject analysed (*N* = 84[b])						
≤ 300	15	18	14,367 (48419)	870	42–264,325	441–2356
301–600	15	18				
601–900	13	15				
901–1200	5	6				
1201–1500	5	6				
1501–1800	2	2				
> 1800	25	30				
Not reported	4	5				
Planned ICC for sample size (*N* = 86)						
0.00–0.02	18	21	0.065 (0.082)	0.05	0.0002–0.5	0.0258–0.0700
> 0.02–0.05	33	38				
> 0.05–0.08	9	11				
> 0.08–0.11	8	9				
> 0.11–0.14	2	2				
> 0.14	6	7				
Not reported	10	12				
Planned design effect (*N* = 86)						
0.00–2.99	47	55	4.5 (8.90)	1.96	1.03–70.5	1.384–4.600
3.00–5.99	12	14				
6.00–8.99	10	12				

Table 4 Cluster and sample size characteristics of the trials included in the review *(Continued)*

Characteristics	n	%	Mean (SD)	Median	Range	IQR
9.00–11.99	1	1				
≥ 12	3	3				
Not reported	13	15				
Observed ICC of analysed primary outcome (*N* = 100)						
− 0.02 to 0.02	35	35	0.06 (0.12)	0.02	-0.02 to 0.63	0.0010–0.0600
> 0.02–0.07	9	9				
> 0.07–0.12	3	3				
> 0.12–0.17	6	6				
> 0.17–0.22	2	2				
> 0.22	3	3				
Not reported	42	42				

[b]Two trials were excluded because the analysed subjects were measured in person-years. *N* = total number of trials and/or primary outcomes; *n* = counts observed in each level of a category; SD = standard deviation. Not reported means that the information of interest was not considered and/or provided in the trial

clustering assumed an exchangeable correlation structure, which is similar to the findings of this current review (5/5, 100%; one study did not report their correlation structure [42]). Overall, they found that a lower proportion 79% (68/88) of the trials accounted for clustering compared to our review which observed a higher proportion 95% (95/100).

It is worth noting that while the use of appropriate statistical methods is high, none of the trials considered the recent potentially improved statistical methods developed in other study designs where clustered data do arise, such as the quadratic inference function (QIF), the alternating logistic regression (ALR) and the targeted maximum likelihood (tMLE). These recent methods are improvements over the standard GEE method for estimating the regression coefficients in the model [60]. The results of our study revealed that the number of clusters randomised in a cRCT could be as large as 922 in a study where the clusters were households [55] and as few as 7 clusters [42]. This result is different from the findings of Arnup et al. [61] where they focused on cluster randomised cross-over trials, one reason for choosing a cross-over design is if the number of the prospective clusters is small. In their study, the lowest number of clusters randomised was 2 while 25% of the number of clusters randomised was below 4.

In practice, active controls are mostly used when assessing the effect of non-pharmacological interventions (86/86, 100%). As revealed in our results, most times, it is impractical to conduct studies where the participants are blinded to the experimental arms they are allocated to. However, to some extent, masking is achieved by blinding either the person randomising the subjects, the assessor and/or the statistician that will analyse the data. To carry out a robust cluster trial, it is preferable to conduct an internal pilot/feasibility study (84%, 72/86) to assess the viability of the items/phases of the trial, such as the data collection tools, the understanding (and safety) and acceptance of the intervention by the participants and the ability to recruit to target before proceeding with the main trial.

Recruiting participants (for clusters, see Table 3) into a trial seems not to be a problem in cluster trials, particularly when compared to iRCTs (see Table 6). In 87% of the cluster trials, researchers were able to recruit ≥ 80% of their final planned participant recruitment targets. This result also applies to the number of clusters recruited/randomised, where 76% of the trials were able to recruit ≥ 80% of their planned clusters recruitment target (see Table 6).

In Table 6, we compared the ability of cRCTs and iRCTs to recruit to their target the number of

Table 5 Comparison of the non-adherence in the reporting of observed ICC for each primary outcome before and after CONSORT 2010 statement for cRCTs (published in Sept 2012)

	Year of publication		
	Before 1997–2012	After 2013–2021	All 1997–2021
Number of trials	11	75	86
Number of primary outcomes	13	87	100
Number of primary outcomes with the observed ICC not reported (%)	4 (31)	38 (44)	42(42)

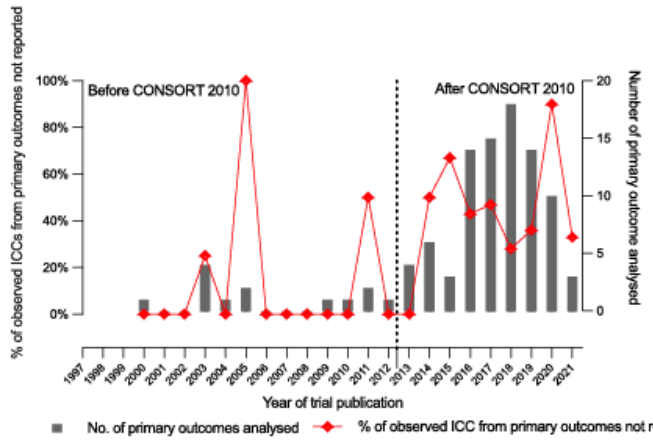


Fig. 2 Plot comparing the trend of not reporting the observed ICCs of analysed primary outcomes in cRCTs before and after CONSORT 2010 statement with the first published cRCT in NIHR Journals library recorded in 2000

participants. In terms of recruiting to 100% of the original participant target, cluster trials seem more successful than iRCTs (66% vs 55%). This is confirmed by the fact that in cluster trials, the originally planned sample sizes are rarely revised (24%) and tend to be revised upward (57%, 12/21) rather than downward (43%). When compared to iRCTs, the number (and percentage) of upward revisions were higher in cluster trials (57% vs 36%). Even with the most upward revisions, cluster trial recruitment periods are rarely extended to meet up with recruitment targets compared to iRCTs (13%, 11/86 vs 54%, 65/122).

We also found that in practice the completely randomised parallel-group cluster trial design is the most used design involving two experimental arms in its simplest form. This cluster design is easy to set up, implement and analyse. Our results indicated that all the trials reviewed, except one, were superiority trials involving contrasting experimental arms. For the sample size calculation, our results indicated that the median assumed ICC value, used in the calculation, was 0.05, while the average was 0.065. However, we found observed that the ICC assumed in the sample size calculation could be as low as 0.0002 (Table 4). Our results also indicated that a

Table 6 Comparing the ability to recruit to the target the number of participants between cRCTs and iRCTs using results of previous studies that reviewed iRCTs

Review	McDonald et al.	Sully et al.	Walters et al.	This study
Recruitment period	**1994–2002**	**2002–2008**	**2004–2016**	**1997–2021**
Number of trials in the study	***N* = 122 iRCTs**	***N* = 73 iRCTs**	***N* = 151 iRCTs**	***N* = 86 cRCTs**
Recruited 100% of original target	38 of 122 (31%)	40 of 73 (55%)	61 of 151 (40%)	**57 of 86 (66%)**
Original target was revised	42 of 122 (34%)	14 of 73 (19%)	52 of 151 (34%)	21[c] of 86 (24%)
Original target revised upward	6 of 42 (14%)	5 of 14 (36%)	11 of 52 (21%)	**12 of 21 (57%)**
Original target revised downward	36 of 42 (86%)	9 of 14 (64%)	41 of 52 (79%)	9 of 21 (43%)
Recruited 80% of original target	67 of 122 (55%)	57 of 73 (78%)	95 of 151 (63%)	**71 of 86 (83%)**
Recruited 100% of revised target	19 of 42 (45%)	10 of 14 (71%)	28 of 52 (54%)	**16 of 21 (76%)**
Recruited 80% of revised target	34 of 42 (80%)	13 of 14 (93%)	48 of 52 (92%)	**21 of 21 (100%)**
Extended their recruitment	65 of 122 (54%)	33 of 73 (45%)	49 of 151(32%)	**11 of 86 (13%)**

Source: Adapted (and modified) from Walters et al. [16]
[c]Was supposed to be 25 trials but 2 trials did not report their original target that was revised, and another two trials did not report their final revised target and the number of participants recruited respectively; they were excluded since comparison cannot be done

disappointing trend of not reporting the observed ICC for each primary outcome is happening. About 4 out of 10 of the observed ICCs from the analysed primary outcomes in cRCTs are not being reported. The implication of not reporting the ICC cannot be overemphasised; the ICC is an important item in designing/planning future cluster trials as it is needed for sample size calculation. It is reasonable to make it available for researchers planning to undertake similar study. The importance of reporting the ICC was reemphasised by the development of a framework specifying how and what should be reported in association with the ICC to facilitate understanding and the planning of future cluster trials [62].

Surprisingly, this occurs more in recent times despite the availability and publicity of the CONSORT 2010 statement extension for cluster trials, although 2005 had the highest percentage of this disappointing practice (100%), but with only two analysed primary outcomes recorded.

It is worth noting that this was also after the publication of the CONSORT 2004 statement [8] extension for cluster trials. Ivers et al. [11] assessed the impact of the CONSORT 2004 statement extension for cluster trials on quality of reporting and study methodology, and one of the criteria compared was the "reporting of an estimated intracluster correlation". They found only 18% of the 300 manuscripts reported an ICC estimate and 22% vs 14% before and after CONSORT 2004 statement respectively. This result indicated a decline in the practice of reporting indicated a decline in the practice of reporting the observed ICC which is similar to our current study. We found a 13% increase (change in non-adherence before to after CONSORT 2010) in non-adherence to the CONSORT reporting guidelines with regards to reporting the observed ICC for each primary outcome analysed, using CONSORT 2010 statement as the basis for comparison. CONSORT statements extensions for cluster trials are published to facilitate improved quality reporting of cluster trials. If used properly, they are supposed to help in the understanding, assessing and replicating of cluster trials by all stakeholders of clinical trials. Hence, all authors intending to write up the report for their cluster trial(s) should make good use of the updated CONSORT 2010 statement.

The observations made in this review is that in practice there are important issues in cRCTs that are still being ignored or handled inadequately or not reported.

Firstly, missing data is not adequately handled most of the time. The majority (79/86, 92%) of the studies reviewed acknowledged the existence of missing data, which is obvious due to inevitable loss to follow-up in a closed cohort follow-up study; however, the majority still went ahead to analyse only available observations (84%). To assess the robustness of the findings in the trials, especially when missing data was not technically handled, most researchers resorted to conducting sensitivity analysis. However, if they had handled the problem of missing data technically (e.g., using statistical imputations) in the original analysis, it could have improved the inferences made in the study.

Secondly, there appears to be a slow uptake of improved statistical methods developed in other study designs where clustered data can arise, such as the QIF, tMLE and ALR that are potentially better in certain situations than the popular statistical methods used currently for analysing cluster trials. It would be ideal if these methods are publicised by methodologists of cluster trials so that researchers can use them when necessary to make optimal inferences [60].

Limitation

We acknowledge that searching and retrieving cluster trial reports from one source could lead to publication bias. We optimized the study by including all cluster trials instead of a random sample, and the reports published in the NIHR Journals Library were also published independently as result articles in other journals; hence, reports included in this review represent a collection of articles from several journals independent of NIHR Journals Library.

Conclusion and recommendation

In practice, most of the publicly funded cluster trials adjusted for clustering using an appropriate/recognised statistical method with most of the primary analyses done at the individual level using generalized linear mixed models. However, the inadequate analysis and poor reporting of cluster trials, particularly not reporting the observed ICC for the analysed primary outcomes is still happening in recent times despite the availability of the CONSORT reporting guidelines extension for cluster trials published over a decade ago. One way of addressing this issue is to encourage journal editors and peer reviewers to insist that authors should adhere to CONSORT reporting guidelines for cluster trials when submitting their manuscripts. This review will serve as a reference tool in conducting systematic reviews of statistical methods used in practice and statistical methods available in the literature for analysing cluster trials.

Patient and public involvement Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

Abbreviations

cRCT: Cluster randomised controlled trial; CONSORT: Consolidated standards of reporting trials; NIHR: National Institute for health research; SD: Standard deviation; ICC: Intracluster correlation coefficient; IQR: Interquartile range; RCT: Randomised controlled trial; IRCT: Individually randomised controlled

trial; GEE: Generalized estimating equation; GLMM: Generalized linear mixed model; QIF: Quadratic inference function; tMLE: Targeted maximum likelihood; ALR: Alternating logistic regression; PHR: Public health research; HSDR: Health services and delivery research; EME: Efficacy and mechanism evaluation; PGfAR: Programme grants for applied research; HTA: Health technology assessment; NR: Not reported

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13063-022-06025-1.

Additional file 1. Data collection tool (NIHR).
Additional file 2. PRISMA 2020 checklist.
Additional file 3. List of included reports.

Authors' contributions

BCO conceptualised the idea for this review, prepared the data collection form, carried out the search and data extraction, analysed the data, wrote the first draft of the manuscript and revised and edited the manuscript. SJW and RMJ conceptualised the idea for this review, prepared the data collection tool, supervised and validated the search and extraction of the data, and critically revised and edited the manuscript. The authors read and approved the final manuscript.

Authors' information

BCO is a PhD candidate in Medical Statistics. SJW is a Professor of Medical Statistics. RMJ is a Senior lecturer in Medical Statistics.

Funding

Offorha is sponsored by the Nigerian Tertiary Education Trust Fund (TETFund). Prof. Walters, and Dr. Jacques, received funding across various projects by NIHR. Prof. Walters is a National Institute for Health Research (NIHR) Senior Investigator (NF-SI-0617-10012) supported by the NIHR for this research project. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care. These organisations had no role in the study design; in the collection, analysis, and interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

Availability of data and materials

The information extracted in this review is based on published trials in the NIHR Journals library. The data extracted from the NIHR Journals Library supporting the finding of this study is available upon reasonable request from the corresponding author.

Declarations

Ethics approval and consent to participate

The information extracted in this review is based on published NIHR trials where ethics approvals were obtained by the original trial teams. This review does not involve recruiting new participants or analysing individual participants, and the original participants cannot be identified from this review.

Consent for publication

This review does not involve recruiting new participants or analysing individual participants' data. Individual informed consent was obtained to take part in the original trials by primary investigators.

Competing interests

The PhD study of Offorha is financially sponsored by the Nigerian Tertiary Education Trust Fund (TETFund) (Grant No. TETF/ES/UNIW/UTURU/TSA/2019). Prof. Walters and Dr. Jacques received funding across various projects by NIHR. Prof. Walters is a National Institute for Health Research (NIHR) Senior Investigator (NF-SI-0617-10012) supported by the NIHR for this research project. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care. These organisations had no role in the study design; in the collection, analysis, and interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

Received: 19 September 2021 **Accepted:** 13 January 2022
Published online: 04 February 2022

References

1. Campbell MJ, Walters SJ. How to design, analyse and report cluster randomised trials in medicine and health related research: Wiley; 2014. https://doi.org/10.1002/9781118763452.
2. Christie J, OHalloran P, Stevenson M. Planning a cluster randomized controlled trial. Nurs Res. 2009;58(2):128–34. https://doi.org/10.1097/NNR.0b013e3181900cb5.
3. Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. Acad Emerg Med. 2002;9(4):330–41. https://doi.org/10.1197/aemj.9.4.330.
4. Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. BMC Med Res Methodol. 2004;4(1):2–7. https://doi.org/10.1186/1471-2288-4-21.
5. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. J Am Med Assoc. 1996;276(8):637–9. https://doi.org/10.1001/jama.276.8.637.
6. Elbourne DR, Campbell MK. D.R. E. Extending the CONSORT statement to cluster randomized trials: For discussion. Stat Med. 2001;20(3):489–96. https://doi.org/10.1002/1097-0258(20010215)20:3<489:AID-SIM806>3.0.CO;2-S
7. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. BMJ. 2004;328(7441):702 LP–708. https://doi.org/10.1136/bmj.328.7441.702.
8. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ Online. 2012;345(7881):1–21. https://doi.org/10.1136/bmje5661.
9. Schulz KF, Altman DG, Moher D, the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med. 2010;8(1):18. https://doi.org/10.1186/1741-7015-8-18.
10. 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 Online. 2011;343(7827):1–14. https://doi.org/10.1136/bmj.d5886.
11. 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.
12. Agbla SC, DiazOrdaz K. Reporting non-adherence in cluster randomised trials: a systematic review. Clin Trials. 2018;15(3):294–304. https://doi.org/10.1177/1740774518761666.
13. Gogtay NJ. Reporting of randomized controlled trials: will it ever improve? Perspect Clin Res. 2019;10(2):49–50. https://doi.org/10.4103/picr.PICR_11_19.
14. NIHR. National Institute for Health Research. Annu Rep. 2019;(December):1–50.
15. Turner J, Nicholl J, Webber L, Cox H, Dixon S, Yates D. A randomised controlled trial of prehospital intravenous fluid. Health Technol Assess Winch Engl. 2000;4(31):1–57.
16. Walters SJ, Dos Anjos Henriques-Cadby IB, Bortolami O, et al. Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom Health Technology Assessment Programme. BMJ Open. 2017;7(3):1–10. https://doi.org/10.1136/bmjopen-2016-015276.
17. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. The BMJ. 2021;372. https://doi.org/10.1136/bmj.n71.
18. Kitchener HC, Gittins M, Rivero-Arias O, et al. A cluster randomised trial of strategies to increase cervical screening uptake at first invitation (STRATEGI C). Health Technol Assess. 2016;20(68). https://doi.org/10.3310/hta20680.
19. Christian MS, EL Evans C, Cade JE. Does the Royal Horticultural Society Campaign for School Gardening increase intake of fruit and vegetables in children? Results from two randomised controlled trials. Public Health Res. 2014;2(4):1–162. https://doi.org/10.3310/phr02040.
20. Raine R, Atkin W, Von Wagner C, et al. Testing innovative strategies to reduce the social gradient in the uptake of bowel cancer screening: a programme of four qualitatively enhanced randomised controlled trials.

Programme Grants Appl Res. 2017;5(8):338–02. https://doi.org/10.3310/pgfar05080.
21. Foy R, Willis T, Glidewell L, et al. Developing and evaluating packages to support implementation of quality indicators in general practice: the ASPIRE research programme, including two cluster RCTs. Programme Grants Appl Res. 2020;8(4). https://doi.org/10.3310/pgfar08040.
22. Ballard C, Orrell M, Moniz-Cook E, et al. Improving mental health and reducing antipsychotic use in people with dementia in care homes: the WHELD research programme including two RCTs. Programme Grants Appl Res. 2020;8(6). https://doi.org/10.3310/pgfar08060.
23. Ramsay CR, Clarkson JE, Duncan A, Lamont TJ, Heasman PA, Boyers D, et al. Improving the quality of dentistry (IQuaD): a cluster factorial randomised controlled trial comparing the effectiveness and cost-benefit of oral hygiene advice and/or periodontal instrumentation with routine care for the prevention and management of perio. Health Technol Assess. 2018;22(38): vii–143. https://doi.org/10.3310/hta22380.
24. MacArthur C, Winter HR, Bick DE, Lilford, RJ, Lancashire RJ, Knowles H, et al. Redesigning postnatal care: a randomised controlled trial of protocol-based midwifery-led care focused on individual women's physical and psychological health needs.Health Technol Assess. 2003;7(37):1–98.
25. Lawlor DA, Kipping RR, Anderson EL, Howe LD, Chittleborough CR, Moure-Fernandez A, et al. Active for Life Year 5: a cluster randomised controlled trial of a primary school-based intervention to increase levels of physical activity, decrease sedentary behaviour and improve diet. Public Health Res. 2016;4(7):1–156. https://doi.org/10.3310/phr04070.
26. Sumnall H, Agus A, Cole J, Doherty P, Foxcroft D, Harvey S, et al. Steps Towards Alcohol Misuse Prevention Programme (STAMPP): a school- and community-based cluster randomised controlled trial. Public Health Res. 2017;5(2):1–154. https://doi.org/10.3310/phr05020.
27. Connolly P, Miller S, Kee F, Sloan S, Gildea A, McIntosh E, et al. A cluster randomised controlled trial and evaluation and cost-effectiveness analysis of the Roots of Empathy schools-based programme for improving social and emotional well-being outcomes among 8- to 9-year-olds in Northern Ireland. Public Health Res. 2018;6(4):1–108. https://doi.org/10.3310/phr06040.
28. Thompson DG, O'Brien S, Kennedy A, Rogers A, Whorwell P, Lovell K, et al. A randomised controlled trial, cost-effectiveness and process evaluation of the implementation of self-management for chronic gastrointestinal disorders in primary care, and linked projects on identification and risk assessment. Programme Grants Appl Res. 2018;6(1):1–154. https://doi.org/10.3310/pgfar06010.
29. Wykes T, Csipke E, Rose D, Craig T, McCrone P, Williams P, et al. Patient involvement in improving the evidence base on mental health inpatient care: the PERCEIVE programme. Programme Grants Appl Res. 2018;6(7):1–182. https://doi.org/10.3310/pgfar06070.
30. Gaughran F, Stahl D, Patel A, et al. A health promotion intervention to improve lifestyle choices and health outcomes in people with psychosis: a research programme including the IMPaCT RCT. Programme Grants Appl Res. 2020;8(1). https://doi.org/10.3310/pgfar08010.
31. Wright J, Lawton R, O'Hara J, Armitage G, Sheard L, Marsh C, et al. Improving patient safety through the involvement of patients: development and evaluation of novel interventions to engage patients in preventing patient safety incidents and protecting them against unintended harm. Programme Grants Appl Res. 2016;4(15):1–296. https://doi.org/10.3310/pgfar04150.
32. Snooks H, Bailey-Jones K, Burge-Jones D, Dale J, Davies J, Evans B, et al. Predictive risk stratification model: a randomised stepped-wedge trial in primary care (PRISMATIC). Health Serv Deliv Res. 2018;6(1):1–164. https://doi.org/10.3310/hsdr06010.
33. Peden CJ, Stephens T, Martin G, Kahan BC, Thomson A, Everingham K, et al. A national quality improvement programme to improve survival after emergency abdominal surgery: the EPOCH stepped-wedge cluster RCT. Health Serv Deliv Res. 2019;7(32):1–96. https://doi.org/10.3310/hsdr07320.
34. Speed C, Heaven B, Adamson A, et al. LIFELAX – diet and LIFEstyle versus LAXatives in the management of chronic constipation in older people : randomised controlled trial. 2010;14(52). https://doi.org/10.3310/hta14520.
35. Simmons RK, Borch-Johnsen K, Lauritzen T, Rutten GEHM, Sandbaek A, van den Donk M, et al. A randomised trial of the effect and cost-effectiveness of early intensive multifactorial therapy on 5-year cardiovascular outcomes in individuals with screen-detected type 2 diabetes: the Anglo–Danish–Dutch Study of Intensive treatment in people with scr. Health Technol Assess. 2016;20(64):1–86. https://doi.org/10.3310/hta20640.
36. Davies MJ, Gray LJ, Ahrabian D, Carey M, Farooqi A, Gray A, et al. A community-based primary prevention programme for type 2 diabetes mellitus integrating identification and lifestyle intervention for prevention: a cluster randomised controlled trial. Programme Grants Appl Res. 2017;5(2): 1–290. https://doi.org/10.3310/pgfar05020.
37. Priebe S, Bremner SA, Lauber C, Henderson C, Burns T. Financial incentives to improve adherence to antipsychotic maintenance medication in non-adherent patients: a cluster randomised controlled trial. Health Technol Assess. 2016;20(70):x–121. https://doi.org/10.3310/hta20700.
38. Cameron ST, Glasier A, McDaid L, Radley A, Patterson S, Baraitser P, et al. Provision of the progestogen-only pill by community pharmacies as bridging contraception for women receiving emergency contraception: the Bridge-it RCT. Health Technol Assess Winch Engl. 2021;25(27):1–92. https://doi.org/10.3310/hta25270.
39. Morgan K, Dixon S, Mathers N, Thompson J, Tomeny M. Psychological treatment for insomnia in the regulation of long-term hypnotic drug use. Health Technol Assess. 2004;8(8). https://doi.org/10.3310/hta8080.
40. Gates S, Lall R, Quinn T, Deakin CD, Cooke MW, Horton J, et al. Prehospital randomised assessment of a mechanical compression device in out-of-hospital cardiac arrest (PARAMEDIC): a pragmatic, cluster randomised trial and economic evaluation. Health Technol Assess. 2017;21(11):1–175. https://doi.org/10.3310/hta21110.
41. Perez J, Russo DA, Stochl J, et al. Understanding causes of and developing effective interventions for schizophrenia and other psychoses. Programme Grants Appl Res. 2016;4(2). https://doi.org/10.3310/pgfar04020.
42. Heller S, Lawton J, Amiel S, Cooke D, Mansell P, Brennan A, et al. Improving management of type 1 diabetes in the UK: the Dose Adjustment For Normal Eating (DAFNE) programme as a research test-bed. A mixed-method analysis of the barriers to and facilitators of successful diabetes self-management, a health economic analys. Programme Grants Appl Res. 2014; 2(5):1–188. https://doi.org/10.3310/pgfar02050.
43. Salisbury C, Man MS, Chaplin K, Mann C, Bower P, Brookes S, et al. A patient-centred intervention to improve the management of multimorbidity in general practice: the 3D RCT. Health Serv Deliv Res. 2019;7(5):1–238. https://doi.org/10.3310/hsdr07050.
44. Mouncey PR, Wade D, Richards-Belle A, Sadique Z, Wulff J, Grieve R, et al. A nurse-led, preventive, psychological intervention to reduce PTSD symptom severity in critically ill patients: the POPPI feasibility study and cluster RCT. Health Serv Deliv Res. 2019;7(30):1–174. https://doi.org/10.3310/hsdr07300.
45. Killaspy H, King M, Holloway F, Craig TJ, Cook S, Mundy T, et al. The Rehabilitation Effectiveness for Activities for Life (REAL) study: a national programme of research into NHS inpatient mental health rehabilitation services across England. Programme Grants Appl Res. 2017;5(7):1–284. https://doi.org/10.3310/pgfar05070.
46. Moniz-Cook E, Hart C, Woods B, Whitaker C, James I, Russell I, et al. Challenge Demcare: management of challenging behaviour in dementia at home and in care homes – development, evaluation and implementation of an online individualised intervention for care homes; and a cohort study of specialist community mental health car. Programme Grants Appl Res. 2017;5(15):1–290. https://doi.org/10.3310/pgfar05150.
47. Surr CA, Holloway I, Walwyn REA, Griffiths AW, Meads D, Kelley R, et al. Dementia care mapping™ to reduce agitation in care home residents with dementia: the epic cluster rct. Health Technol Assess. 2020;24(16):1–174. https://doi.org/10.3310/hta24160.
48. Lamb SE, Williams MA, Williamson EM, Gates S, Withers EJ, Mt-Isa S, et al. Managing injuries of the neck trial (minit): a randomised controlled trial of treatments for whiplash injuries. Health Technol Assess. 2012;16(49):1–141. https://doi.org/10.3310/hta16490.
49. Iliffe S, Kendrick D, Morris R, Masud T, Gage H, Skelton D, et al. Multicentre cluster randomised trial comparing a community group exercise programme and home-based exercise with usual care for people aged 65 years and over in primary care. Health Technol Assess. 2014;18(49):1–105. https://doi.org/10.3310/hta18490.
50. Snooks HA, Anthony R, Chatters R, Dale J, Fothergill R, Gaze S, et al. Support and assessment for fall emergency referrals (SAFER) 2: a cluster randomised trial and systematic review of clinical effectiveness and cost-effectiveness of new protocols for emergency ambulance paramedics to assess older people following a fall wi. Health Technol Assess. 2017;21(13):1–218. https://doi.org/10.3310/hta21130.
51. Heller S, White D, Lee E, Lawton J, Pollard D, Waugh N, et al. A cluster randomised trial, cost-effectiveness analysis and psychosocial evaluation of

insulin pump therapy compared with multiple injections during flexible intensive insulin therapy for type 1 diabetes: The REPOSE Trial. Health Technol Assess. 2017;21(20):1–277. https://doi.org/10.3310/hta21200.
52. Ring H, Howlett J, Pennington M, et al. Training nurses in a competency framework to support adults with epilepsy and intellectual disability: the EpAID cluster RCT. Health Technol Assess. 2018;22(10). https://doi.org/10.3310/hta22100.
53. Humphrey N, Hennessey A, Lendrum A, Wigelsworth M, Turner A, Panayiotou M, et al. The PATHS curriculum for promoting social and emotional well-being among children aged 7–9 years: a cluster RCT. Public Health Res. 2018;6(10):1–116. https://doi.org/10.3310/phr06100.
54. Gulliford MC, Juszczyk D, Prevost AT, Soames J, McDermott L, Sultana K, et al. Electronically delivered interventions to reduce antibiotic prescribing for respiratory infections in primary care: cluster RCT using electronic health records and cohort study. Health Technol Assess. 2019;23(11):11–70. https://doi.org/10.3310/hta23110.
55. Harris T, Kerry S, Victor C, Iliffe S, Ussher M, Fox-Rushby J, et al. A pedometer-based walking intervention in 45- to 75-year-olds, with and without practice nurse support: The PACE-UP three-arm cluster RCT. Health Technol Assess. 2018;22(37):1–273. https://doi.org/10.3310/hta22370.
56. Morrell CJ, Warner R, Slade P, et al. Psychological interventions for postnatal depression: cluster randomised trial and economic evaluation. The PoNDER trial. Health Technol Assess. 2009;13(30). https://doi.org/10.3310/hta13300.
57. Dormandy E, Bryan S, Gulliford MC, Roberts TE, Ades AE, Calnan M, et al. Antenatal screening for haemoglobinopathies in primary care: a cohort study and cluster randomised trial to inform a simulation model. The screening for haemoglobinopathies in first trimester (SHIFT) trial. Health Technol Assess. 2010;14(20):1–160. https://doi.org/10.3310/hta14200.
58. Harrington DM, Davies MJ, Bodicoat D, Charles JM, Chudasama YV, Gorely T, et al. A school-based intervention ('Girls Active') to increase physical activity levels among 11- to 14-year-old girls: cluster RCT. Public Health Res. 2019;7(5):1–162. https://doi.org/10.3310/phr07050.
59. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials. 2016;17(1):1–10. https://doi.org/10.1186/s13063-016-1201-z.
60. Turner EL, Li F, Gallis JA, Prague M, Murray DM. Review of recent methodological developments in group-randomized trials: Part 1 - Design. Am J Public Health. 2017;107(6):907–15. https://doi.org/10.2105/AJPH.2017.303706.
61. Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials. J Clin Epidemiol. 2016;74:40–50. https://doi.org/10.1016/j.jclinepi.2015.11.013.
62. Campbell MK, Grimshaw JM, Elbourne DR. Intracluster correlation coefficients in cluster randomized trials: empirical insights into how should they be reported. BMC Med Res Methodol. 2004;5(1):1–5. https://doi.org/10.1186/1471-2288-4-9.

Publisher's Note

Appendix 4

SAS Syntax and R codes for fitting the statistical models to the PoNDER trial dataset (Chapter 7)

```
*************************************************************;
*     SAS SYNTAX TO FIT GzLMM AND QIF (CONTINUOUS OUTCOME) ;
*                                                           ;
*************************************************************;

/* Read in PoNDER data from stored location*/
proc import out = work.ponder2
datafile    =    "C:\Users\cmp17bco\Desktop\PhD_YR2\Work    Package    1\WP_1_
SAS\DATA\ponder2.sav"
  dbms =spss replace;
  run;

/* FIT UNADJUSTED GzLMM-MLE MODEL USING GLMMIX PROC*/
 proc glimmix data=ponder2 method = quad(qpoints=10);
  class group2(ref="control") clusteri;
  model epds_6mo = group2 / dist=normal ddfm=bw cl;
random intercept / subject = clusteri s type=vc g;
run;

/* FIT ADJUSTED GzLMM-MLE   MODEL USING GLMMIX*/
  proc glimmix data=ponder2 method = quad(qpoints=10);
  class group2(ref="control") clusteri;
  model  epds_6mo  =  group2  epds_6we  alone  history  any_life/  dist=normal
ddfm=bw cl;
random intercept / subject = clusteri s type=vc g;
run;

*Upload QIF macro from stored location;
options mautosource sasautos = ('C:\Users\cmp17bco\Desktop\WP_1_SAS\MACROS',
sasautos);
%qif

/* FIT UNADJUSTED QIF MODEL USING QIF MACRO*/
%qif(data=ponder2,
            yvar=epds_6mo,
            xvar=group2,
            id=clusteri, dist=normal, corr=exch,
            print=y, outpar=par1, outqif=qif1, outcov=cov1, outres=resid1);
run;

/* FITTING ADJUSTED QIF MODEL USING QIF MACRO*/
%qif(data=ponder2,
            yvar=epds_6mo,
            xvar=group2 epds_6we alone history any_life,
            id=clusteri, dist=normal, corr=exch,
            print=y, outpar=par2, outqif=qif2, outcov=cov2, outres=resid2);
run;
```

```
*************************************************************;
*     SAS SYNTAX TO FIT GzLMM AND QIF (BINARY OUTCOME)      ;
*                                                           ;
*************************************************************;


/* FITTING UNADJUSTED GzLMM-MLE GLIMMIX*/
 proc glimmix data=ponder2 method = quad(qpoints=10);
  class group2(ref="control") clusteri;
  model atrisk6m = group / dist=bin s ddfm=bw cl;
random intercept / subject = clusteri type= vc g;
run;

/* FITTING ADJUSTED GzLMM-MLE USING GLIMMIX*/
 proc glimmix data=ponder2 method = quad(qpoints=10);
  class group2(ref="control") clusteri;
  model atrisk6m = group2 epds_6we alone history any_life/ dist=bin s ddfm=bw
cl;
random intercept / subject = clusteri type=vc g;
run;


/* FITTING UNADJUSTED QIF USING QIF MACRO*/
%qif(data=ponder2,
            yvar=atrisk6m,
            xvar=group2,
            id=clusteri, dist=bin, corr=exch, descend=y,
            print=y, outpar=par1, outqif=qif1, outcov=cov1, outres=resid1);
run;

/* FITTING ADJUSTED QIF USING QIF MACRO*/
%qif(data=ponder2,
            yvar=atrisk6m,
            xvar=group2 epds_6we alone history any_life ,
            id=clusteri, dist=bin, corr=exch, descend=y,
            print=y, outpar=par2, outqif=qif2, outcov=cov2, outres=resid2);
run;
```

```
#####################################################################
################ R CODES TO FIT GEE1 AND GEE2####################
#####################################################################

########## INSTALL REQUIRED PACKAGES ###################
install.packages(c("lme4",  “lmtest","afex",  “geepack",  “qif",  "dplyr",  "tidyverse","MLmetrics",
"MESS"))

########## LOAD INSTALL PACKAGES ######################
library(pacman)
pacman::p_load(afex,lmtest,lme4,geepack,qif,dplyr,tidyverse,MLmetrics,MESS)

######Import the PoNDER dataset #################
p.data<-read.csv("PONDER data  n=2659 with 12 and 18m followups.csv", header = T)
detach(p.data)
attach(p.data)

###Form new data frame for only covariates needed
PONDER<-select(p.data,epds_6mo,clusteri,group2,
          epds_6we,alone,history,any_life,atrisk6m)
##Converting intergers to factors
name<-c(2,3,5,6,7)
PONDER[,name]<-lapply(PONDER[,name],factor)

#########################################################################
############ (CONTINUOUS OUTCOME MODELS) ##########################
#########################################################################

####################GEE1 MODELS#################################
#####Unadjusted model
UnAdjusted_Ponder_GEE1_Cont<-geeglm(epds_6mo~group2,family=gaussian,data=PONDER,
                  id=clusteri,corstr = "exc")
summary(UnAdjusted_Ponder_GEE1_Cont)

######Adjusted model
#Select covariates needed
PONDER_Adjusted_Cont<-select(PONDER,epds_6mo,clusteri,group2,
                    epds_6we,alone,history,any_life)
Adjusted_Ponder_GEE1_Cont<-geeglm(epds_6mo~group2+epds_6we+
                  alone +history +any_life,family=gaussian,data=na.omit(PONDER_Adjusted_Cont),
                  id=clusteri,corstr = "exc")
summary(Adjusted_Ponder_GEE1_Cont)

####################GEE2 MODELS#################################
#####Unadjusted model
```

```
UnAdjusted_Ponder_GEE2_Cont<-geese(epds_6mo~group2,data=PONDER,
id=clusteri,family=gaussian(link="identity"),
                    corstr="exchangeable",cor.link = "fisherz",
                    sca.link = "identity")
summary(UnAdjusted_Ponder_GEE2_Cont)

#####Adjusted model
Adjusted_Ponder_GEE2_Cont<-geese(epds_6mo~group2+epds_6we+alone+history+any_life,
data= na.omit(PONDER_Adjusted_Cont),id=clusteri,family=gaussian(link="identity"),
                  corstr="exchangeable",cor.link = "fisherz", sca.link = "identity")
summary(Adjusted_Ponder_GEE2_Cont)

#########################################################################
####################BINARY OUTCOMES#####################################
#########################################################################

########################GEE1 MODELS#################################
#####Unadjusted model
UnAdjusted_Ponder_GEE1_Bi<-geeglm(atrisk6m~group2,family=binomial(link                    =
"logit"),data=PONDER,
                  id=clusteri,corstr = "exc")
summary(UnAdjusted_Ponder_GEE1_Bi)

#####Adjusted analysis
#Select covariates needed
PONDER_Adjusted_Bi<-select(PONDER, atrisk6m,clusteri,group2,
                    epds_6we,alone,history,any_life)

##### Fit the adjusted model
Adjusted_Ponder_GEE1_Bi<-geeglm(atrisk6m~group2+epds_6we+alone+history
+any_life,family=binomial(link = "logit"), data= na.omit(PONDER_Adjusted_Bi),
                id=clusteri,corstr = "exc")
summary(Adjusted_Ponder_GEE1_Bi)

########################GEE2 MODELS#################################
#####Unadjusted model
UnAdjusted_Ponder_GEE2_Bi<-geese(atrisk6m~group2,data=PONDER,
id=clusteri,family=binomial(link="logit"),
              corstr="exchangeable",cor.link = "fisherz",
              sca.link = "identity")
summary(UnAdjusted_Ponder_GEE2_Bi)

#####Adjusted model
Adjusted_Ponder_GEE2_Bi<-geese(atrisk6m~group2+epds_6we+alone +history +any_life,
data= na.omit(PONDER_Adjusted_Bi), id=clusteri,family=binomial(link="logit"),
                    corstr="exchangeable",cor.link = "fisherz",
                    sca.link = "identity")
summary(Adjusted_Ponder_GEE2_Bi)
```

Appendix 5

R codes for the simulation study (Chapter 9)

```
##############Install all required packages
#install.packages(c("dplyr", "tidyverse", "rsimsum", "qif", "geepack", "lme4","afex","broom","purrr" , "pacman"))

##############Load all required packages
library(pacman)
pacman::p_load(dplyr, tidyverse, rsimsum, qif, geepack, lme4,afex,broom,purrr)

#########################################################################
########Define the parameters of the simulation function:
# n_i is cluster size
# n_cl is number of clusters per treatment arm
# rho is intracluster correlation coefficient (ICC)
# b_1 is true intervention effect/treatment effect

############################## MAIN SIMULATION FUNCTION ###################################
## function to simulate the continuous outcome y, fit the 3 methods and extract the parameter estimates
sim_fun<-function(i,dgm,n_cl,n_i,rho,b_1){
 # create the total observations
 # create the clusters
 # create the participant unique ID
 #generate the indicator variable; x_1=0 control, x_1=1 intervention
 n_obs<-(2*n_cl)*n_i
 cl<-as.factor(rep(1:(2*n_cl),each=n_i))
 id<-(1:n_obs)
 x_1<-rep(0:1,n_cl,each=n_i)

 #generate the cluster level random effects
 #generate the individual subjects cluster level random effects
 #generate the individual level residuals
 #add up all the random effects
 cl_error<-rnorm((2*n_cl),0,sd=sqrt(rho/(1-rho)))
 ind_cl_error<-rep(cl_error, each=n_i)
 ind_error<-rnorm(n_obs,0,1)
 error<-ind_error+ind_cl_error
 #generate the vector of the continuous outcomes y
 #put all relevant variables in a data frame
 y <- (b_1*x_1)+error
 simul<-data.frame(id,x_1,cl,y)

 # analyse simulated data set with MLE
 # extract parameter estimates
 fit0.mle<- tryCatch(lmer(y~x_1+(1|cl),data = simul, REML = F),
```

```
                error= function(e) NULL,
                warning = function(w) NULL
)
if(is.null(fit0.mle)){
  out=data.frame(
    i=i,
    dgm=dgm,
    method=1,
    theta_true=NA,
    theta=NA,
    se_theta=NA,
    ll=NA,ul=NA,p=NA,icc=NA)
}
else{
  fit1.mle<-summary(fit0.mle)[10]
  theta<-fit1.mle$coefficients[2]
  se_theta<-fit1.mle$coefficients[4]
  ll<-confint(fit0.mle, method = "Wald")[4,1]
  ul<-confint(fit0.mle, method = "Wald")[4,2]
  p<-fit1.mle$coefficients[10]
  icc<-summary(fit0.mle)$varcor$cl[1]/(summary(fit0.mle)$varcor$cl[1]+(sigma(fit0.mle))^2)

  out=data.frame(
    i=i,
    dgm=dgm,
    method=1,
    theta_true=b_1,
    theta,
    se_theta,
    ll,ul,p,icc)
}
 #---------------------------------------------------------------------
####analyse simulated dataset with GEE1
#extract parameter estimates

fit0.gee<- tryCatch(summary(geeglm(y~x_1,data = simul, family = gaussian, corstr = "exchangeable",id=cl)),
                error= function(e) NULL,
                warning = function(w) NULL
)
  if(is.null(fit0.gee)){
  out2=data.frame(
    i=i,
    dgm=dgm,
    method=1,
    theta_true=NA,
    theta=NA,
    se_theta=NA,
    ll=NA,ul=NA,p=NA,icc=NA)
```

```
}
else{
 theta<-coef(fit0.gee)[2,1]
 se_theta<-coef(fit0.gee)[2,2]
 ll<-theta+(qnorm(0.025)*se_theta)
 ul<-theta+(qnorm(0.975)*se_theta)
 p<-fit0.gee$coefficients[2,4]
 icc<-fit0.gee$corr$Estimate

    out2=data.frame(
  i=i,
  dgm=dgm,
  method=2,
  theta_true=b_1,
  theta=theta,
  se_theta=se_theta,
  ll,ul,p,icc)
}

 #------------------------------------------------------------------------
#analyse simulated dataset with QIF
#extract parameter estimates

fit0.qif<- tryCatch( summary(qif(y~x_1,data = simul, family = gaussian, corstr = "exchangeable",id=cl, invfun =
"ginv")),
                error= function(e) NULL,
                warning = function(w) NULL
)
if(is.null(fit0.qif )){
 out3=data.frame(
  i=i,
  dgm=dgm,
  method=1,
  theta_true=NA,
  theta=NA,
  se=NA,
  ll=NA,up=NA,p=NA,icc=NA)
}
else{
 theta<-coef(fit0.qif)[2]
 se_theta<-coef(fit0.qif)[4]
 ll<-theta+(qnorm(0.025)*se_theta)
 ul<-theta+(qnorm(0.975)*se_theta)
 p<-coef(fit0.qif)[8]
 icc<-0

 out3=data.frame(
  i=i,
```

```
    dgm=dgm,
    method=3,
    theta_true=b_1,
    theta=theta,
    se_theta=se_theta,
    ll,ul,p,icc)
 }
 return(result<-list(out,out2,out3))
}
#Set seed
#create empty storage
set.seed(674563)
n_sim<-4000
data_ests0<-vector(mode = "list",length = n_sim)

#------------------------------------------------------------------------------
#loop for n_sim repetitions for each dgm
# Save random state of each repetition
 for (i in 1:n_sim)
 {
  attr(df,"seed")<-.Random.seed
  data_ests0[[i]]=sim_fun(i,dgm=1,5,150,0.001,0)
  data_ests0[[(i+n_sim)]]=sim_fun(i,dgm=2,5,150,0.01,0)
  data_ests0[[(i+2*n_sim)]]=sim_fun(i,dgm=3,5,150,0.05,0)
  data_ests0[[(i+3*n_sim)]]=sim_fun(i,dgm=4,5,150,0.25,0)
  data_ests0[[(i+4*n_sim)]]=sim_fun(i,dgm=5,5,250,0.001,0)
  data_ests0[[(i+5*n_sim)]]=sim_fun(i,dgm=6,5,250,0.01,0)
  data_ests0[[(i+6*n_sim)]]=sim_fun(i,dgm=7,5,250,0.05,0)
  data_ests0[[(i+7*n_sim)]]=sim_fun(i,dgm=8,5,250,0.25,0)

  data_ests0[[(i+8*n_sim)]]=sim_fun(i,dgm=9,10,50,0.001,0)
  data_ests0[[(i+9*n_sim)]]=sim_fun(i,dgm=10,10,50,0.01,0)
  data_ests0[[(i+10*n_sim)]]=sim_fun(i,dgm=11,10,50,0.05,0)
  data_ests0[[(i+11*n_sim)]]=sim_fun(i,dgm=12,10,50,0.25,0)
  data_ests0[[(i+12*n_sim)]]=sim_fun(i,dgm=13,10,150,0.001,0)
  data_ests0[[(i+13*n_sim)]]=sim_fun(i,dgm=14,10,150,0.01,0)
  data_ests0[[(i+14*n_sim)]]=sim_fun(i,dgm=15,10,150,0.05,0)
  data_ests0[[(i+15*n_sim)]]=sim_fun(i,dgm=16,10,150,0.25,0)

  data_ests0[[(i+16*n_sim)]]=sim_fun(i,dgm=17,20,20,0.001,0)
  data_ests0[[(i+17*n_sim)]]=sim_fun(i,dgm=18,20,20,0.01,0)
  data_ests0[[(i+18*n_sim)]]=sim_fun(i,dgm=19,20,20,0.05,0)
  data_ests0[[(i+19*n_sim)]]=sim_fun(i,dgm=20,20,20,0.25,0)
  data_ests0[[(i+20*n_sim)]]=sim_fun(i,dgm=21,20,100,0.001,0)
  data_ests0[[(i+21*n_sim)]]=sim_fun(i,dgm=22,20,100,0.01,0)
  data_ests0[[(i+22*n_sim)]]=sim_fun(i,dgm=23,20,100,0.05,0)
  data_ests0[[(i+23*n_sim)]]=sim_fun(i,dgm=24,20,100,0.25,0)
  #
  data_ests0[[(i+24*n_sim)]]=sim_fun(i,dgm=25,25,20,0.001,0)
  data_ests0[[(i+25*n_sim)]]=sim_fun(i,dgm=26,25,20,0.01,0)
  data_ests0[[(i+26*n_sim)]]=sim_fun(i,dgm=27,25,20,0.05,0)
```

```
    data_ests0[[(i+27*n_sim)]]=sim_fun(i,dgm=28,25,20,0.25,0)
    data_ests0[[(i+28*n_sim)]]=sim_fun(i,dgm=29,25,50,0.001,0)
    data_ests0[[(i+29*n_sim)]]=sim_fun(i,dgm=30,25,50,0.01,0)
    data_ests0[[(i+30*n_sim)]]=sim_fun(i,dgm=31,25,50,0.05,0)
    data_ests0[[(i+31*n_sim)]]=sim_fun(i,dgm=32,25,50,0.25,0)

    data_ests0[[(i+32*n_sim)]]=sim_fun(i,dgm=33,60,10,0.001,0)
    data_ests0[[(i+33*n_sim)]]=sim_fun(i,dgm=34,60,10,0.01,0)
    data_ests0[[(i+34*n_sim)]]=sim_fun(i,dgm=35,60,10,0.05,0)
    data_ests0[[(i+35*n_sim)]]=sim_fun(i,dgm=36,60,10,0.25,0)
    data_ests0[[(i+36*n_sim)]]=sim_fun(i,dgm=37,60,80,0.001,0)
    data_ests0[[(i+37*n_sim)]]=sim_fun(i,dgm=38,60,80,0.01,0)
    data_ests0[[(i+38*n_sim)]]=sim_fun(i,dgm=39,60,80,0.05,0)
    data_ests0[[(i+39*n_sim)]]=sim_fun(i,dgm=40,60,80,0.25,0)
    # #-------------------------------------------------->>>>>>>>
    #-------------------------------------------------->>>>>>>>
    data_ests0[[(i+40*n_sim)]]=sim_fun(i,dgm=41,5,150,0.001,0.2)
    data_ests0[[(i+41*n_sim)]]=sim_fun(i,dgm=42,5,150,0.01,0.2)
    data_ests0[[(i+42*n_sim)]]=sim_fun(i,dgm=43,5,150,0.05,0.2)
    data_ests0[[(i+43*n_sim)]]=sim_fun(i,dgm=44,5,150,0.25,0.2)
    data_ests0[[(i+44*n_sim)]]=sim_fun(i,dgm=45,5,250,0.001,0.2)
    data_ests0[[(i+45*n_sim)]]=sim_fun(i,dgm=46,5,250,0.01,0.2)
    data_ests0[[(i+46*n_sim)]]=sim_fun(i,dgm=47,5,250,0.05,0.2)
    data_ests0[[(i+47*n_sim)]]=sim_fun(i,dgm=48,5,250,0.25,0.2)
#
    data_ests0[[(i+48*n_sim)]]=sim_fun(i,dgm=49,10,50,0.001,0.2)
    data_ests0[[(i+49*n_sim)]]=sim_fun(i,dgm=50,10,50,0.01,0.2)
    data_ests0[[(i+50*n_sim)]]=sim_fun(i,dgm=51,10,50,0.05,0.2)
    data_ests0[[(i+51*n_sim)]]=sim_fun(i,dgm=52,10,50,0.25,0.2)
    data_ests0[[(i+52*n_sim)]]=sim_fun(i,dgm=53,10,150,0.001,0.2)
    data_ests0[[(i+53*n_sim)]]=sim_fun(i,dgm=54,10,150,0.01,0.2)
    data_ests0[[(i+54*n_sim)]]=sim_fun(i,dgm=55,10,150,0.05,0.2)
    data_ests0[[(i+55*n_sim)]]=sim_fun(i,dgm=56,10,150,0.25,0.2)

    data_ests0[[(i+56*n_sim)]]=sim_fun(i,dgm=57,20,20,0.001,0.2)
    data_ests0[[(i+57*n_sim)]]=sim_fun(i,dgm=58,20,20,0.01,0.2)
    data_ests0[[(i+58*n_sim)]]=sim_fun(i,dgm=59,20,20,0.05,0.2)
    data_ests0[[(i+59*n_sim)]]=sim_fun(i,dgm=60,20,20,0.25,0.2)
    data_ests0[[(i+60*n_sim)]]=sim_fun(i,dgm=61,20,100,0.001,0.2)
    data_ests0[[(i+61*n_sim)]]=sim_fun(i,dgm=62,20,100,0.01,0.2)
    data_ests0[[(i+62*n_sim)]]=sim_fun(i,dgm=63,20,100,0.05,0.2)
    data_ests0[[(i+63*n_sim)]]=sim_fun(i,dgm=64,20,100,0.25,0.2)
    #
    data_ests0[[(i+64*n_sim)]]=sim_fun(i,dgm=65,25,20,0.001,0.2)
    data_ests0[[(i+65*n_sim)]]=sim_fun(i,dgm=66,25,20,0.01,0.2)
    data_ests0[[(i+66*n_sim)]]=sim_fun(i,dgm=67,25,20,0.05,0.2)
    data_ests0[[(i+67*n_sim)]]=sim_fun(i,dgm=68,25,20,0.25,0.2)
    data_ests0[[(i+68*n_sim)]]=sim_fun(i,dgm=69,25,50,0.001,0.2)
    data_ests0[[(i+69*n_sim)]]=sim_fun(i,dgm=70,25,50,0.01,0.2)
    data_ests0[[(i+70*n_sim)]]=sim_fun(i,dgm=71,25,50,0.05,0.2)
    data_ests0[[(i+71*n_sim)]]=sim_fun(i,dgm=72,25,50,0.25,0.2)

    data_ests0[[(i+72*n_sim)]]=sim_fun(i,dgm=73,60,10,0.001,0.2)
    data_ests0[[(i+73*n_sim)]]=sim_fun(i,dgm=74,60,10,0.01,0.2)
    data_ests0[[(i+74*n_sim)]]=sim_fun(i,dgm=75,60,10,0.05,0.2)
```

```
  data_ests0[[(i+75*n_sim)]]=sim_fun(i,dgm=76,60,10,0.25,0.2)
  data_ests0[[(i+76*n_sim)]]=sim_fun(i,dgm=77,60,80,0.001,0.2)
  data_ests0[[(i+77*n_sim)]]=sim_fun(i,dgm=78,60,80,0.01,0.2)
  data_ests0[[(i+78*n_sim)]]=sim_fun(i,dgm=79,60,80,0.05,0.2)
  data_ests0[[(i+79*n_sim)]]=sim_fun(i,dgm=80,60,80,0.25,0.2)
  # ------------------------------------------------------------>>>>>>>>>
  # #------------------------------------------------------------>>>>>>>>>
  data_ests0[[(i+80*n_sim)]]=sim_fun(i,dgm=81,5,150,0.001,0.3)
  data_ests0[[(i+81*n_sim)]]=sim_fun(i,dgm=82,5,150,0.01,0.3)
  data_ests0[[(i+82*n_sim)]]=sim_fun(i,dgm=83,5,150,0.05,0.3)
  data_ests0[[(i+83*n_sim)]]=sim_fun(i,dgm=84,5,150,0.25,0.3)
  data_ests0[[(i+84*n_sim)]]=sim_fun(i,dgm=85,5,250,0.001,0.3)
  data_ests0[[(i+85*n_sim)]]=sim_fun(i,dgm=86,5,250,0.01,0.3)
  data_ests0[[(i+86*n_sim)]]=sim_fun(i,dgm=87,5,250,0.05,0.3)
  data_ests0[[(i+87*n_sim)]]=sim_fun(i,dgm=88,5,250,0.25,0.3)
  #
  data_ests0[[(i+88*n_sim)]]=sim_fun(i,dgm=89,10,50,0.001,0.3)
  data_ests0[[(i+89*n_sim)]]=sim_fun(i,dgm=90,10,50,0.01,0.3)
  data_ests0[[(i+90*n_sim)]]=sim_fun(i,dgm=91,10,50,0.05,0.3)
  data_ests0[[(i+91*n_sim)]]=sim_fun(i,dgm=92,10,50,0.25,0.3)
  data_ests0[[(i+92*n_sim)]]=sim_fun(i,dgm=93,10,150,0.001,0.3)
  data_ests0[[(i+93*n_sim)]]=sim_fun(i,dgm=94,10,150,0.01,0.3)
  data_ests0[[(i+94*n_sim)]]=sim_fun(i,dgm=95,10,150,0.05,0.3)
  data_ests0[[(i+95*n_sim)]]=sim_fun(i,dgm=96,10,150,0.25,0.3)
  #
  data_ests0[[(i+96*n_sim)]]=sim_fun(i,dgm=97,20,20,0.001,0.3)
  data_ests0[[(i+97*n_sim)]]=sim_fun(i,dgm=98,20,20,0.01,0.3)
  data_ests0[[(i+98*n_sim)]]=sim_fun(i,dgm=99,20,20,0.05,0.3)
  data_ests0[[(i+99*n_sim)]]=sim_fun(i,dgm=100,20,20,0.25,0.3)
  data_ests0[[(i+100*n_sim)]]=sim_fun(i,dgm=101,20,100,0.001,0.3)
  data_ests0[[(i+101*n_sim)]]=sim_fun(i,dgm=102,20,100,0.01,0.3)
  data_ests0[[(i+102*n_sim)]]=sim_fun(i,dgm=103,20,100,0.05,0.3)
  data_ests0[[(i+103*n_sim)]]=sim_fun(i,dgm=104,20,100,0.25,0.3)
  #
  data_ests0[[(i+104*n_sim)]]=sim_fun(i,dgm=105,25,20,0.001,0.3)
  data_ests0[[(i+105*n_sim)]]=sim_fun(i,dgm=106,25,20,0.01,0.3)
  data_ests0[[(i+106*n_sim)]]=sim_fun(i,dgm=107,25,20,0.05,0.3)
  data_ests0[[(i+107*n_sim)]]=sim_fun(i,dgm=108,25,20,0.25,0.3)
  data_ests0[[(i+108*n_sim)]]=sim_fun(i,dgm=109,25,50,0.001,0.3)
  data_ests0[[(i+109*n_sim)]]=sim_fun(i,dgm=110,25,50,0.01,0.3)
  data_ests0[[(i+110*n_sim)]]=sim_fun(i,dgm=111,25,50,0.05,0.3)
  data_ests0[[(i+111*n_sim)]]=sim_fun(i,dgm=112,25,50,0.25,0.3)
  #
  data_ests0[[(i+112*n_sim)]]=sim_fun(i,dgm=113,60,10,0.001,0.3)
  data_ests0[[(i+113*n_sim)]]=sim_fun(i,dgm=114,60,10,0.01,0.3)
  data_ests0[[(i+114*n_sim)]]=sim_fun(i,dgm=115,60,10,0.05,0.3)
  data_ests0[[(i+115*n_sim)]]=sim_fun(i,dgm=116,60,10,0.25,0.3)
  data_ests0[[(i+116*n_sim)]]=sim_fun(i,dgm=117,60,80,0.001,0.3)
  data_ests0[[(i+117*n_sim)]]=sim_fun(i,dgm=118,60,80,0.01,0.3)
  data_ests0[[(i+118*n_sim)]]=sim_fun(i,dgm=119,60,80,0.05,0.3)
  data_ests0[[(i+119*n_sim)]]=sim_fun(i,dgm=120,60,80,0.25,0.3)

}
#Combine parameter estimates from all runs - long format
```

#Save estimates data sets
data_est_long=bind_rows(data_ests0,.id = NULL)
saveRDS(data_ests0, file = "simulated_data_hpc.rds")

##
######################## ANALYSING THE SIMULATED DATASETS (R MARKDOWN)#############
##

#####Load all required packages

```{r echo=FALSE}
library(pacman)
pacman::p_load(dplyr, tidyverse, rsimsum, qif, geepack, lme4,caTools, bitops,
        afex,broom,purrr,cowplot,readr,lavaan,ggplot2,PupillometryR,blandr,ggthemes)
```

```{r results='hide'}
#Load all saved RDS data
 dt<-readRDS("U:/ManW10/Desktop/PhD_3YR_MAIN/SIMULATION_HPC_BESSEMER/NEW SIMULATION FILES/data_hpc_full.rds")
#Combine all lists to a large data frame
df_ests_hpc=bind_rows(dt,.id = NULL)
```

#Performing EDA on estimates

```{r result="hide"}
df_ests_hpc$method<-recode_factor(df_ests_hpc$method, `1`="GzLMM",`2`="GEE1", `3`="QIF")
```

```{r echot=F, results='hide'}
#####Check non convergence
mis_df <- df_ests_hpc[rowSums(is.na(df_ests_hpc)) > 0,]
summary(mis_df)
length((mis_df))
lapply(mis_df,summary)
```

```
##################################################
#Check raw estimates for outliers, skewness and distribution-----------------------------------------
##################################################

```{r,  fig.show= "hide"}
#Rainclouds or half-half plots
ggplot(subset(df_ests_hpc,dgm==78), aes(x=method, y= theta, fill=method,colour=method))+
geom_flat_violin(position=position_nudge(x=0.25,y=0), adjust=2)+
  geom_point(position=position_jitter(width=0.15),size=0.25)+
   geom_boxplot(aes(x=as.numeric(method)+0.02,y=theta),outlier.shape = NA,alpha=0.3,width=.1,colour="black")+
    ylab(" Estimates of θ")+xlab("Method")+coord_flip()+theme_cowplot()+guides(fill="none", colour="none")+
ggtitle( "DGM 78: N = 120, n_i = 80, ICC = 0.01, θ = 0.2"    ) +
 theme_cowplot(font_size  = 16)
```

##
#Compute performance measures and their MCSEs using rsimsum package----------------------------
##

```{r}
s<- subset(df_ests_hpc,dgm%in% 41:80)
```
#Automated performance measures
```{r }
perf_sim <- simsum(s,
                   estvarname = "theta",
                   se = "se_theta",
                   true = "theta_true",
                   by = "dgm",
                   methodvar = "method",
                  ref="GzLMM",
                   x=TRUE)
```
```

```
####Extraxt estimated bias for all methods----------------------------------------------------------------
```{r include=FALSE}
df_theta<-subset(perf_sim$summ, stat=="cover")

######Prepare performance measures for plotting----------------------------------------------
df_theta<-df_theta%>%
mutate( N=
 case_when(
 dgm%in% 1:8~10,dgm%in% 9:16~20,dgm %in% 17:24~40,dgm%in% 25:32~50,dgm %in% 33:40~120,
 dgm %in% 41:48~10,dgm %in% 49:56~20,dgm %in% 57:64~40,dgm %in% 65:72~50,dgm %in% 73:80~120,
  dgm %in% 81:88~10,dgm %in%89:96~20,dgm %in% 97:104~40,dgm%in% 105:112~50,dgm %in% 113:120~120
 )
)

#----------------------------------------------------------------
df_theta<-df_theta%>%
mutate( n =
 case_when(
 dgm==1~150, dgm==2~150,dgm==3~150,dgm==4~150,dgm==5~250,dgm==6~250,
 dgm==7~250, dgm==8~250,dgm==9~50,dgm==10~50,dgm==11~50,dgm==12~50,
 dgm==13~150, dgm==14~150,dgm==15~150,dgm==16~150,dgm==17~20,dgm==18~20,
 dgm==19~20, dgm==20~20,dgm==21~100,dgm==22~100,dgm==23~100,dgm==24~100,
 dgm==25~20, dgm==26~20,dgm==27~20,dgm==28~20,dgm==29~50,dgm==30~50,

 dgm==31~50, dgm==32~50,dgm==33~10,dgm==34~10,dgm==35~10,dgm==36~10,
 dgm==37~80, dgm==38~80,dgm==39~80,dgm==40~80,dgm==41~150,dgm==42~150,
 dgm==43~150, dgm==44~150,dgm==45~250,dgm==46~250,dgm==47~250,dgm==48~250,
 dgm==49~50, dgm==50~50,dgm==51~50,dgm==52~50,dgm==53~150,dgm==54~150,
 dgm==55~150, dgm==56~150,dgm==57~20,dgm==58~20,dgm==59~20,dgm==60~20,

 dgm==61~100, dgm==62~100,dgm==63~100,dgm==64~100,dgm==65~20,dgm==66~20,
 dgm==67~20, dgm==68~20,dgm==69~50,dgm==70~50,dgm==71~50,dgm==72~50,
 dgm==73~10, dgm==74~10,dgm==75~10,dgm==76~10,dgm==77~80,dgm==78~80,
 dgm==79~80, dgm==80~80,dgm==81~150,dgm==82~150,dgm==83~150,dgm==84~150,
 dgm==85~250, dgm==86~250,dgm==87~250,dgm==88~250,dgm==89~50,dgm==90~50,
```
```

```
 dgm==91~50, dgm==92~50,dgm==93~150,dgm==94~150,dgm==95~150,dgm==96~150,
 dgm==97~20, dgm==98~20,dgm==99~20,dgm==100~20,dgm==101~100,dgm==102~100,
 dgm==103~100, dgm==104~100,dgm==105~20,dgm==106~20,dgm==107~20,dgm==108~20,
 dgm==109~50, dgm==110~50,dgm==111~50,dgm==112~50,dgm==113~10,dgm==114~10,
 dgm==115~10, dgm==116~10,dgm==117~80,dgm==118~80,dgm==119~80,dgm==120~80
  )
)
### Separate the dgm for the effect sizes
df_theta <- df_theta%>%
mutate(θ =
 case_when(
 dgm %in% 1:40~0,dgm %in% 41:80~0.2, dgm %in% 81:120~0.3
 )
)
df_theta[,"ICC"]<-factor(rep(c("0.001","0.01","0.05","0.25"),times=10,each=3))
```

#####Convert parameters to factors

```{r include = FALSE}
df_theta<- mutate_at(df_theta, vars(N, n, θ), as.factor)
```

##For descriptive statistics

```{r}
df_mean<-aggregate(est~method+N+ICC+θ,df_theta,mean)
```

```
#######################################
#Plotting performance measures rsimsum###
#######################################
```

Boxplot for each performance measure

```{r include=false}
autoplot(summary(perf_sim), stats="mse")+ggplot2::theme_bw()+
 # geom_hline(yintercept=0.0,color="gray100",linetype="solid")+
 # geom_hline(yintercept=0.05002,color="black",linetype="dashed")+
#geom_hline(yintercept=0.90,color="red",linetype="dashed")+
 #geom_hline(yintercept=0.80,color="black",linetype="dotted")+
```

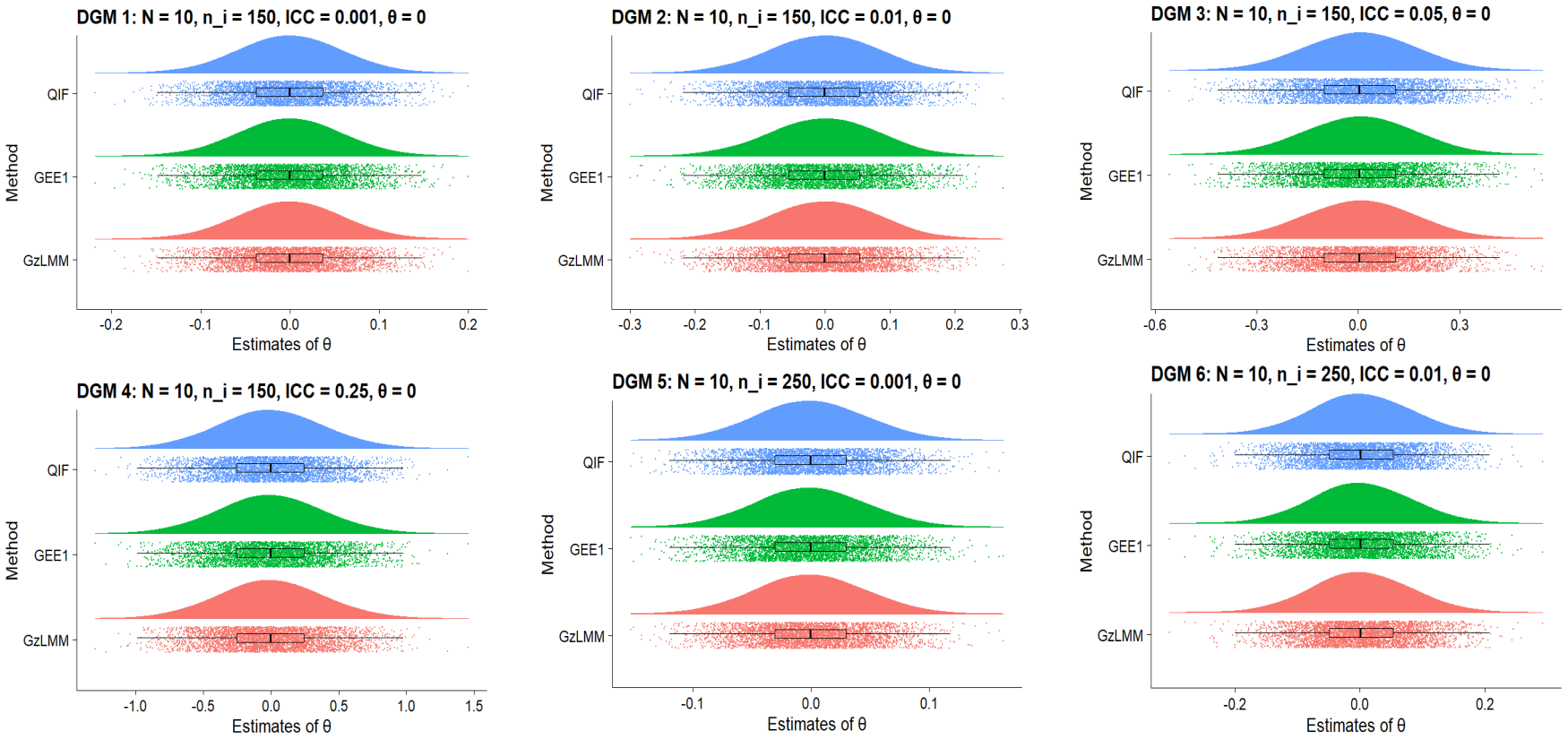
```
 labs(y="Mean square error",x="Statistical method")+
  ggplot2::scale_fill_viridis_c()#+
   #ggtitle("Agreement plot for comparison of the theta of two methods")
```

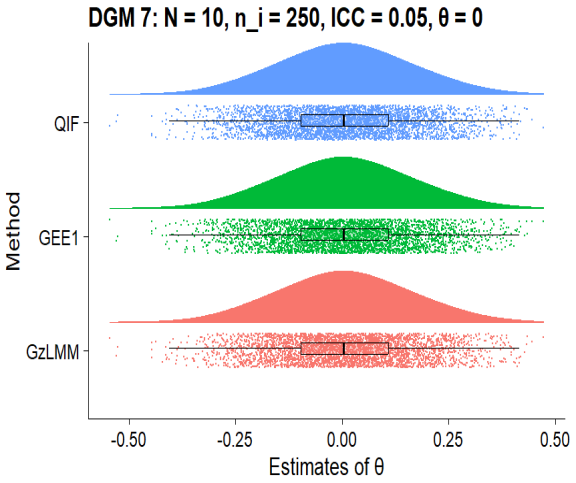
```
 ##### BOXPLOTS BY FACTORS
```{r}
p<-ggplot(data=df_theta, aes(method, est, fill=method))+
  #geom_boxplot()
geom_line(linewidth=1.5,alpha=0)+
 p+theme_bw(base_size = 16)+
#geom_hline(yintercept= 0.0, linetype=3,color= "black", linewidth=1)+
 facet_grid(cols=vars(θ), labeller = labeller(θ=label_both, method=label_value))+
  labs(y="Empirical standard error",x="Statistical method", fill="Method")+
  theme(legend.position = "right",legend.background = element_rect(fill="gray80",  linetype="dotted"))
#shap=c(0,19,3)
#p + scale_shape_manual(values = shap)+
   #scale_color_manual(values=c("blue4","green4","red"))+
#scale_linetype_manual(values=c("twodash","longdash","dotted"))
```
#Trend PLOTs of mean bias BY N, θ, ICC AND method
```{r}
p1<-ggplot(data=df_mean, aes(x=N,y=est,group=interaction(method,θ)))+
  geom_point(size=3,aes(shape=method))+
geom_line(linewidth=1.5,alpha=0.7,aes(color=method,linetype=method))+
 theme_bw(base_size = 20)+  #geom_hline(yintercept= 0.96, linetype=2,color= "black", linewidth=1)+
geom_hline(yintercept= 0.90, linetype=3,color= "red", linewidth=1)+
 geom_hline(yintercept= 0.80, linetype=2,color= "black", linewidth=1)+
 facet_grid(row=vars(θ),cols=vars(ICC),labeller = labeller(ICC=label_both,θ=label_both))+
  labs(y="Mean power",x="Number of clusters")+
  theme(legend.position = "right",legend.background = element_rect(fill="gray70",  linetype="dotted"))
shap=c(0,19,3)
   p1 + scale_shape_manual(values = shap)+
   scale_color_manual(values=c("blue4","green4","red"))+
   scale_linetype_manual(values=c("twodash","longdash","dotted"))
```
#Trend PLOT BY N, n AND THETA
```

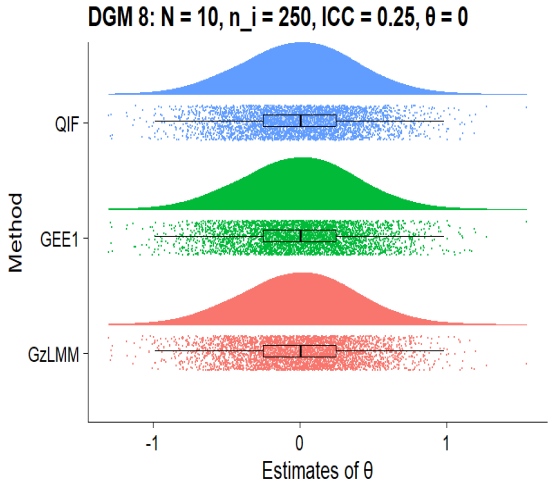
```{r}
p1<-ggplot(data=df_theta, aes(N,est,method,shape=n))+
  geom_point(size=3,aes(shape=n))+
#geom_line(linewidth=1.5,alpha=0.7)+
 theme_bw(base_size = 18)+
 # geom_hline(yintercept= 0.96, linetype=2,color= "black", linewidth=1)+
geom_hline(yintercept= 0.05, linetype=3,color= "red", linewidth=1)+
# geom_hline(yintercept= 0.8, linetype=2,color= "black", linewidth=1)+
 facet_grid(rows=vars(method),cols=vars(ICC),labeller = labeller(ICC=label_both))+
  labs(y="Type I error rate",x="Number of clusters")+
  theme(legend.position = "right",legend.background = element_rect(fill="gray80",  linetype="dotted"))
shap=c(3,8,17,19,5,0,9)
p1 + scale_shape_manual(values = shap)+
scale_color_manual(values=c("black","red3"))
#scale_linetype_manual(values=c("twodash","longdash","dotted"))
```

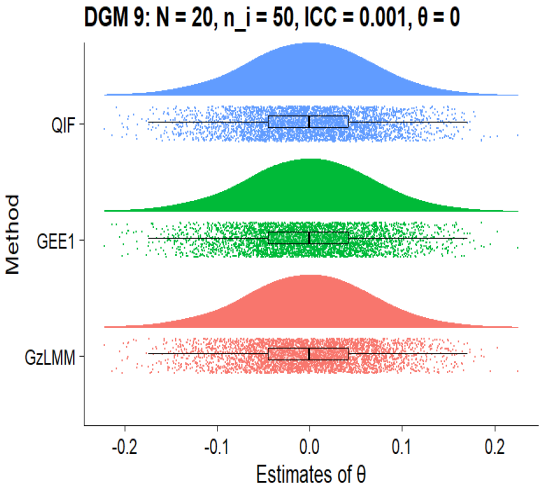
Appendix 6

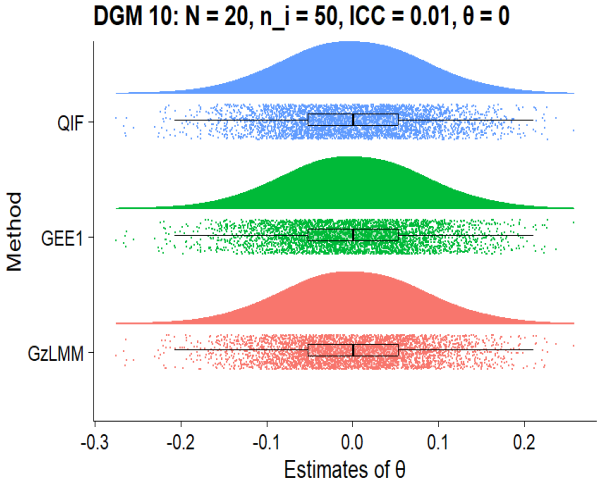
A raincloud plot showing the distribution (top density plot), the specific points (scatter plot) and the basic summary statistics (superimposed boxplot) of the simulated 4000 estimates of the true intervention effect θ from each method for a single scenario.

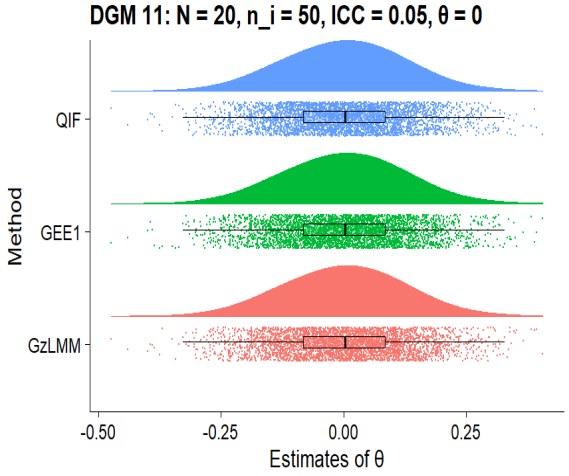


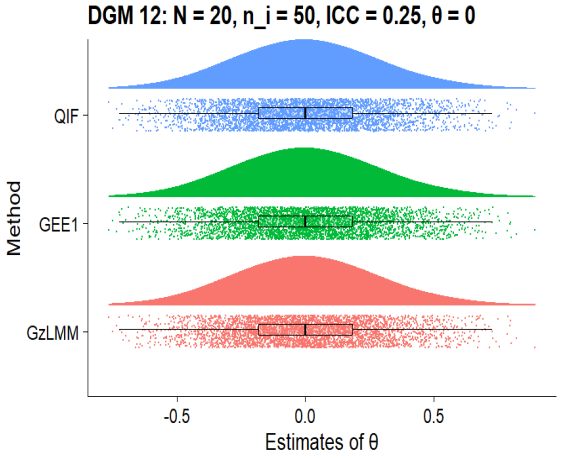

DGM 7: N = 10, n_i = 250, ICC = 0.05, θ = 0
Method
QIF
GEE1
GzLMM
-0.50
-0.25
0.00
0.25
0.50
Estimates of θ

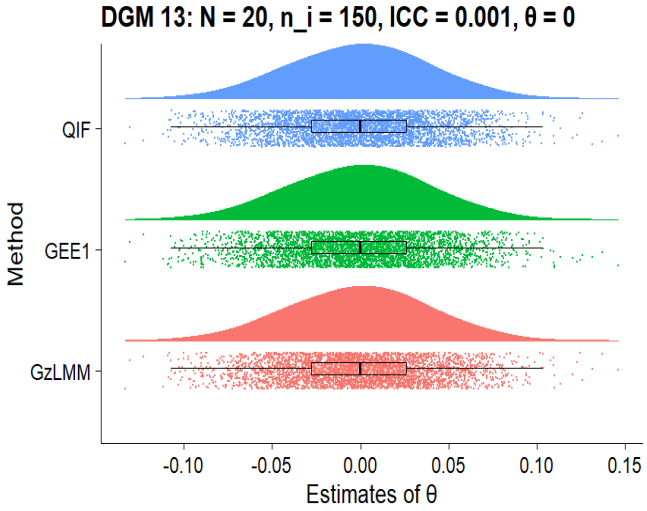

DGM 8: N = 10, n_i = 250, ICC = 0.25, θ = 0
Method
QIF
GEE1
GzLMM
-1
0
1
Estimates of θ

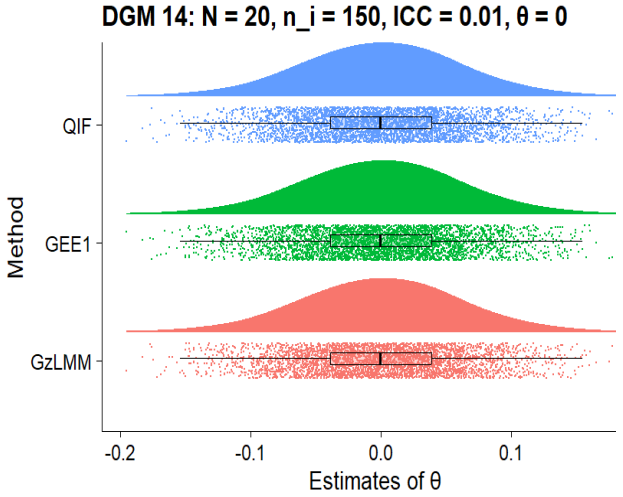

DGM 9: N = 20, n_i = 50, ICC = 0.001, θ = 0
Method
QIF
GEE1
GzLMM
-0.2
-0.1
0.0
0.1
0.2
Estimates of θ

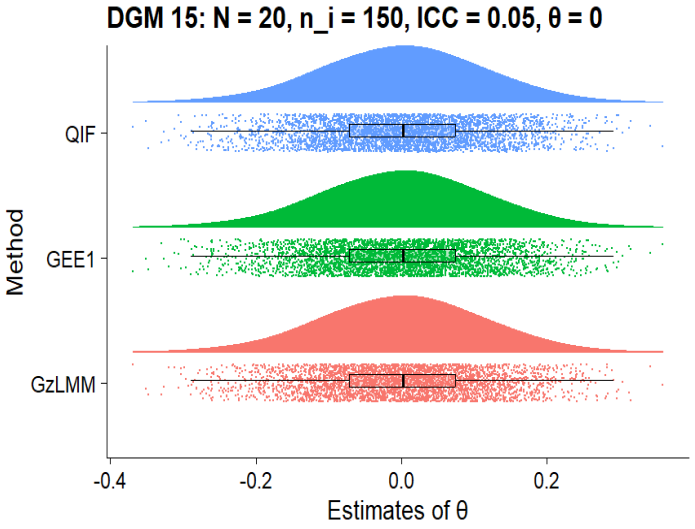

DGM 10: N = 20, n_i = 50, ICC = 0.01, θ = 0
Method
QIF
GEE1
GzLMM
-0.3
-0.2
-0.1
0.0
0.1
0.2
Estimates of θ

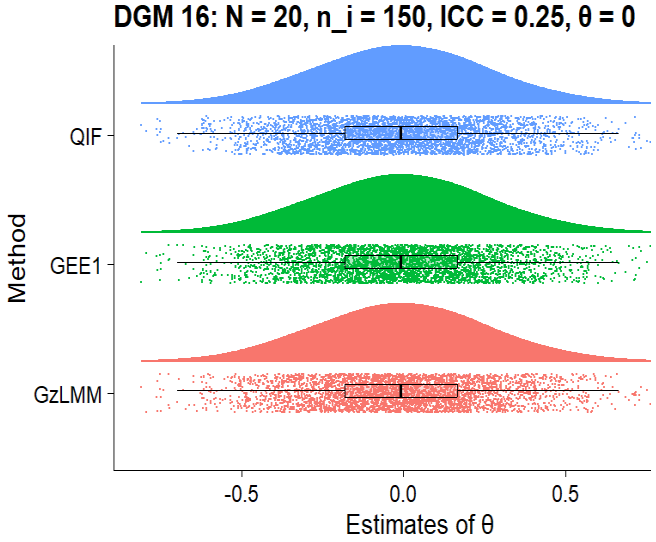

DGM 11: N = 20, n_i = 50, ICC = 0.05, θ = 0
Method
QIF
GEE1
GzLMM
-0.50
-0.25
0.00
0.25
Estimates of θ

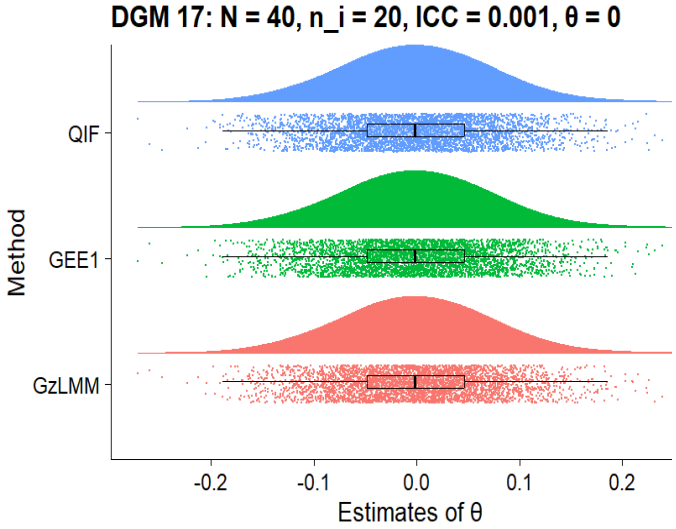

DGM 12: N = 20, n_i = 50, ICC = 0.25, θ = 0
Method
QIF
GEE1
GzLMM
-0.5
0.0
0.5
Estimates of θ

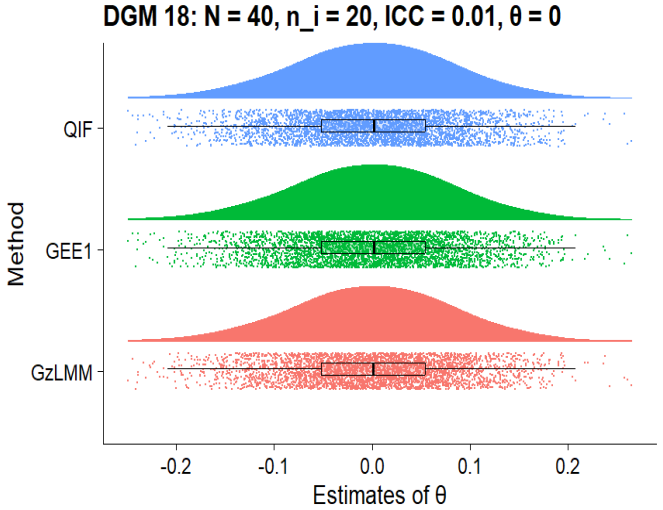

DGM 13: N = 20, n_i = 150, ICC = 0.001, θ = 0
Method
QIF
GEE1
GzLMM
-0.10
-0.05
0.00
0.05
0.10
0.15
Estimates of θ

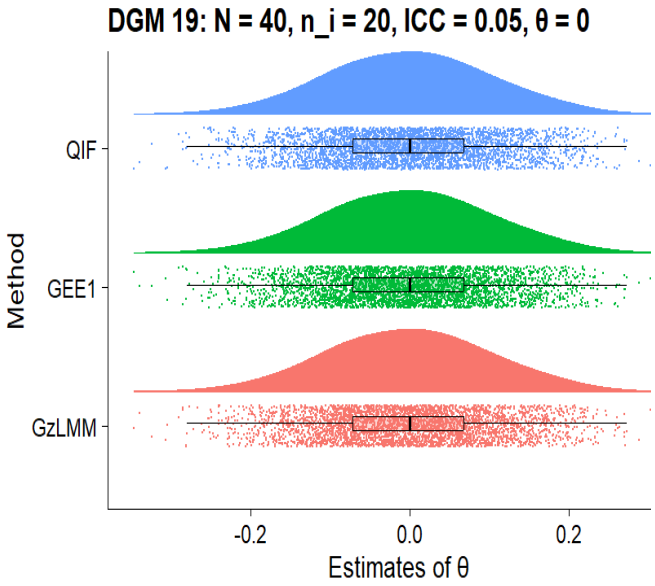

DGM 14: N = 20, n_i = 150, ICC = 0.01, θ = 0
Method
QIF
GEE1
GzLMM
-0.2
-0.1
0.0
0.1
Estimates of θ

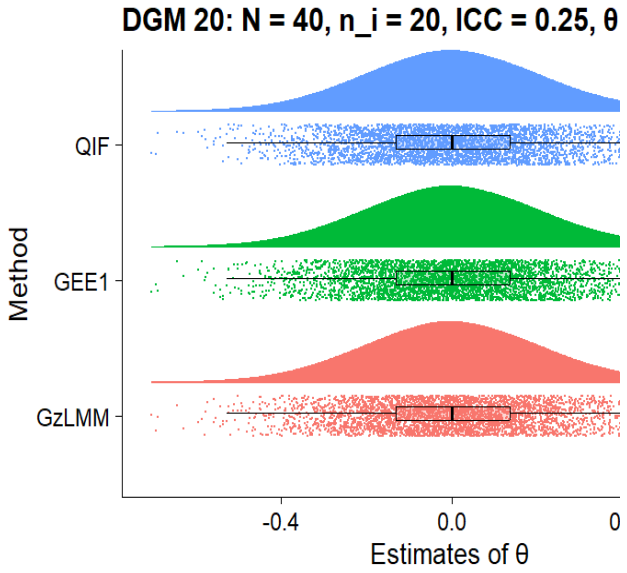

DGM 15: N = 20, n_i = 150, ICC = 0.05, θ = 0
Method
QIF
GEE1
GzLMM
-0.4
-0.2
0.0
0.2
Estimates of θ

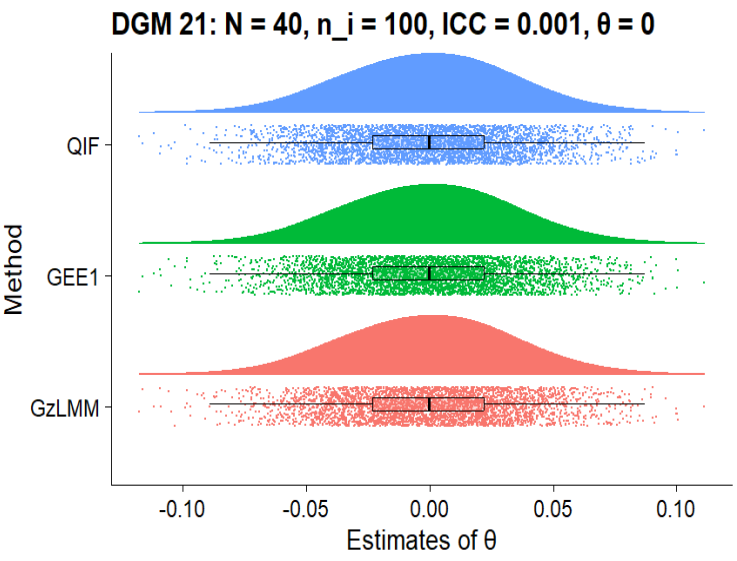

DGM 16: N = 20, n_i = 150, ICC = 0.25, θ = 0
Method
QIF
GEE1
GzLMM
-0.5
0.0
0.5
Estimates of θ

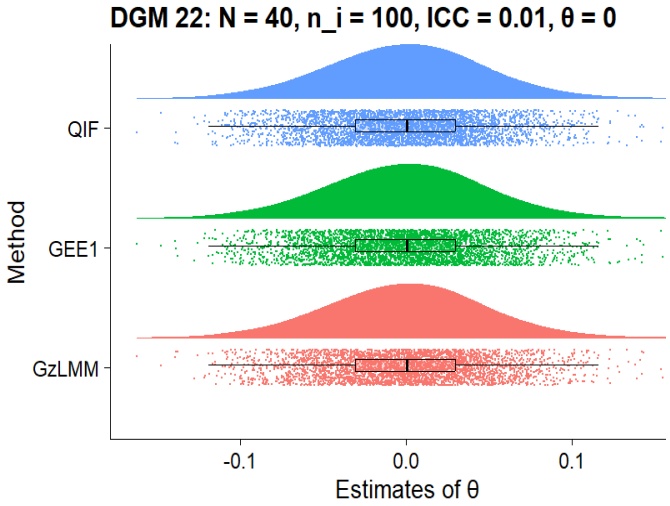

DGM 17: N = 40, n_i = 20, ICC = 0.001, θ = 0
Method
QIF
GEE1
GzLMM
-0.2
-0.1
0.0
0.1
0.2
Estimates of θ

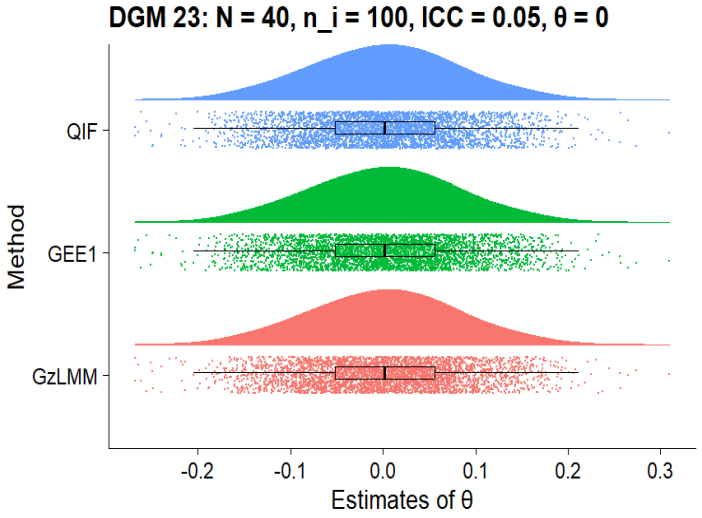

DGM 18: N = 40, n_i = 20, ICC = 0.01, θ = 0
Method
QIF
GEE1
GzLMM
-0.2
-0.1
0.0
0.1
0.2
Estimates of θ

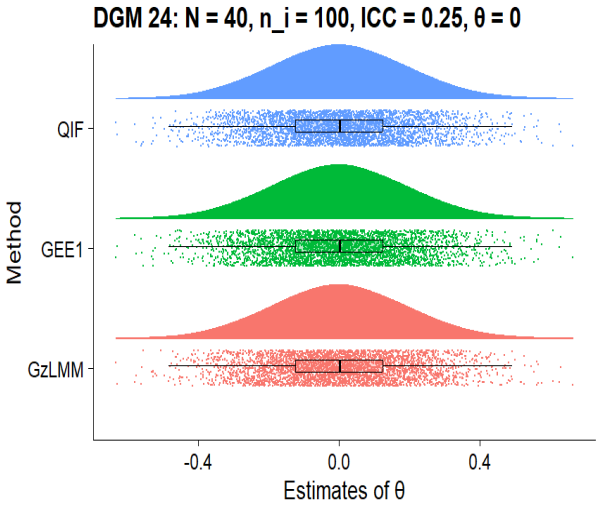

DGM 19: N = 40, n_i = 20, ICC = 0.05, θ = 0
QIF
GEE1
GzLMM
Method
-0.2
0.0
0.2
Estimates of θ

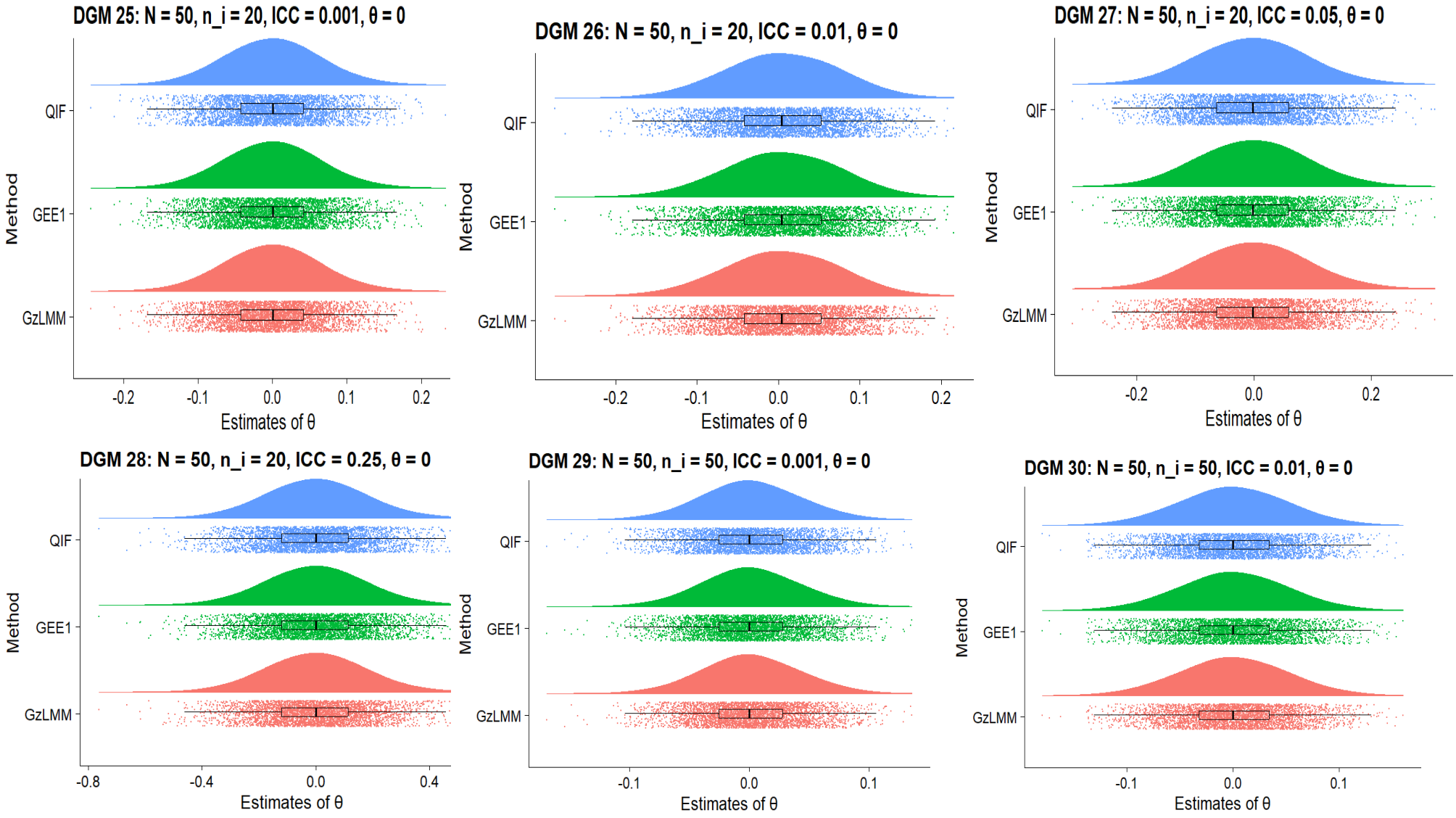

DGM 20: N = 40, n_i = 20, ICC = 0.25, θ
QIF
GEE1
GzLMM
Method
-0.4
0.0
0
Estimates of θ

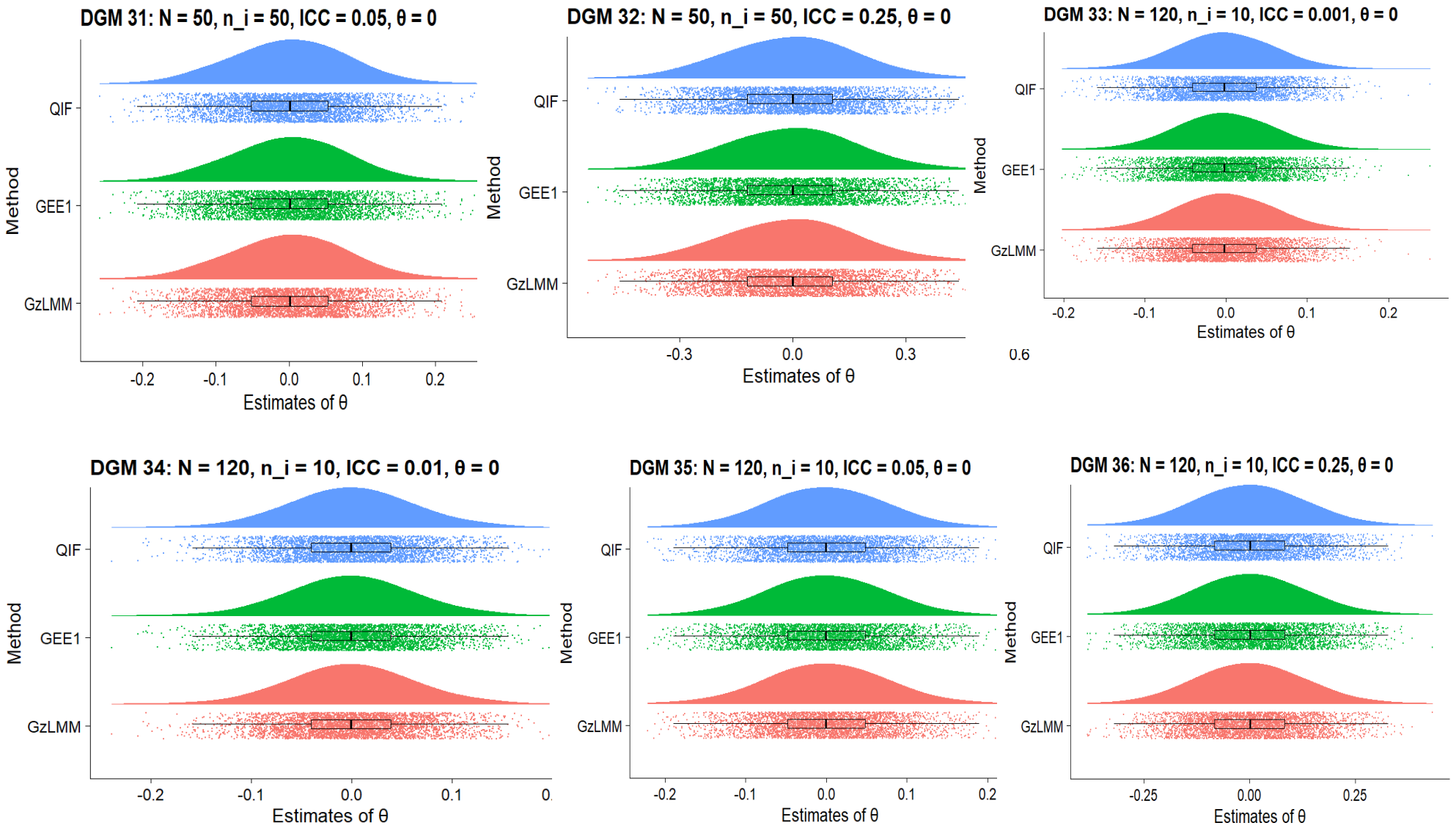

DGM 21: N = 40, n_i = 100, ICC = 0.001, θ = 0
QIF
GEE1
GzLMM
Method
-0.10
-0.05
0.00
0.05
0.10
Estimates of θ

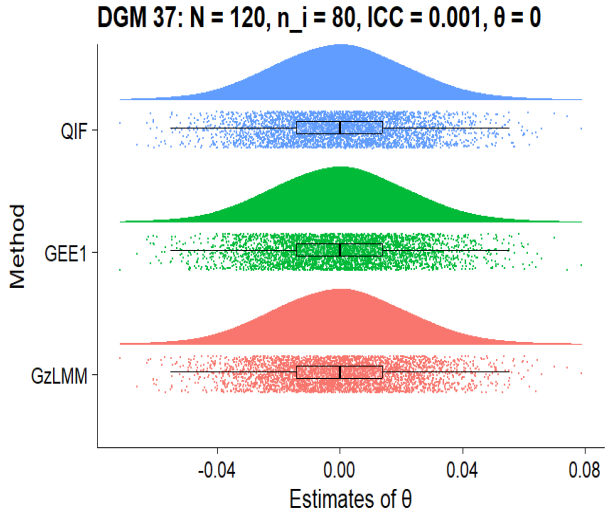

DGM 22: N = 40, n_i = 100, ICC = 0.01, θ = 0
QIF
GEE1
GzLMM
Method
-0.1
0.0
0.1
Estimates of θ

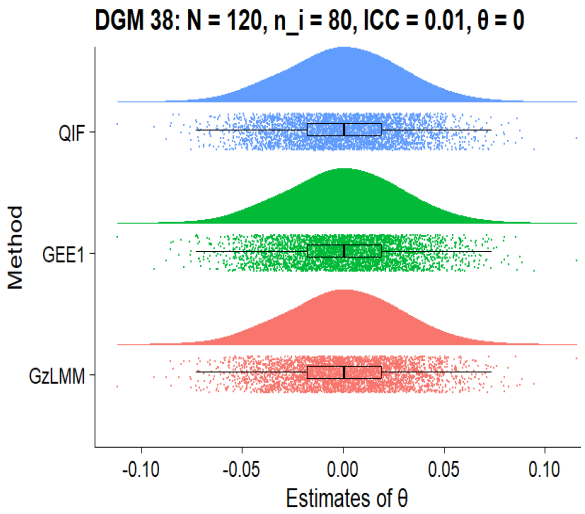

DGM 23: N = 40, n_i = 100, ICC = 0.05, θ = 0
QIF
GEE1
GzLMM
Method
-0.2
-0.1
0.0
0.1
0.2
0.3
Estimates of θ

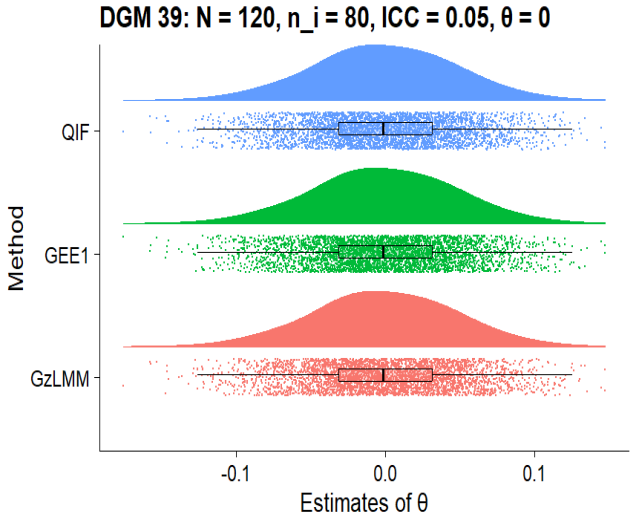

DGM 24: N = 40, n_i = 100, ICC = 0.25, θ = 0
QIF
GEE1
GzLMM
Method
-0.4
0.0
0.4
Estimates of θ

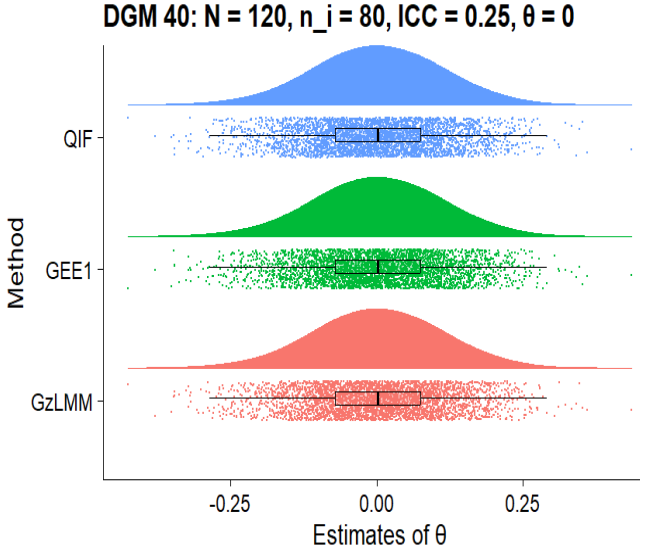

DGM 25: N = 50, n_i = 20, ICC = 0.001, θ = 0
DGM 26: N = 50, n_i = 20, ICC = 0.01, θ = 0
DGM 27: N = 50, n_i = 20, ICC = 0.05, θ = 0
DGM 28: N = 50, n_i = 20, ICC = 0.25, θ = 0
DGM 29: N = 50, n_i = 50, ICC = 0.001, θ = 0
DGM 30: N = 50, n_i = 50, ICC = 0.01, θ = 0
Method
QIF
GEE1
GzLMM
Estimates of θ

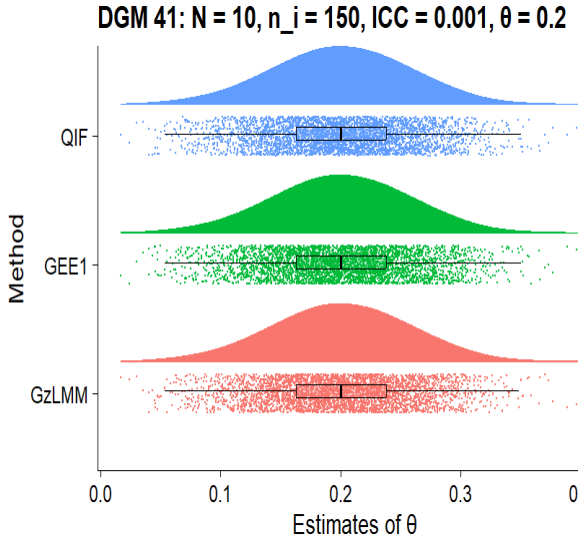

DGM 31: N = 50, n_i = 50, ICC = 0.05, θ = 0
DGM 32: N = 50, n_i = 50, ICC = 0.25, θ = 0
DGM 33: N = 120, n_i = 10, ICC = 0.001, θ = 0
DGM 34: N = 120, n_i = 10, ICC = 0.01, θ = 0
DGM 35: N = 120, n_i = 10, ICC = 0.05, θ = 0
DGM 36: N = 120, n_i = 10, ICC = 0.25, θ = 0
Method
QIF
GEE1
GzLMM
Estimates of θ

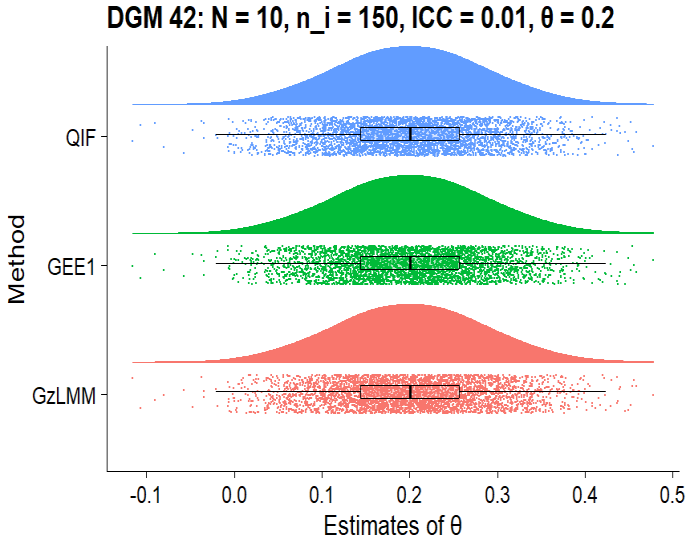

DGM 37: N = 120, n_i = 80, ICC = 0.001, θ = 0
QIF
GEE1
GzLMM
Method
-0.04
0.00
0.04
0.08
Estimates of θ

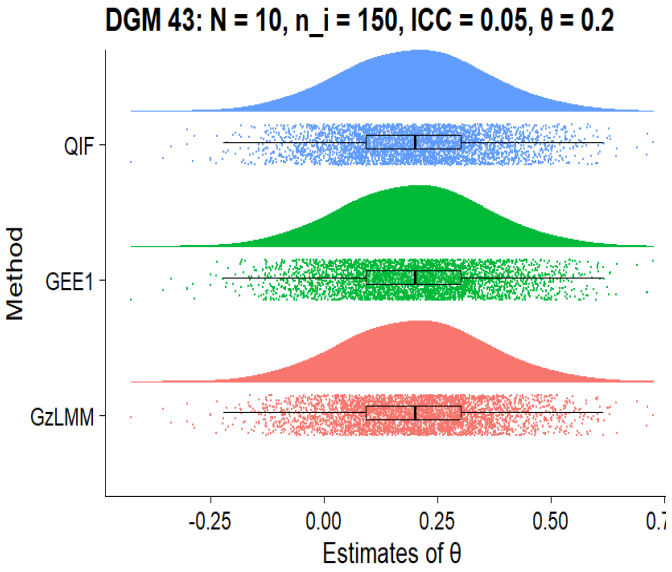

DGM 38: N = 120, n_i = 80, ICC = 0.01, θ = 0
QIF
GEE1
GzLMM
Method
-0.10
-0.05
0.00
0.05
0.10
Estimates of θ

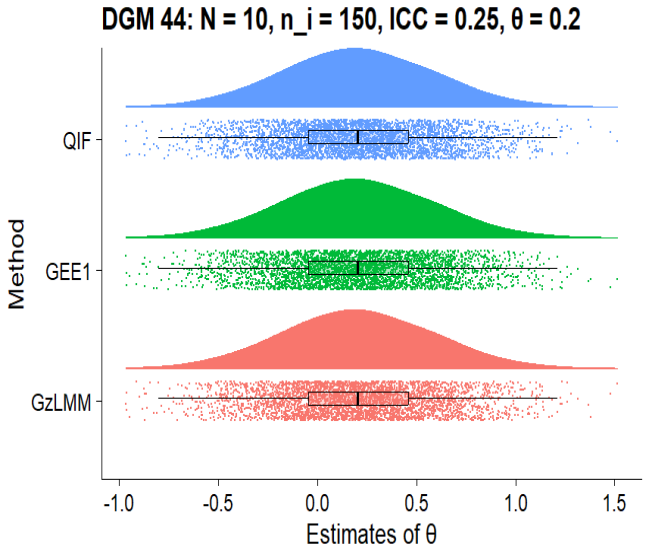

DGM 39: N = 120, n_i = 80, ICC = 0.05, θ = 0
QIF
GEE1
GzLMM
Method
-0.1
0.0
0.1
Estimates of θ

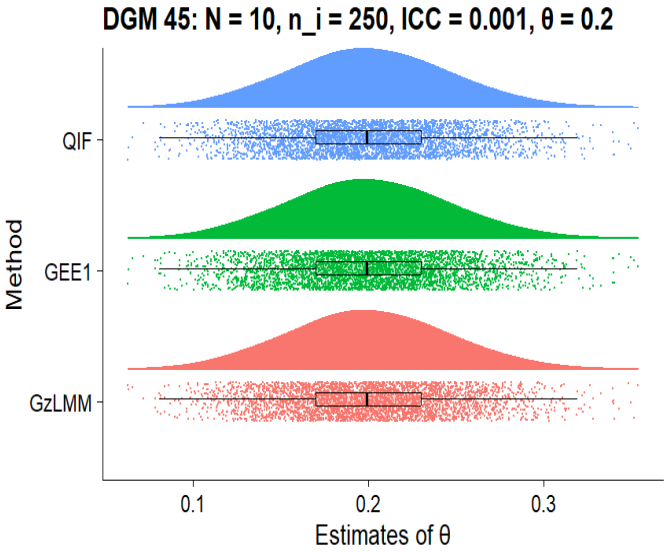

DGM 40: N = 120, n_i = 80, ICC = 0.25, θ = 0
QIF
GEE1
GzLMM
Method
-0.25
0.00
0.25
Estimates of θ

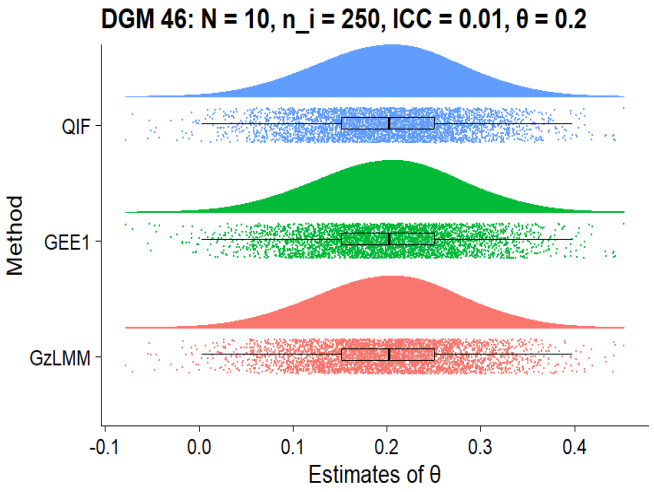

DGM 41: N = 10, n_i = 150, ICC = 0.001, θ = 0.2
QIF
GEE1
GzLMM
Method
0.0
0.1
0.2
0.3
0
Estimates of θ

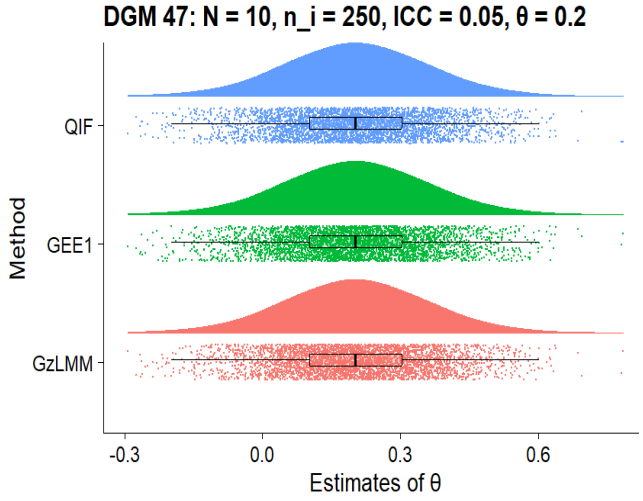

DGM 42: N = 10, n_i = 150, ICC = 0.01, θ = 0.2
QIF
GEE1
GzLMM
Method
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
Estimates of θ

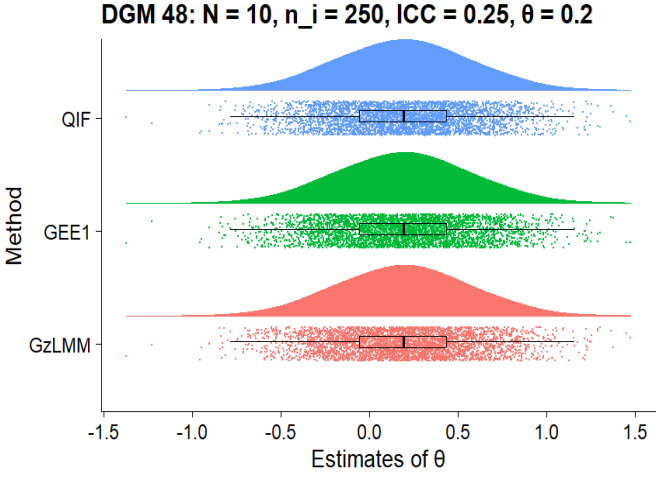

DGM 43: N = 10, n_i = 150, ICC = 0.05, θ = 0.2
QIF
GEE1
GzLMM
Method
-0.25
0.00
0.25
0.50
Estimates of θ

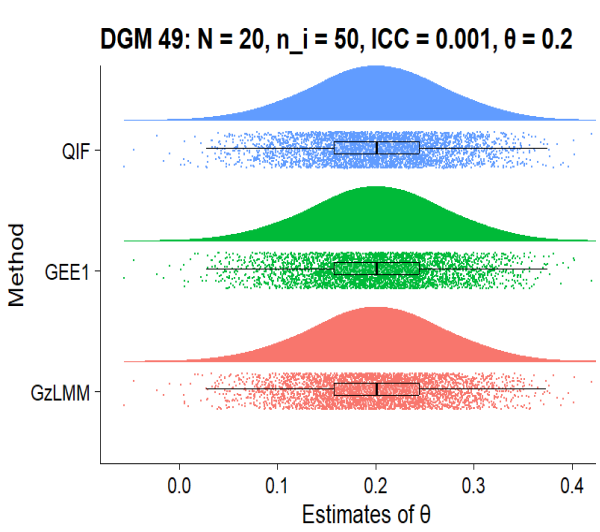

DGM 44: N = 10, n_i = 150, ICC = 0.25, θ = 0.2
QIF
GEE1
GzLMM
Method
-1.0
-0.5
0.0
0.5
1.0
1.5
Estimates of θ

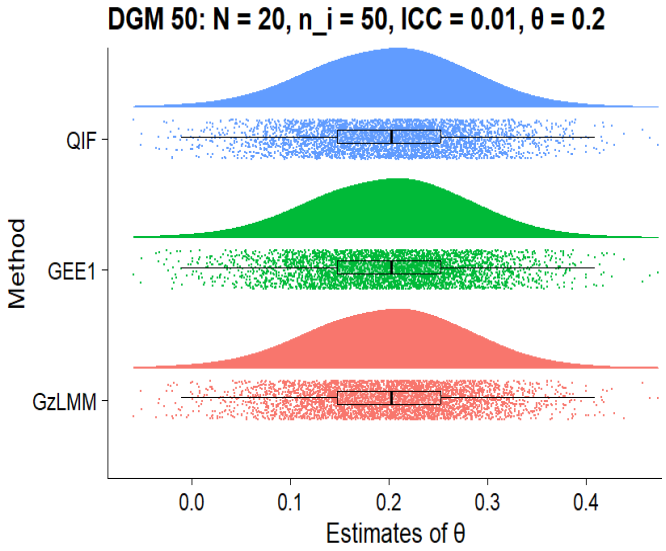

DGM 45: N = 10, n_i = 250, ICC = 0.001, θ = 0.2
QIF
GEE1
GzLMM
Method
0.1
0.2
0.3
Estimates of θ

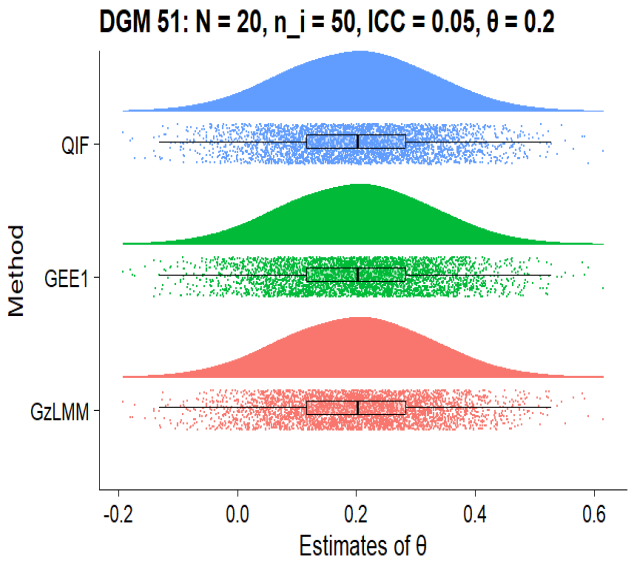

DGM 46: N = 10, n_i = 250, ICC = 0.01, θ = 0.2
QIF
GEE1
GzLMM
Method
-0.1
0.0
0.1
0.2
0.3
0.4
Estimates of θ

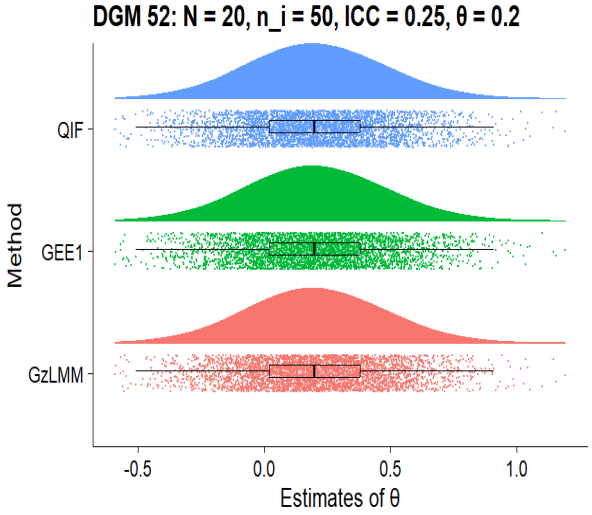

DGM 47: N = 10, n_i = 250, ICC = 0.05, θ = 0.2
QIF
GEE1
GzLMM
Method
-0.3
0.0
0.3
0.6
Estimates of θ

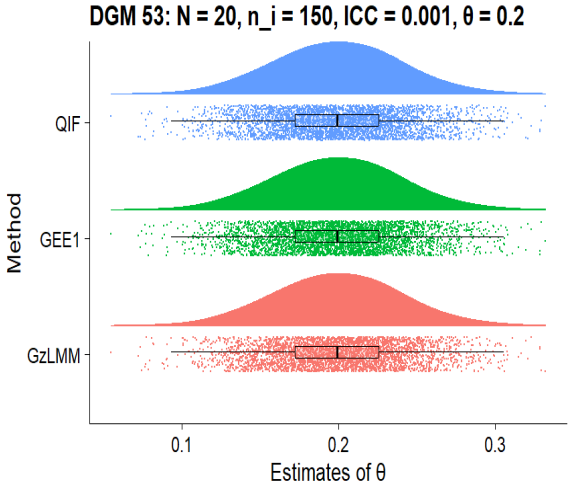

DGM 48: N = 10, n_i = 250, ICC = 0.25, θ = 0.2
QIF
GEE1
GzLMM
Method
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Estimates of θ

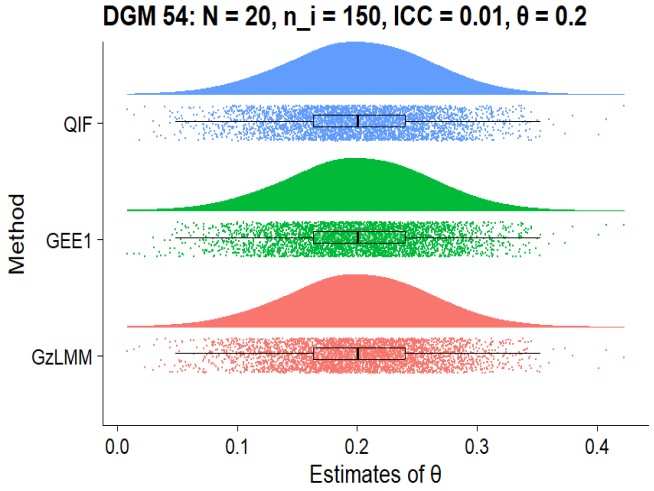

DGM 49: N = 20, n_i = 50, ICC = 0.001, θ = 0.2
Method
QIF
GEE1
GzLMM
0.0
0.1
0.2
0.3
0.4
Estimates of θ

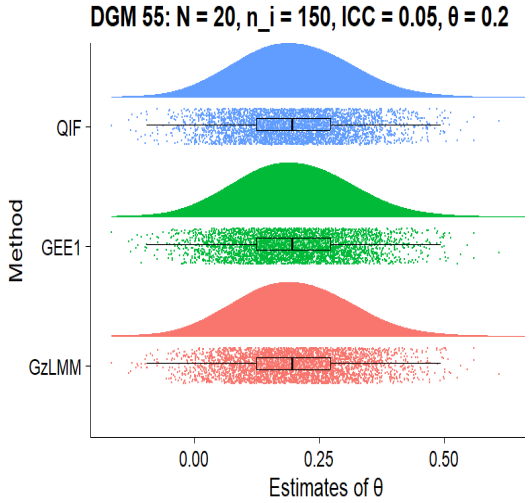

DGM 50: N = 20, n_i = 50, ICC = 0.01, θ = 0.2
Method
QIF
GEE1
GzLMM
0.0
0.1
0.2
0.3
0.4
Estimates of θ

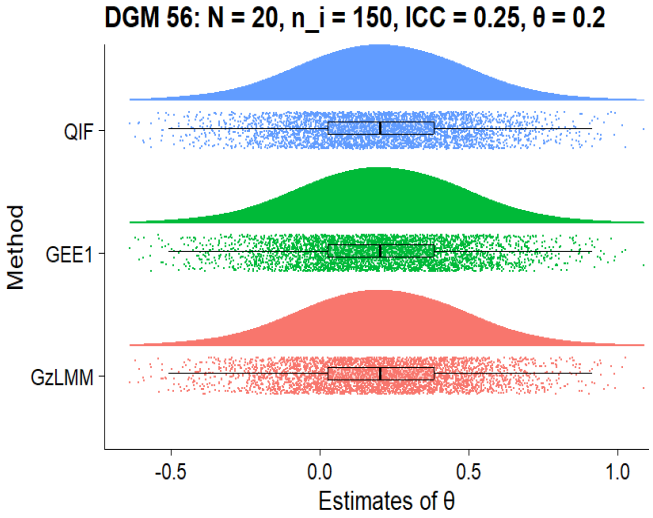

DGM 51: N = 20, n_i = 50, ICC = 0.05, θ = 0.2
Method
QIF
GEE1
GzLMM
-0.2
0.0
0.2
0.4
0.6
Estimates of θ

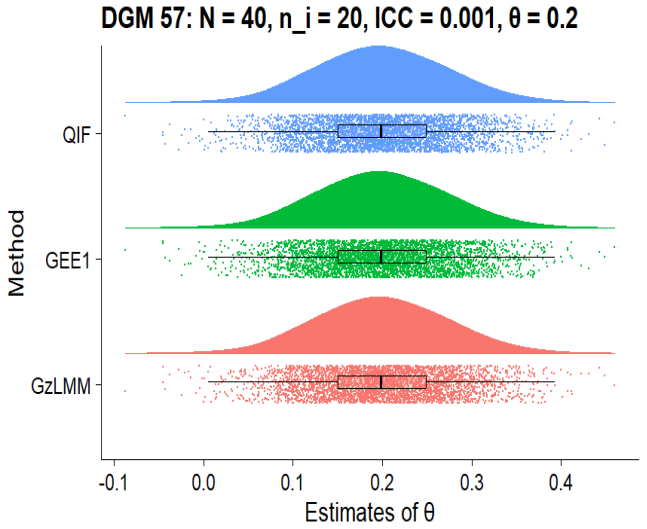

DGM 52: N = 20, n_i = 50, ICC = 0.25, θ = 0.2
Method
QIF
GEE1
GzLMM
-0.5
0.0
0.5
1.0
Estimates of θ

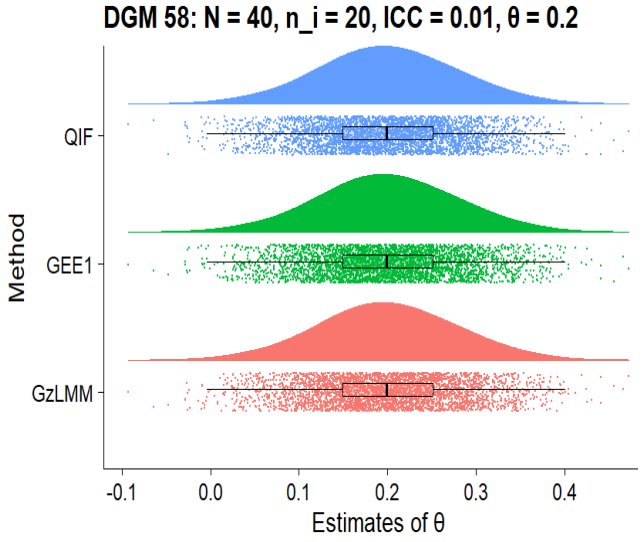

DGM 53: N = 20, n_i = 150, ICC = 0.001, θ = 0.2
Method
QIF
GEE1
GzLMM
0.1
0.2
0.3
Estimates of θ

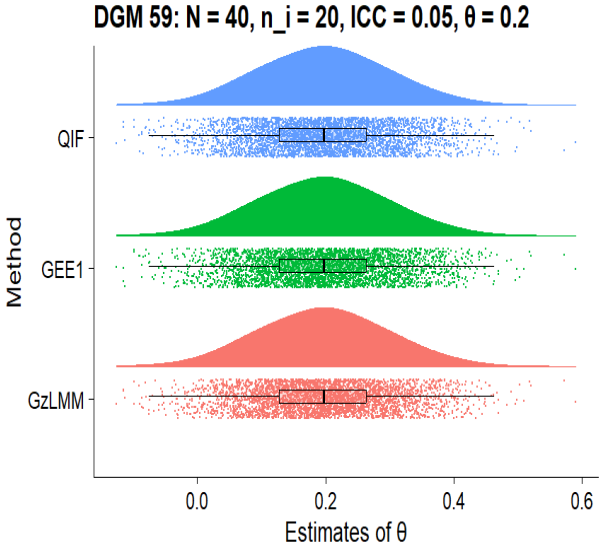

DGM 54: N = 20, n_i = 150, ICC = 0.01, θ = 0.2
Method
QIF
GEE1
GzLMM
0.0
0.1
0.2
0.3
0.4
Estimates of θ

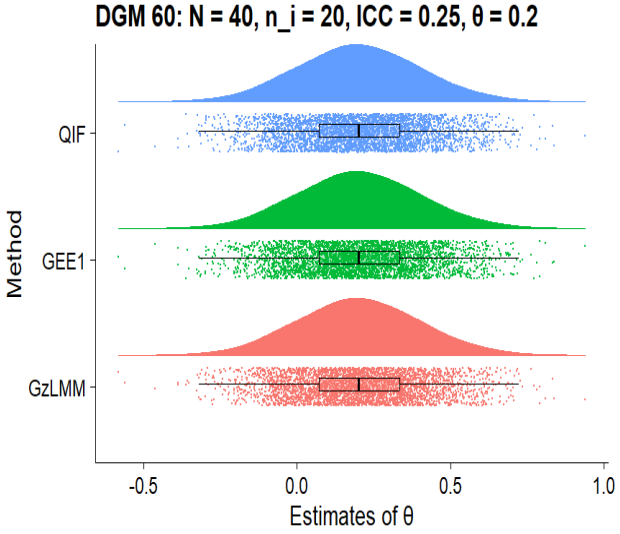

DGM 55: N = 20, n_i = 150, ICC = 0.05, θ = 0.2
QIF
GEE1
GzLMM
Method
0.00
0.25
0.50
Estimates of θ

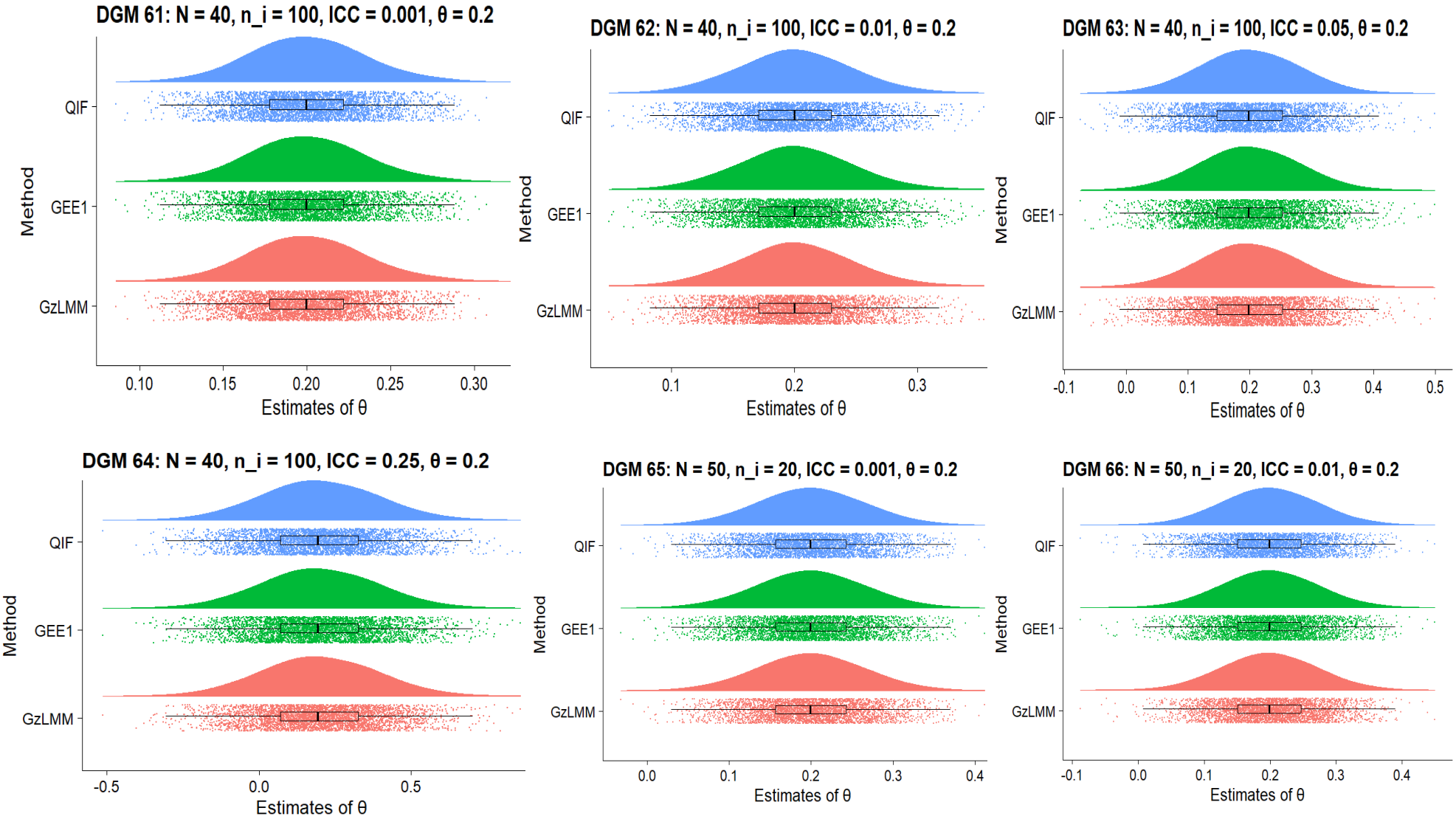

DGM 56: N = 20, n_i = 150, ICC = 0.25, θ = 0.2
QIF
GEE1
GzLMM
Method
-0.5
0.0
0.5
1.0
Estimates of θ

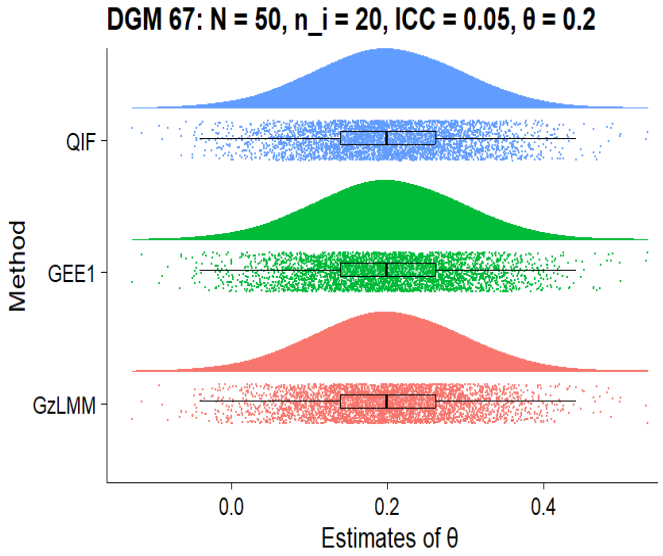

DGM 57: N = 40, n_i = 20, ICC = 0.001, θ = 0.2
QIF
GEE1
GzLMM
Method
-0.1
0.0
0.1
0.2
0.3
0.4
Estimates of θ

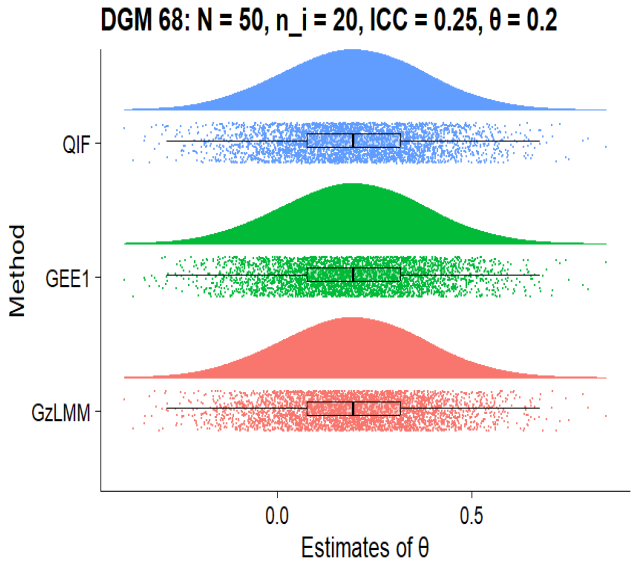

DGM 58: N = 40, n_i = 20, ICC = 0.01, θ = 0.2
QIF
GEE1
GzLMM
Method
-0.1
0.0
0.1
0.2
0.3
0.4
Estimates of θ

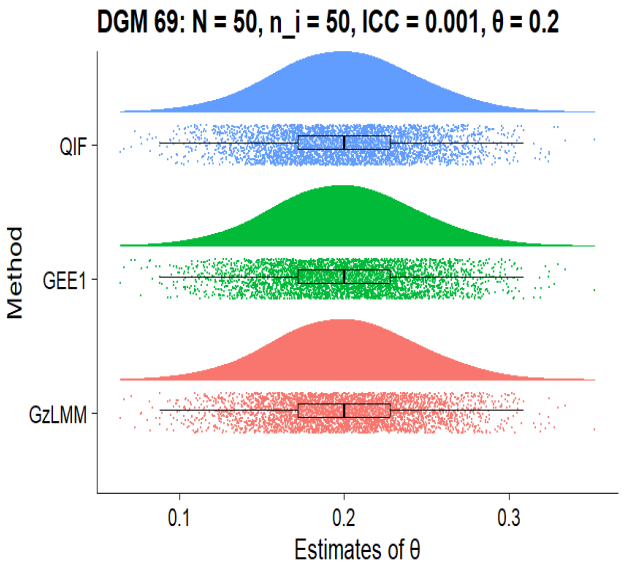

DGM 59: N = 40, n_i = 20, ICC = 0.05, θ = 0.2
QIF
GEE1
GzLMM
Method
0.0
0.2
0.4
0.6
Estimates of θ

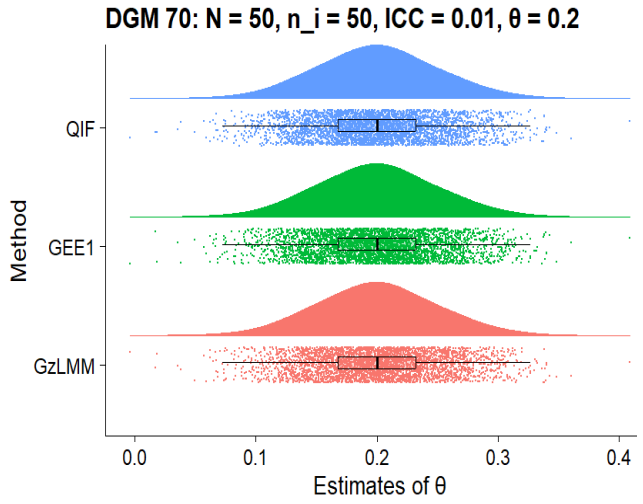

DGM 60: N = 40, n_i = 20, ICC = 0.25, θ = 0.2
QIF
GEE1
GzLMM
Method
-0.5
0.0
0.5
1.0
Estimates of θ

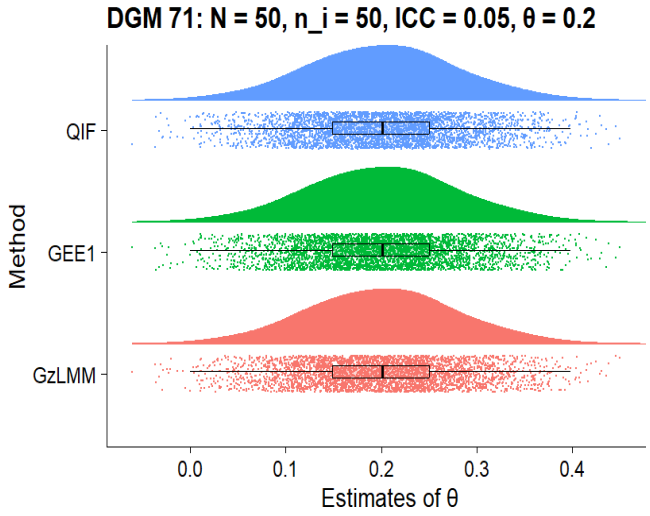

DGM 61: N = 40, n_i = 100, ICC = 0.001, θ = 0.2
DGM 62: N = 40, n_i = 100, ICC = 0.01, θ = 0.2
DGM 63: N = 40, n_i = 100, ICC = 0.05, θ = 0.2
DGM 64: N = 40, n_i = 100, ICC = 0.25, θ = 0.2
DGM 65: N = 50, n_i = 20, ICC = 0.001, θ = 0.2
DGM 66: N = 50, n_i = 20, ICC = 0.01, θ = 0.2
Method
QIF
GEE1
GzLMM
Estimates of θ

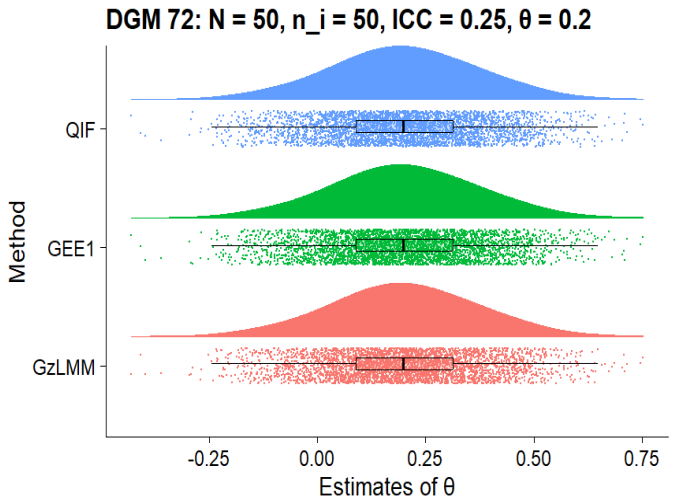

DGM 67: N = 50, n_i = 20, ICC = 0.05, θ = 0.2
QIF
GEE1
GzLMM
Method
0.0
0.2
0.4
Estimates of θ

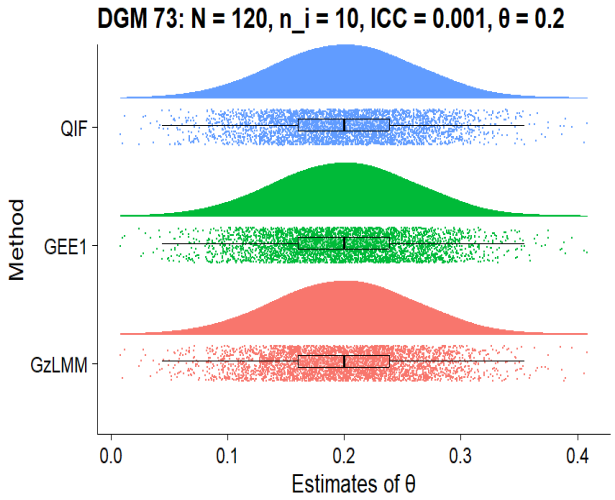

DGM 68: N = 50, n_i = 20, ICC = 0.25, θ = 0.2
QIF
GEE1
GzLMM
Method
0.0
0.5
Estimates of θ

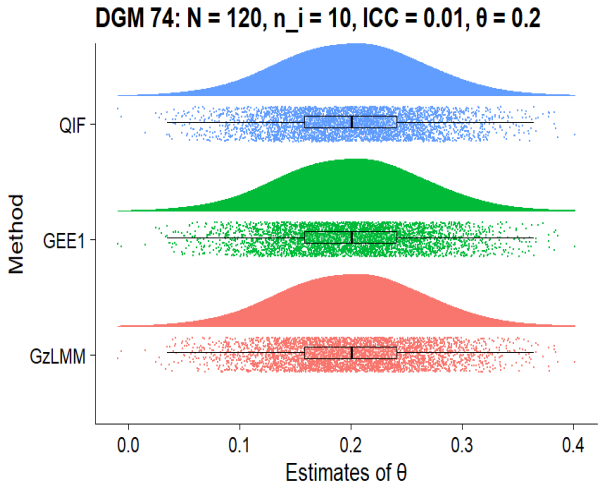

DGM 69: N = 50, n_i = 50, ICC = 0.001, θ = 0.2
QIF
GEE1
GzLMM
Method
0.1
0.2
0.3
Estimates of θ

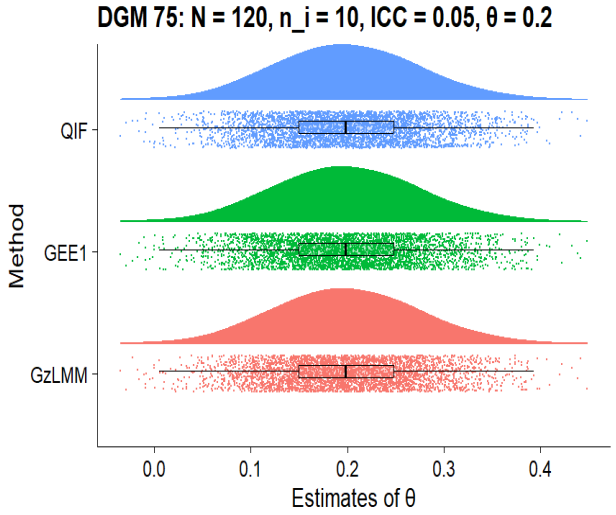

DGM 70: N = 50, n_i = 50, ICC = 0.01, θ = 0.2
QIF
GEE1
GzLMM
Method
0.0
0.1
0.2
0.3
0.4
Estimates of θ

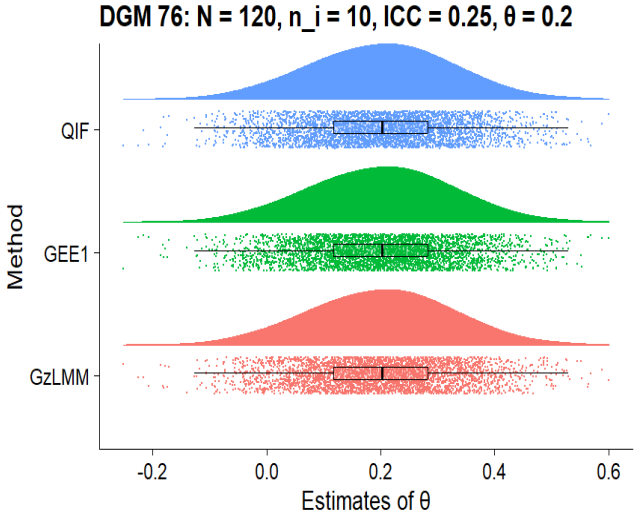

DGM 71: N = 50, n_i = 50, ICC = 0.05, θ = 0.2
QIF
GEE1
GzLMM
Method
0.0
0.1
0.2
0.3
0.4
Estimates of θ

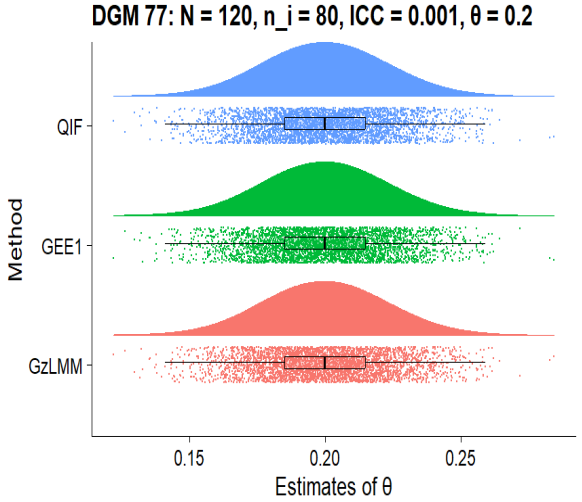

DGM 72: N = 50, n_i = 50, ICC = 0.25, θ = 0.2
QIF
GEE1
GzLMM
Method
-0.25
0.00
0.25
0.50
0.75
Estimates of θ

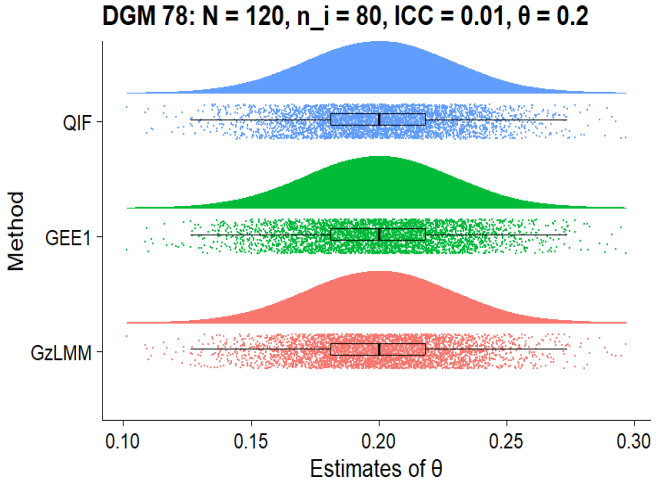

DGM 73: N = 120, n_i = 10, ICC = 0.001, θ = 0.2
Method
QIF
GEE1
GzLMM
0.0
0.1
0.2
0.3
0.4
Estimates of θ

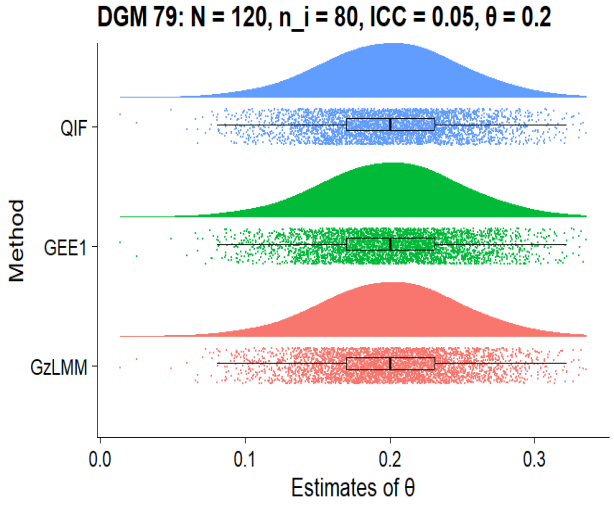

DGM 74: N = 120, n_i = 10, ICC = 0.01, θ = 0.2
Method
QIF
GEE1
GzLMM
0.0
0.1
0.2
0.3
0.4
Estimates of θ

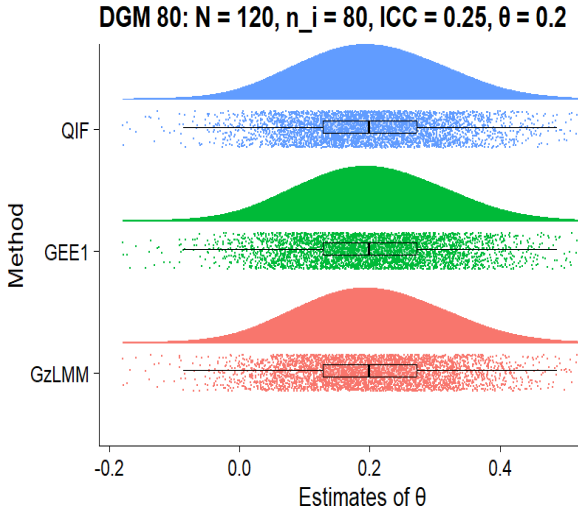

DGM 75: N = 120, n_i = 10, ICC = 0.05, θ = 0.2
Method
QIF
GEE1
GzLMM
0.0
0.1
0.2
0.3
0.4
Estimates of θ

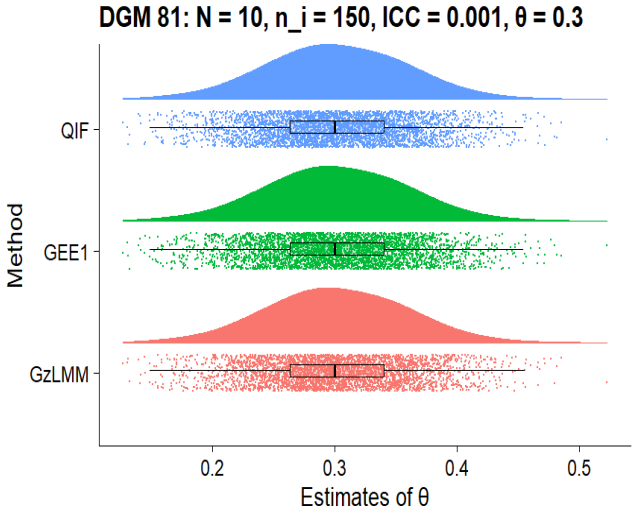

DGM 76: N = 120, n_i = 10, ICC = 0.25, θ = 0.2
Method
QIF
GEE1
GzLMM
-0.2
0.0
0.2
0.4
0.6
Estimates of θ

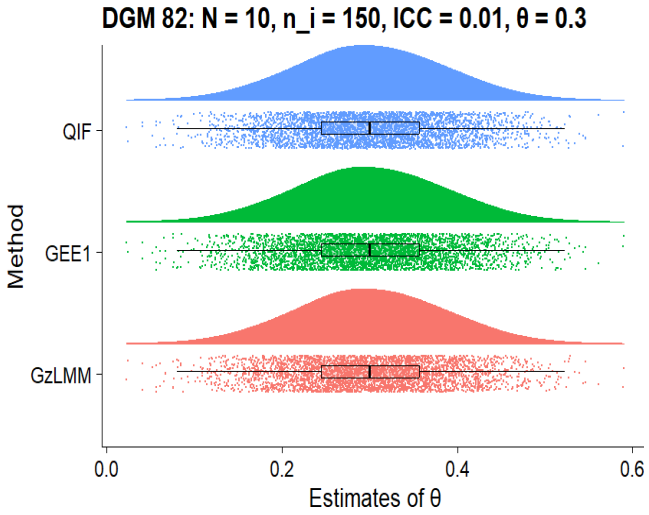

DGM 77: N = 120, n_i = 80, ICC = 0.001, θ = 0.2
Method
QIF
GEE1
GzLMM
0.15
0.20
0.25
Estimates of θ

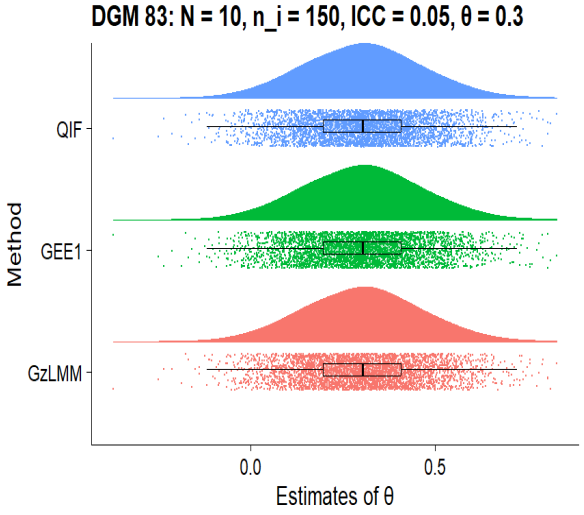

DGM 78: N = 120, n_i = 80, ICC = 0.01, θ = 0.2
Method
QIF
GEE1
GzLMM
0.10
0.15
0.20
0.25
0.30
Estimates of θ

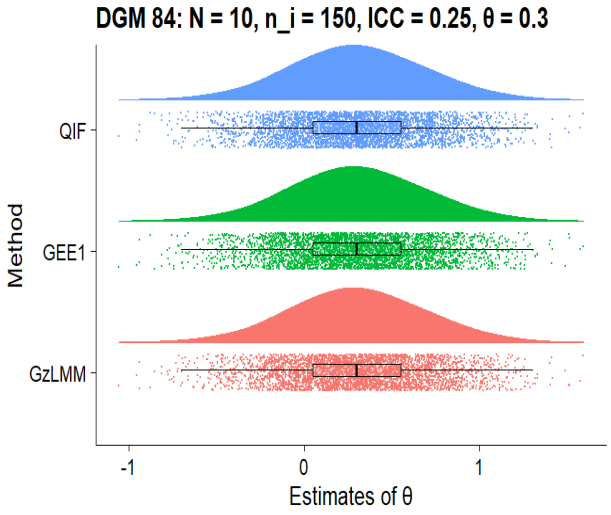

DGM 79: N = 120, n_i = 80, ICC = 0.05, θ = 0.2
QIF
GEE1
GzLMM
Method
0.0
0.1
0.2
0.3
Estimates of θ

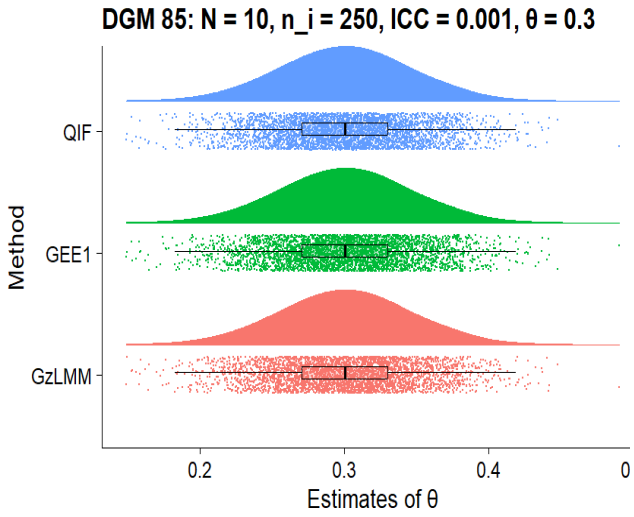

DGM 80: N = 120, n_i = 80, ICC = 0.25, θ = 0.2
QIF
GEE1
GzLMM
Method
-0.2
0.0
0.2
0.4
Estimates of θ

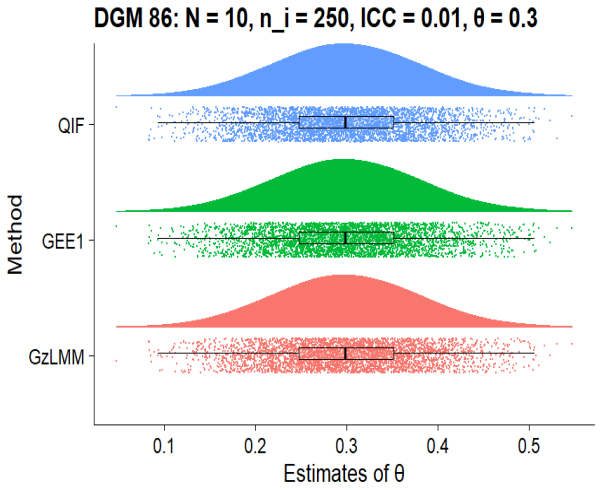

DGM 81: N = 10, n_i = 150, ICC = 0.001, θ = 0.3
QIF
GEE1
GzLMM
Method
0.2
0.3
0.4
0.5
Estimates of θ

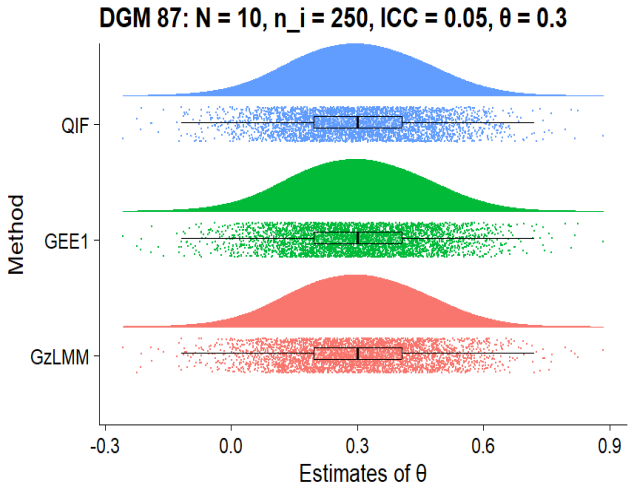

DGM 82: N = 10, n_i = 150, ICC = 0.01, θ = 0.3
QIF
GEE1
GzLMM
Method
0.0
0.2
0.4
0.6
Estimates of θ

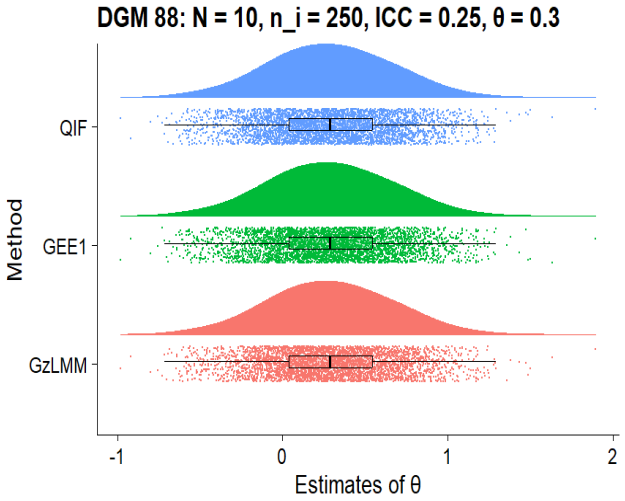

DGM 83: N = 10, n_i = 150, ICC = 0.05, θ = 0.3
QIF
GEE1
GzLMM
Method
0.0
0.5
Estimates of θ

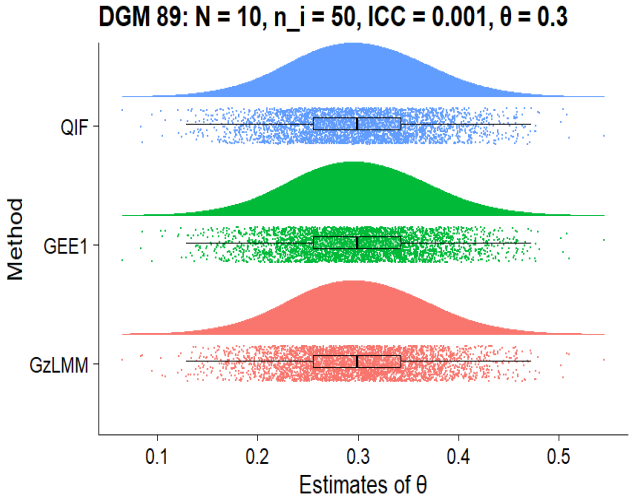

DGM 84: N = 10, n_i = 150, ICC = 0.25, θ = 0.3
QIF
GEE1
GzLMM
Method
-1
0
1
Estimates of θ

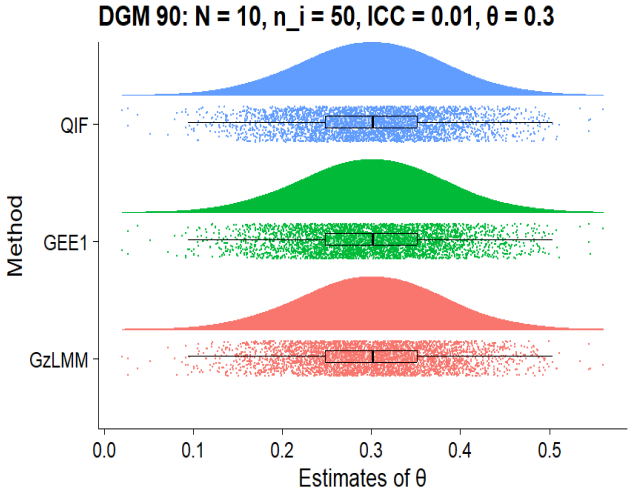

DGM 85: N = 10, n_i = 250, ICC = 0.001, θ = 0.3
QIF
GEE1
GzLMM
Method
0.2
0.3
0.4
0
Estimates of θ

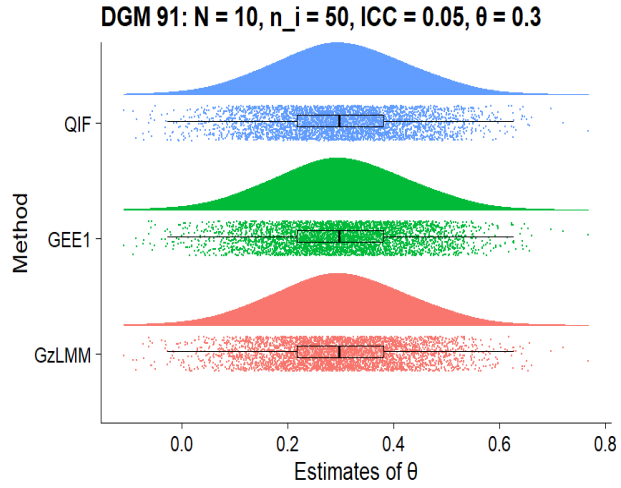

DGM 86: N = 10, n_i = 250, ICC = 0.01, θ = 0.3
QIF
GEE1
GzLMM
Method
0.1
0.2
0.3
0.4
0.5
Estimates of θ

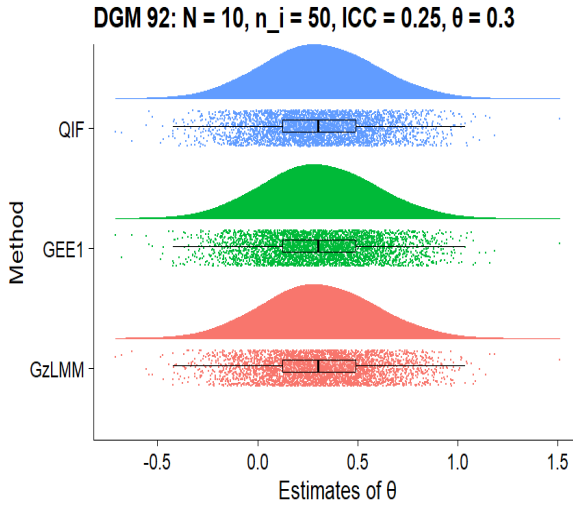

DGM 87: N = 10, n_i = 250, ICC = 0.05, θ = 0.3
QIF
GEE1
GzLMM
Method
-0.3
0.0
0.3
0.6
0.9
Estimates of θ

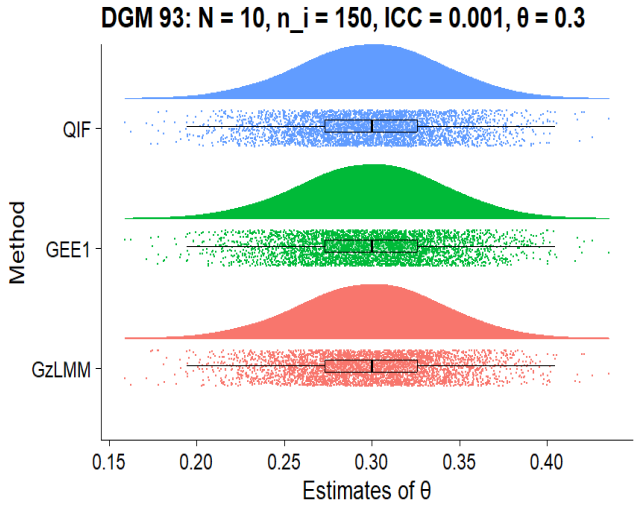

DGM 88: N = 10, n_i = 250, ICC = 0.25, θ = 0.3
QIF
GEE1
GzLMM
Method
-1
0
1
2
Estimates of θ

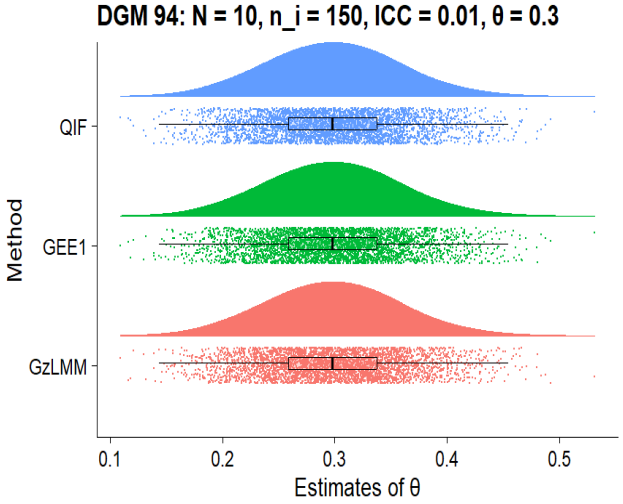

DGM 89: N = 10, n_i = 50, ICC = 0.001, θ = 0.3
QIF
GEE1
GzLMM
Method
0.1
0.2
0.3
0.4
0.5
Estimates of θ

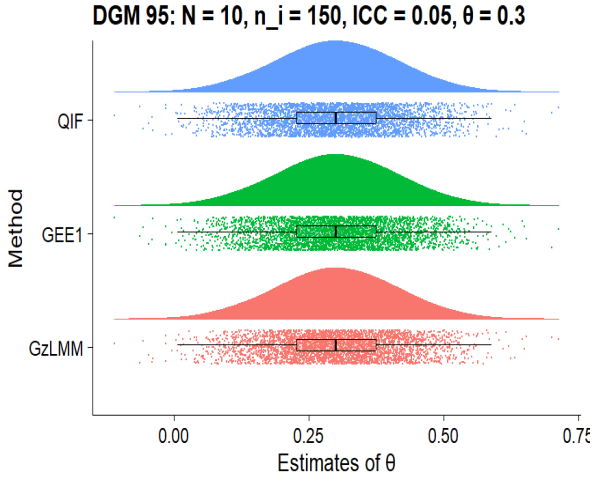

DGM 90: N = 10, n_i = 50, ICC = 0.01, θ = 0.3
QIF
GEE1
GzLMM
Method
0.0
0.1
0.2
0.3
0.4
0.5
Estimates of θ

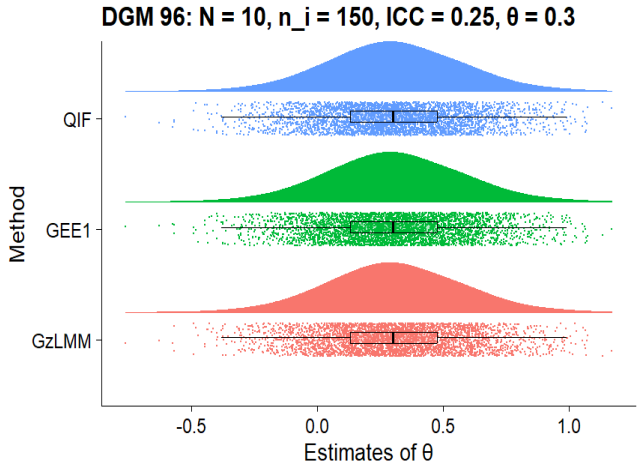

DGM 91: N = 10, n_i = 50, ICC = 0.05, θ = 0.3
QIF
GEE1
GzLMM
Method
0.0
0.2
0.4
0.6
0.8
Estimates of θ

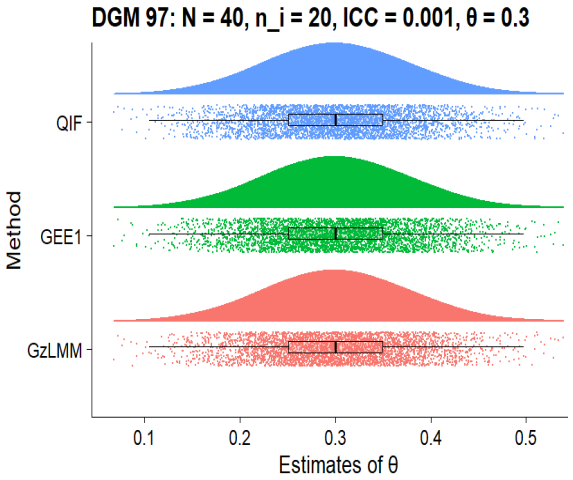

DGM 92: N = 10, n_i = 50, ICC = 0.25, θ = 0.3
QIF
GEE1
GzLMM
Method
-0.5
0.0
0.5
1.0
1.5
Estimates of θ

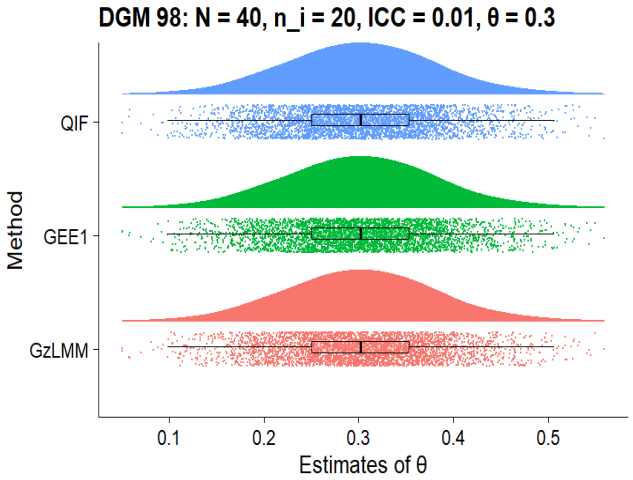

DGM 93: N = 10, n_i = 150, ICC = 0.001, θ = 0.3
QIF
GEE1
GzLMM
Method
0.15
0.20
0.25
0.30
0.35
0.40
Estimates of θ

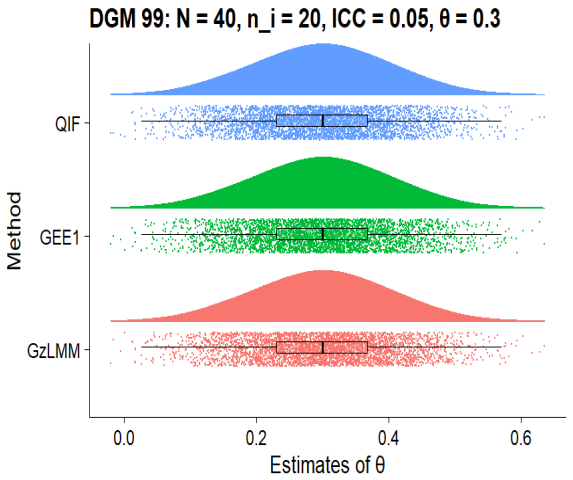

DGM 94: N = 10, n_i = 150, ICC = 0.01, θ = 0.3
QIF
GEE1
GzLMM
Method
0.1
0.2
0.3
0.4
0.5
Estimates of θ

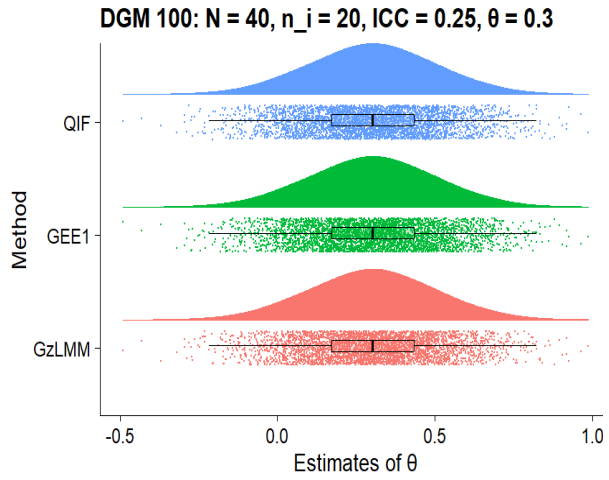

DGM 95: N = 10, n_i = 150, ICC = 0.05, θ = 0.3
QIF
GEE1
GzLMM
Method
0.00
0.25
0.50
Estimates of θ

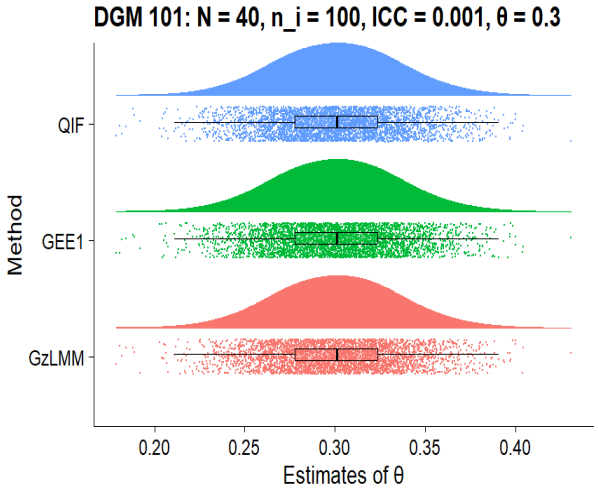

DGM 96: N = 10, n_i = 150, ICC = 0.25, θ = 0.3
QIF
GEE1
GzLMM
Method
-0.5
0.0
0.5
1.0
Estimates of θ

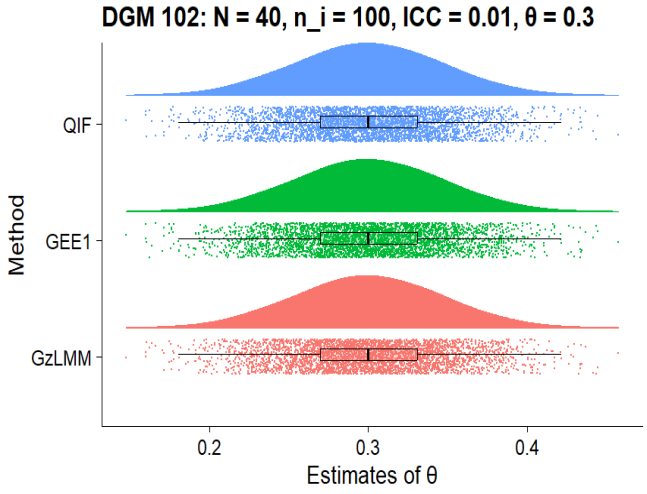

DGM 97: N = 40, n_i = 20, ICC = 0.001, θ = 0.3
Method
QIF
GEE1
GzLMM
0.1
0.2
0.3
0.4
0.5
Estimates of θ

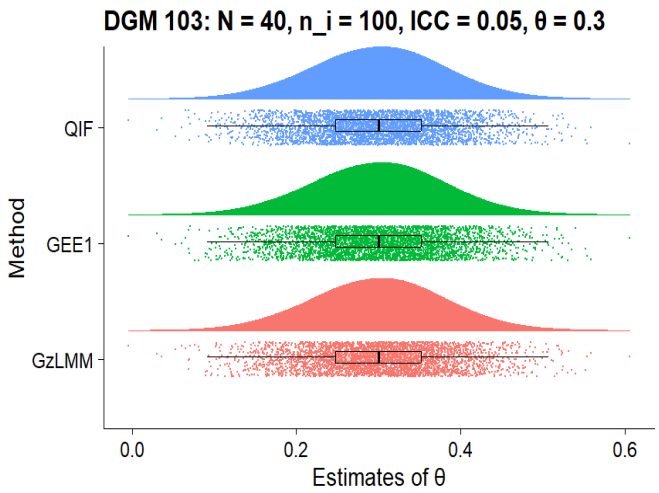

DGM 98: N = 40, n_i = 20, ICC = 0.01, θ = 0.3
Method
QIF
GEE1
GzLMM
0.1
0.2
0.3
0.4
0.5
Estimates of θ

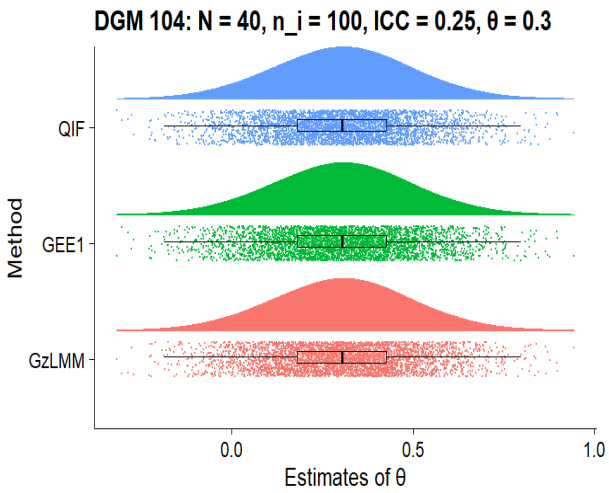

DGM 99: N = 40, n_i = 20, ICC = 0.05, θ = 0.3
Method
QIF
GEE1
GzLMM
0.0
0.2
0.4
0.6
Estimates of θ

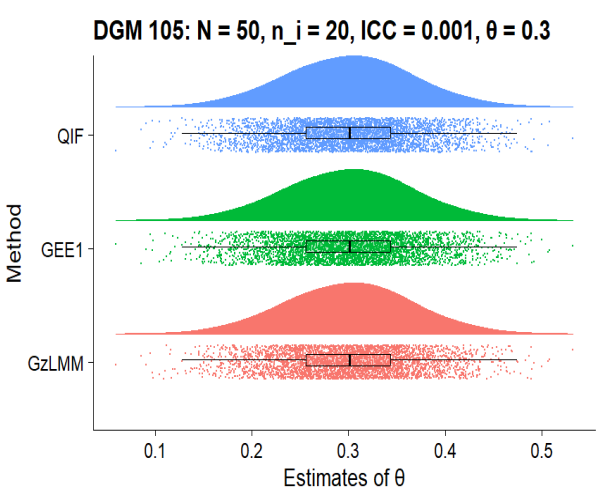

DGM 100: N = 40, n_i = 20, ICC = 0.25, θ = 0.3
Method
QIF
GEE1
GzLMM
-0.5
0.0
0.5
1.0
Estimates of θ

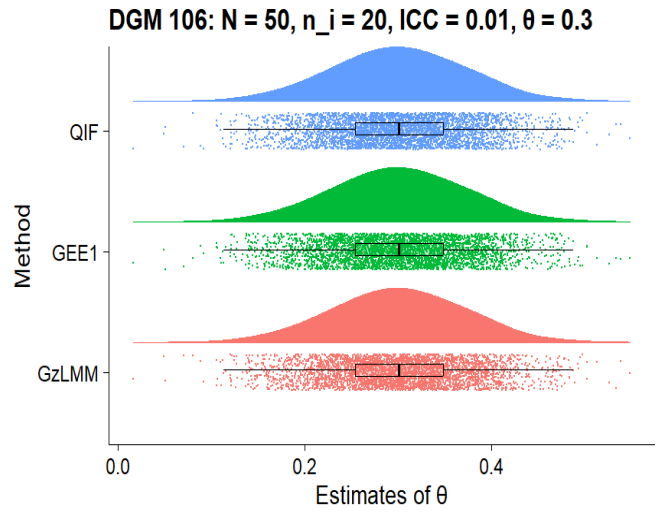

DGM 101: N = 40, n_i = 100, ICC = 0.001, θ = 0.3
Method
QIF
GEE1
GzLMM
0.20
0.25
0.30
0.35
0.40
Estimates of θ

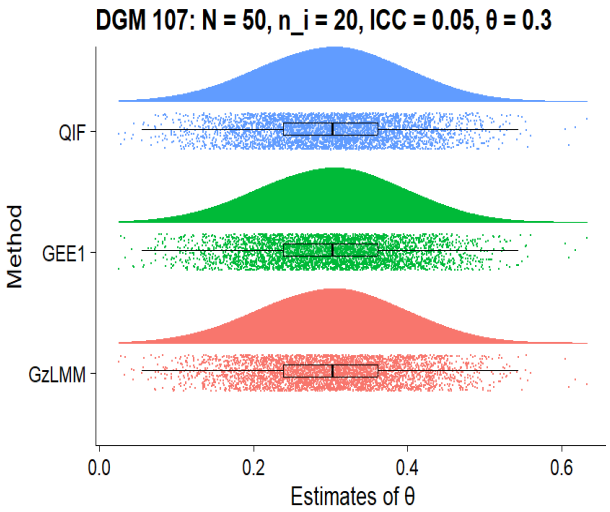

DGM 102: N = 40, n_i = 100, ICC = 0.01, θ = 0.3
Method
QIF
GEE1
GzLMM
0.2
0.3
0.4
Estimates of θ

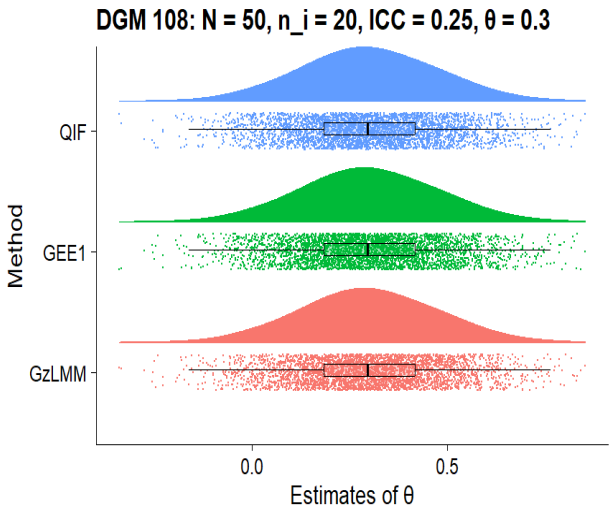

DGM 103: N = 40, n_i = 100, ICC = 0.05, θ = 0.3
Method
QIF
GEE1
GzLMM
0.0
0.2
0.4
0.6
Estimates of θ

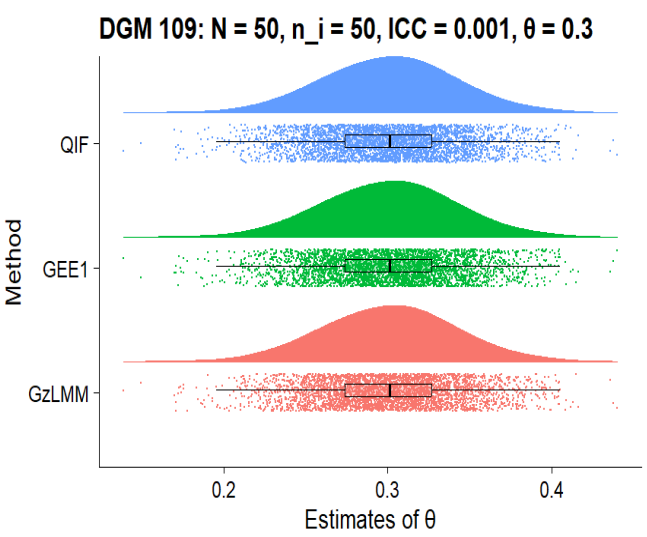

DGM 104: N = 40, n_i = 100, ICC = 0.25, θ = 0.3
Method
QIF
GEE1
GzLMM
0.0
0.5
1.0
Estimates of θ

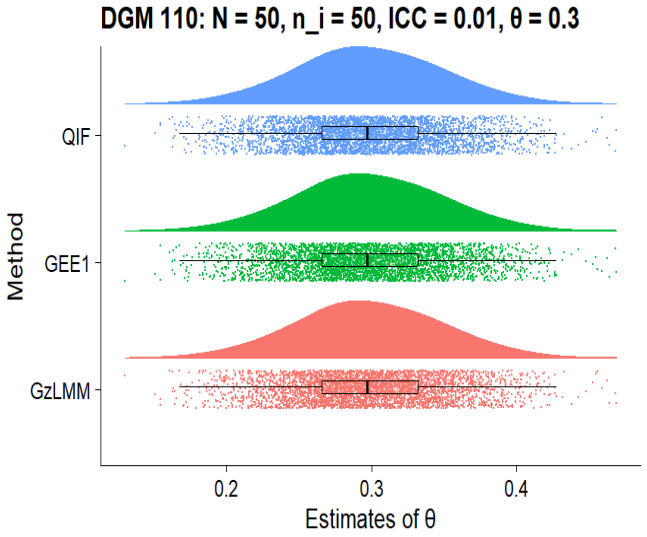

DGM 105: N = 50, n_i = 20, ICC = 0.001, θ = 0.3
Method
QIF
GEE1
GzLMM
0.1
0.2
0.3
0.4
0.5
Estimates of θ

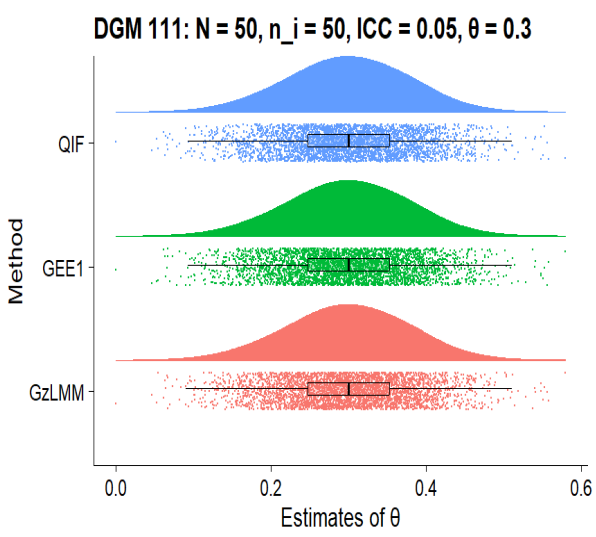

DGM 106: N = 50, n_i = 20, ICC = 0.01, θ = 0.3
Method
QIF
GEE1
GzLMM
0.0
0.2
0.4
Estimates of θ

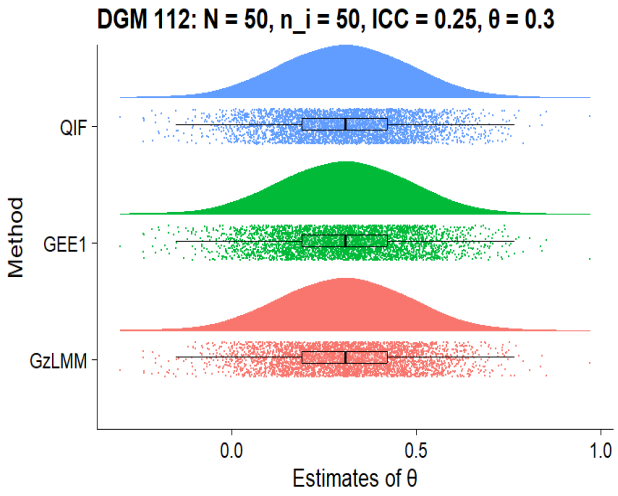

DGM 107: N = 50, n_i = 20, ICC = 0.05, θ = 0.3
Method
QIF
GEE1
GzLMM
0.0
0.2
0.4
0.6
Estimates of θ

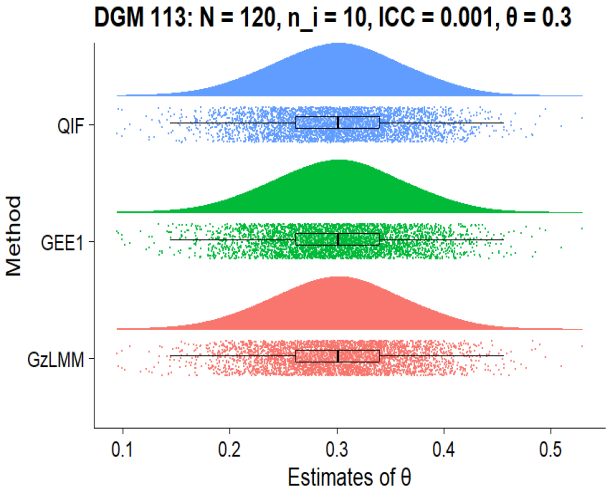

DGM 108: N = 50, n_i = 20, ICC = 0.25, θ = 0.3
Method
QIF
GEE1
GzLMM
0.0
0.5
Estimates of θ

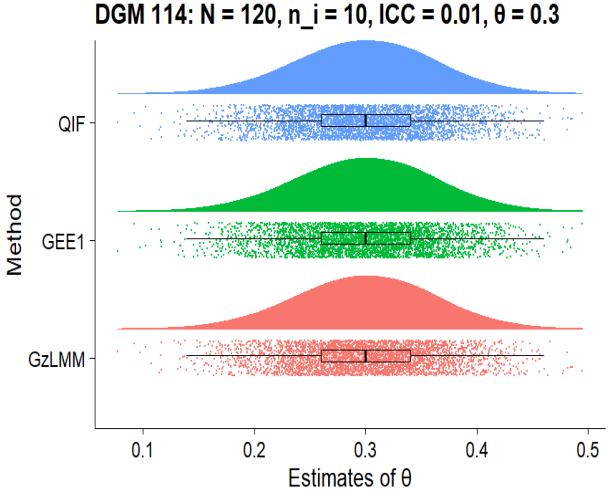

DGM 109: N = 50, n_i = 50, ICC = 0.001, θ = 0.3
QIF
GEE1
GzLMM
Method
0.2
0.3
0.4
Estimates of θ

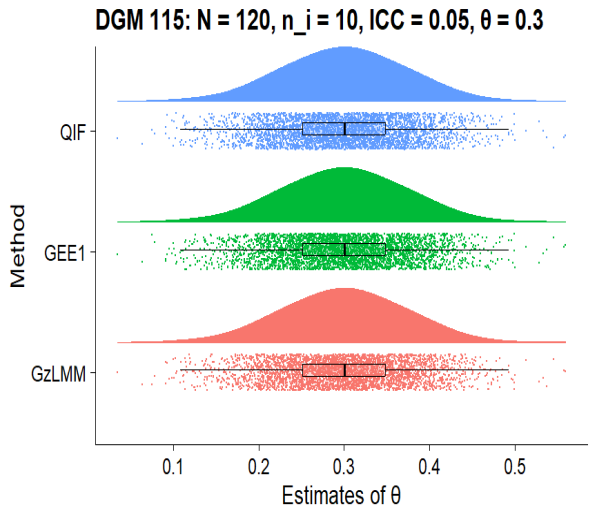

DGM 110: N = 50, n_i = 50, ICC = 0.01, θ = 0.3
QIF
GEE1
GzLMM
Method
0.2
0.3
0.4
Estimates of θ

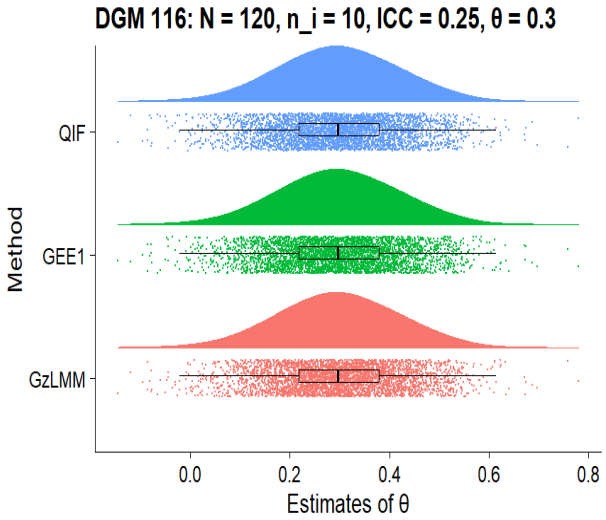

DGM 111: N = 50, n_i = 50, ICC = 0.05, θ = 0.3
QIF
GEE1
GzLMM
Method
0.0
0.2
0.4
0.6
Estimates of θ

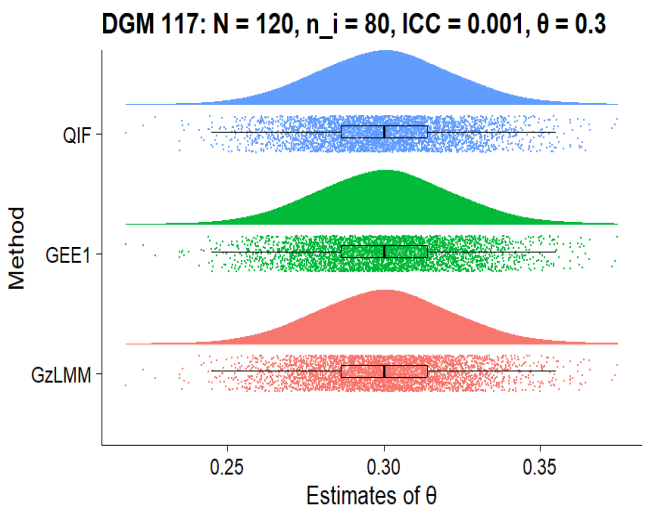

DGM 112: N = 50, n_i = 50, ICC = 0.25, θ = 0.3
QIF
GEE1
GzLMM
Method
0.0
0.5
1.0
Estimates of θ

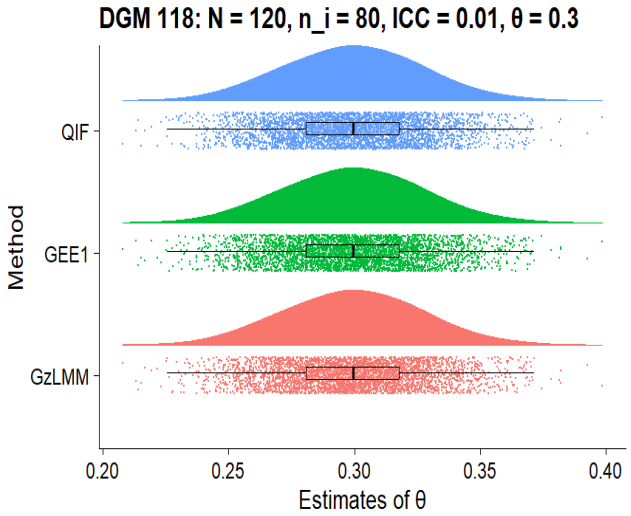

DGM 113: N = 120, n_i = 10, ICC = 0.001, θ = 0.3
QIF
GEE1
GzLMM
Method
0.1
0.2
0.3
0.4
0.5
Estimates of θ

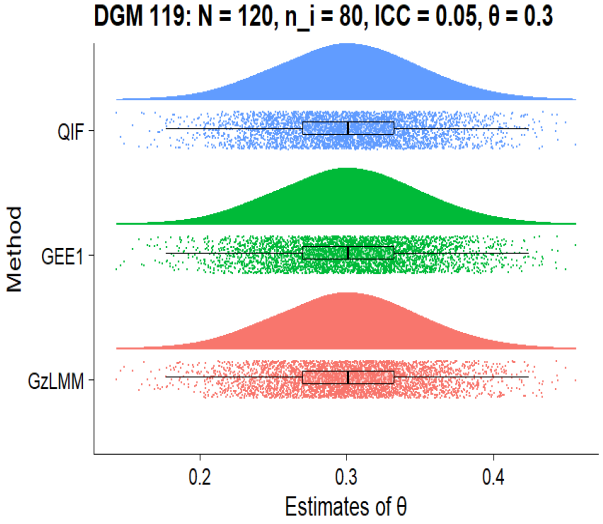

DGM 114: N = 120, n_i = 10, ICC = 0.01, θ = 0.3
QIF
GEE1
GzLMM
Method
0.1
0.2
0.3
0.4
0.5
Estimates of θ

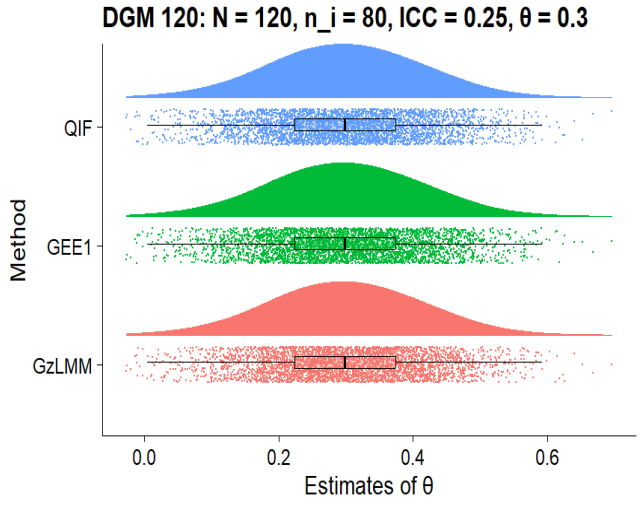

DGM 115: N = 120, n_i = 10, ICC = 0.05, θ = 0.3
Method
QIF
GEE1
GzLMM
0.1
0.2
0.3
0.4
0.5
Estimates of θ


DGM 116: N = 120, n_i = 10, ICC = 0.25, θ = 0.3
Method
QIF
GEE1
GzLMM
0.0
0.2
0.4
0.6
0.8
Estimates of θ


DGM 117: N = 120, n_i = 80, ICC = 0.001, θ = 0.3
Method
QIF
GEE1
GzLMM
0.25
0.30
0.35
Estimates of θ

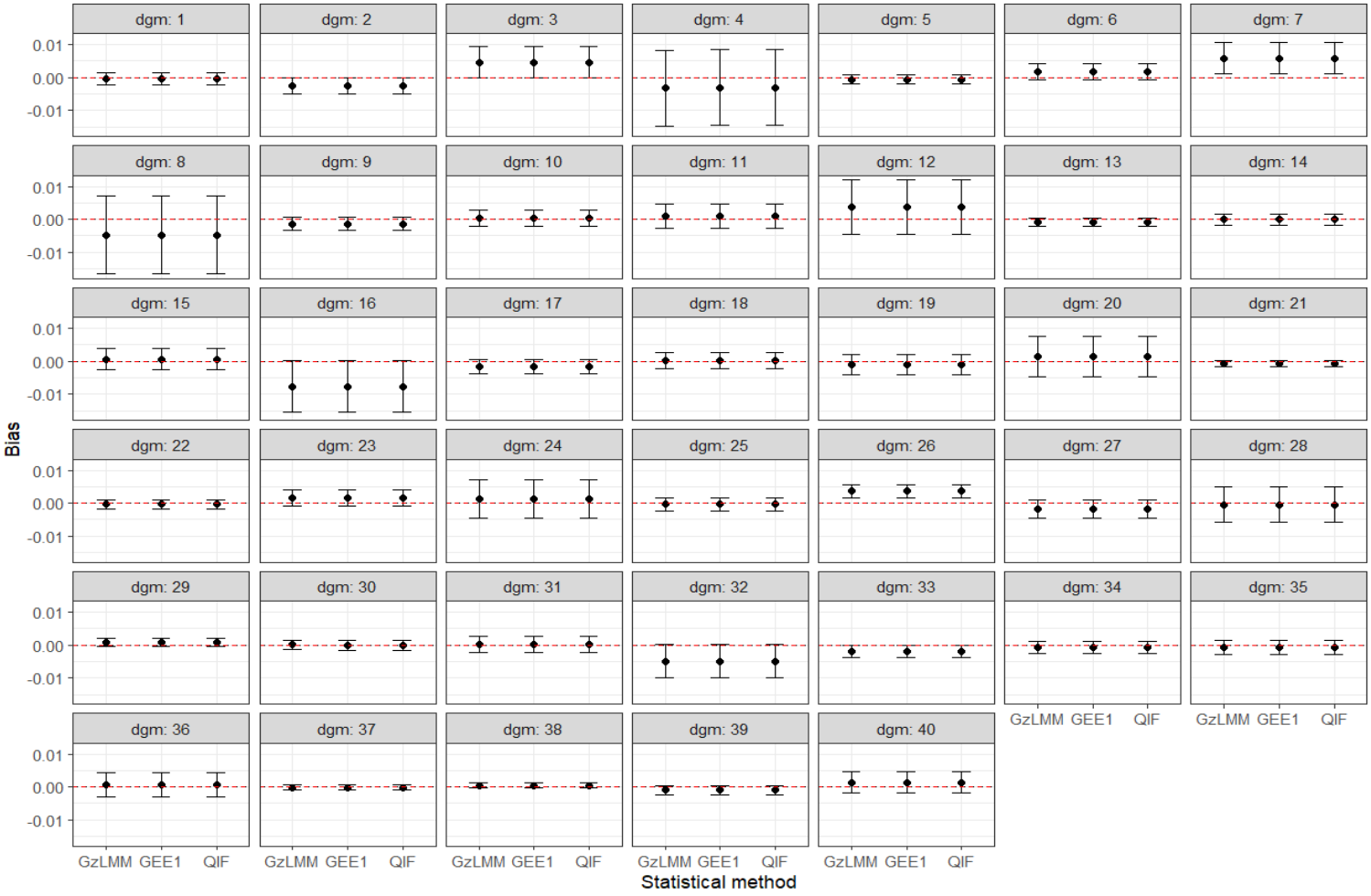

DGM 118: N = 120, n_i = 80, ICC = 0.01, θ = 0.3
Method
QIF
GEE1
GzLMM
0.20
0.25
0.30
0.35
0.40
Estimates of θ

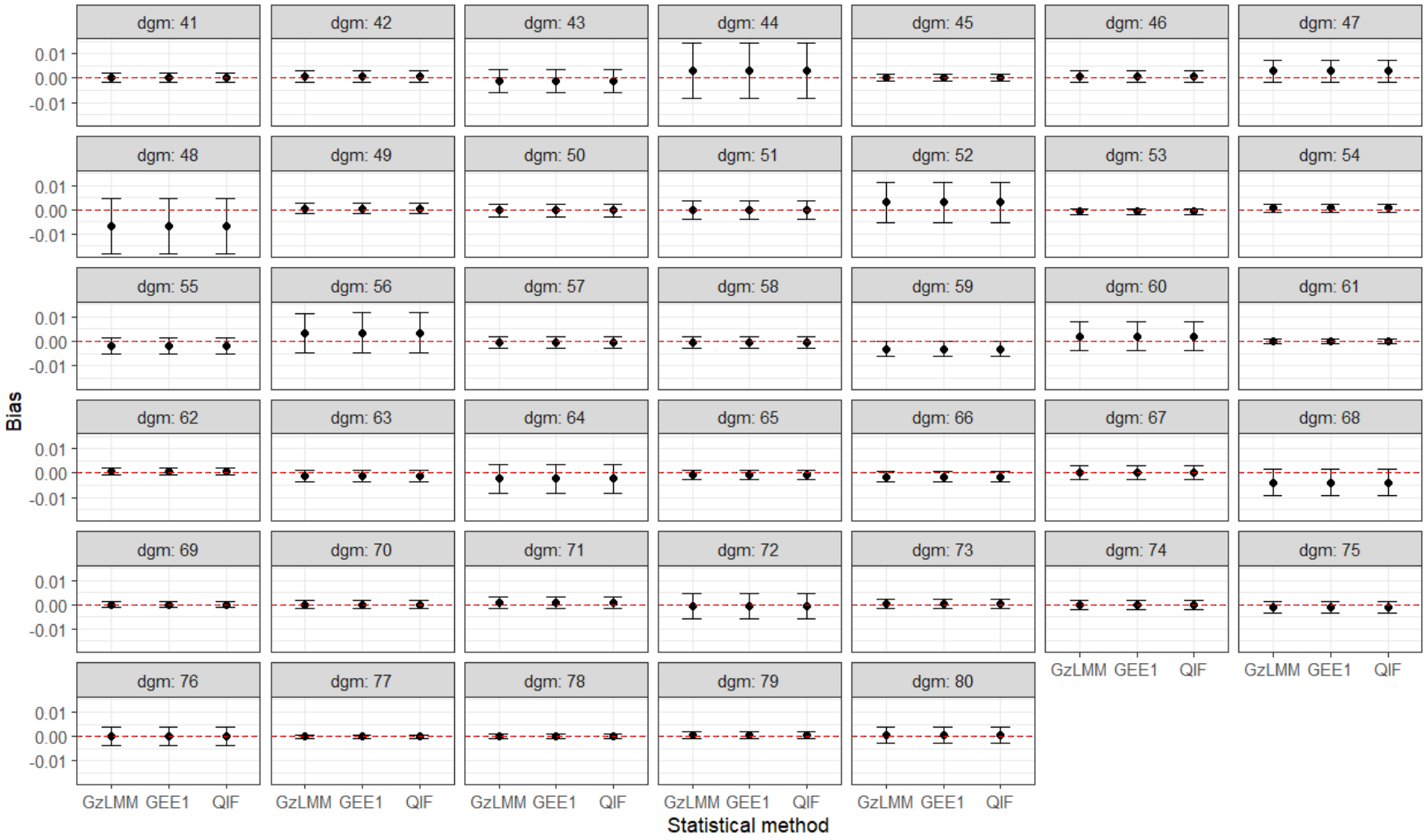

DGM 119: N = 120, n_i = 80, ICC = 0.05, θ = 0.3
Method
QIF
GEE1
GzLMM
0.2
0.3
0.4
Estimates of θ

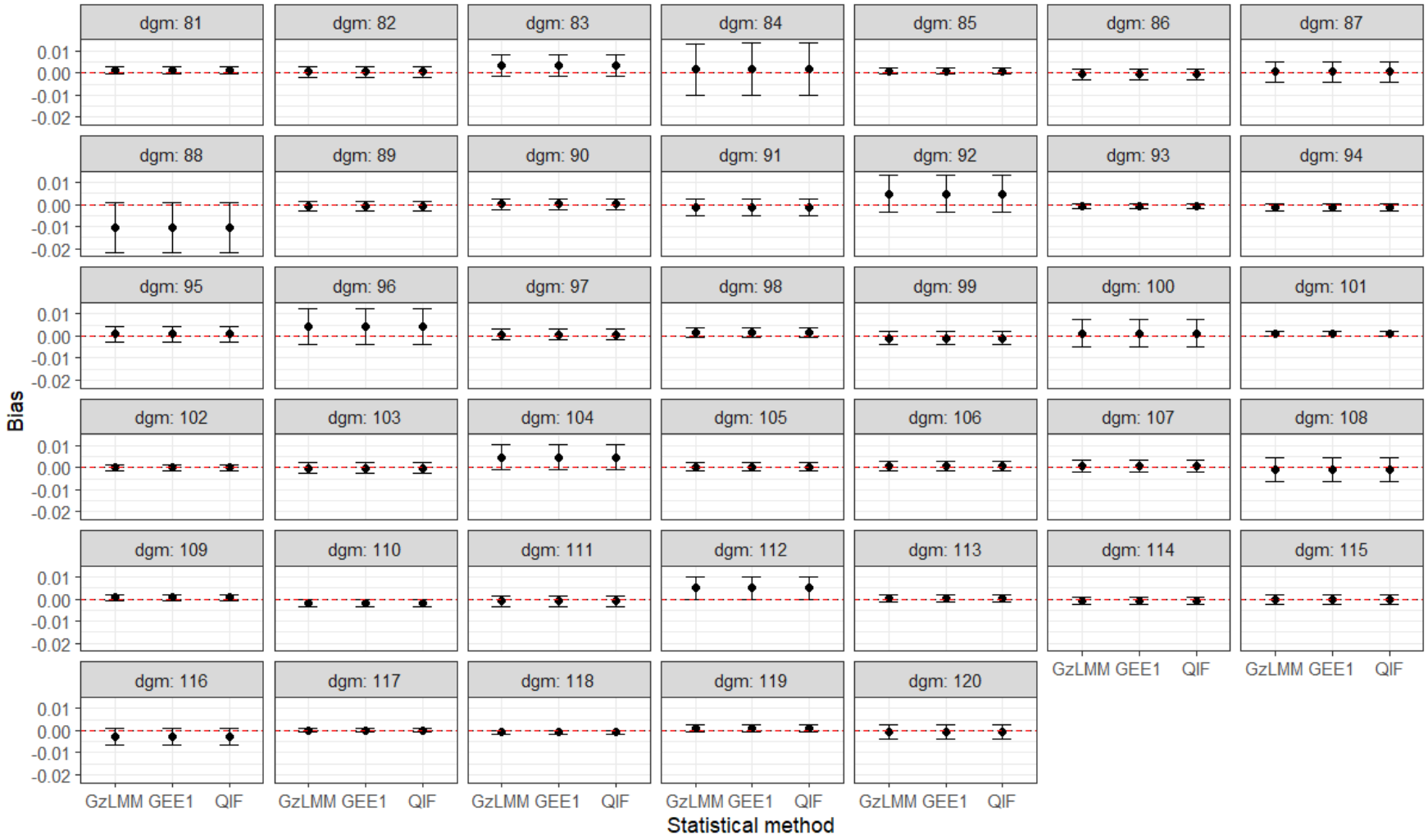

DGM 120: N = 120, n_i = 80, ICC = 0.25, θ = 0.3
Method
QIF
GEE1
GzLMM
0.0
0.2
0.4
0.6
Estimates of θ

Appendix 7

Box plot for each of the 120 scenarios with each containing 4000 estimated biases of the estimates of the intervention effect from GzLMM, GEE1 and QIF

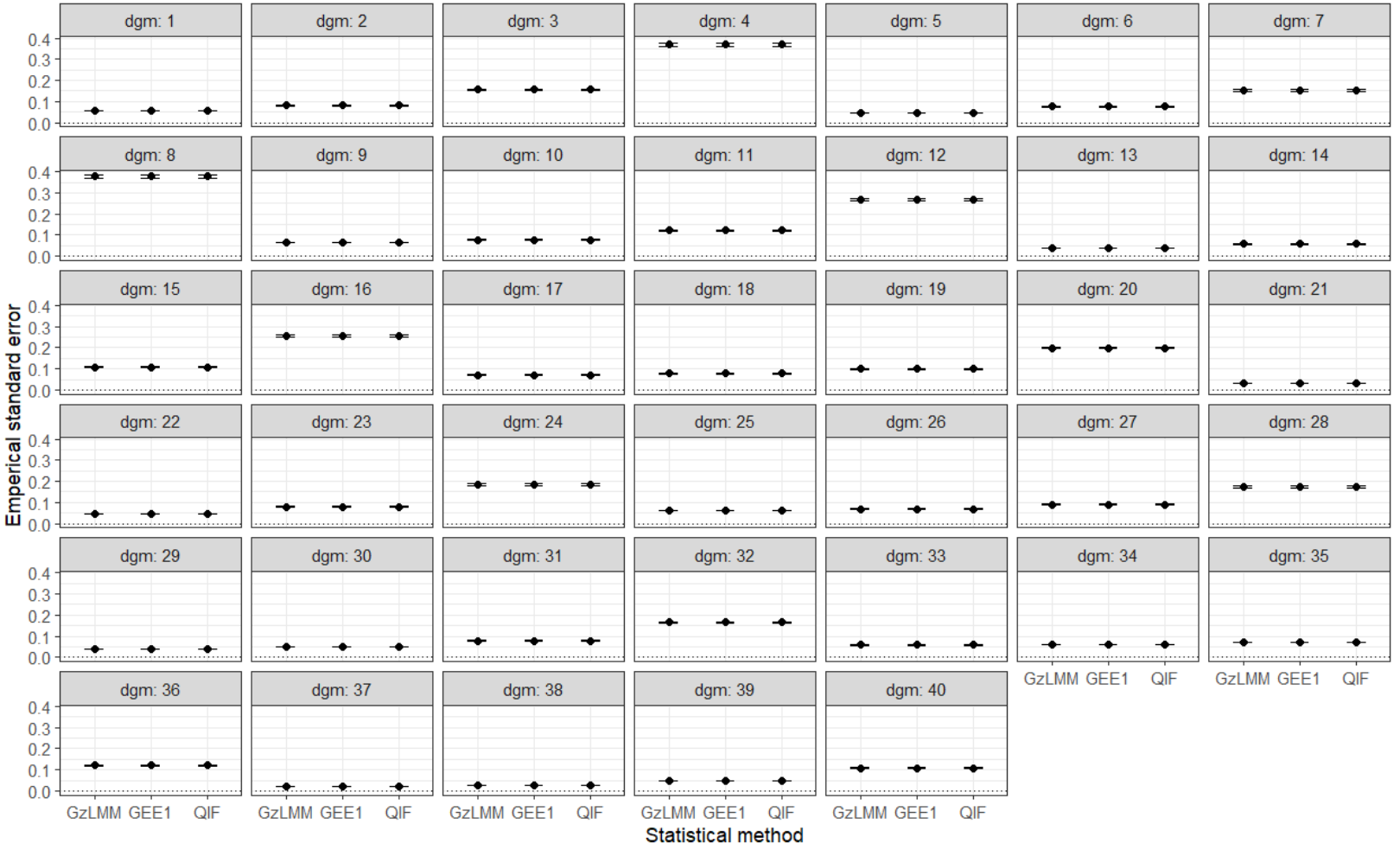


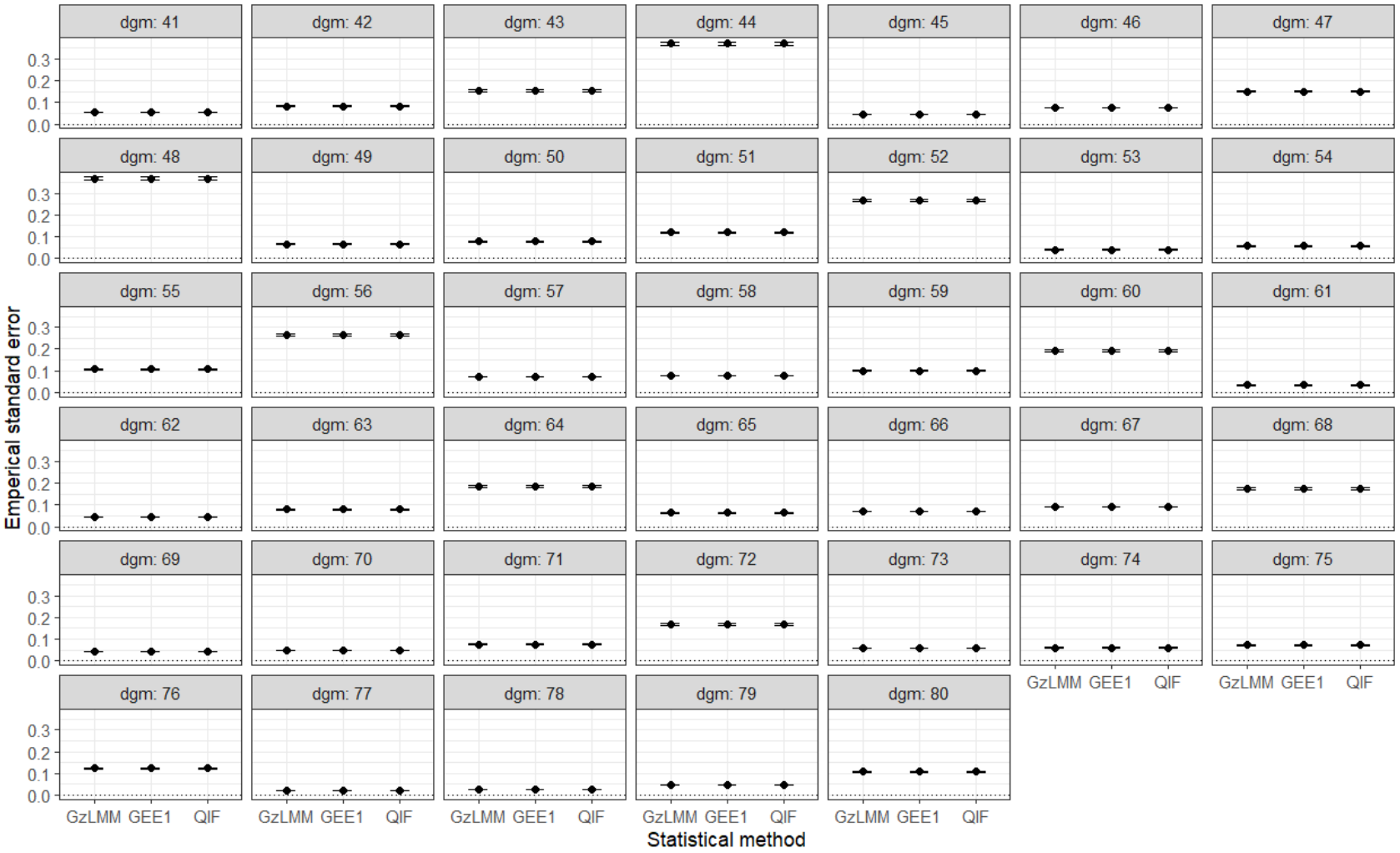

dgm: 41
dgm: 42
dgm: 43
dgm: 44
dgm: 45
dgm: 46
dgm: 47
dgm: 48
dgm: 49
dgm: 50
dgm: 51
dgm: 52
dgm: 53
dgm: 54
dgm: 55
dgm: 56
dgm: 57
dgm: 58
dgm: 59
dgm: 60
dgm: 61
dgm: 62
dgm: 63
dgm: 64
dgm: 65
dgm: 66
dgm: 67
dgm: 68
dgm: 69
dgm: 70
dgm: 71
dgm: 72
dgm: 73
dgm: 74
dgm: 75
dgm: 76
dgm: 77
dgm: 78
dgm: 79
dgm: 80
Bias
0.01
0.00
-0.01
GzLMM
GEE1
QIF
Statistical method

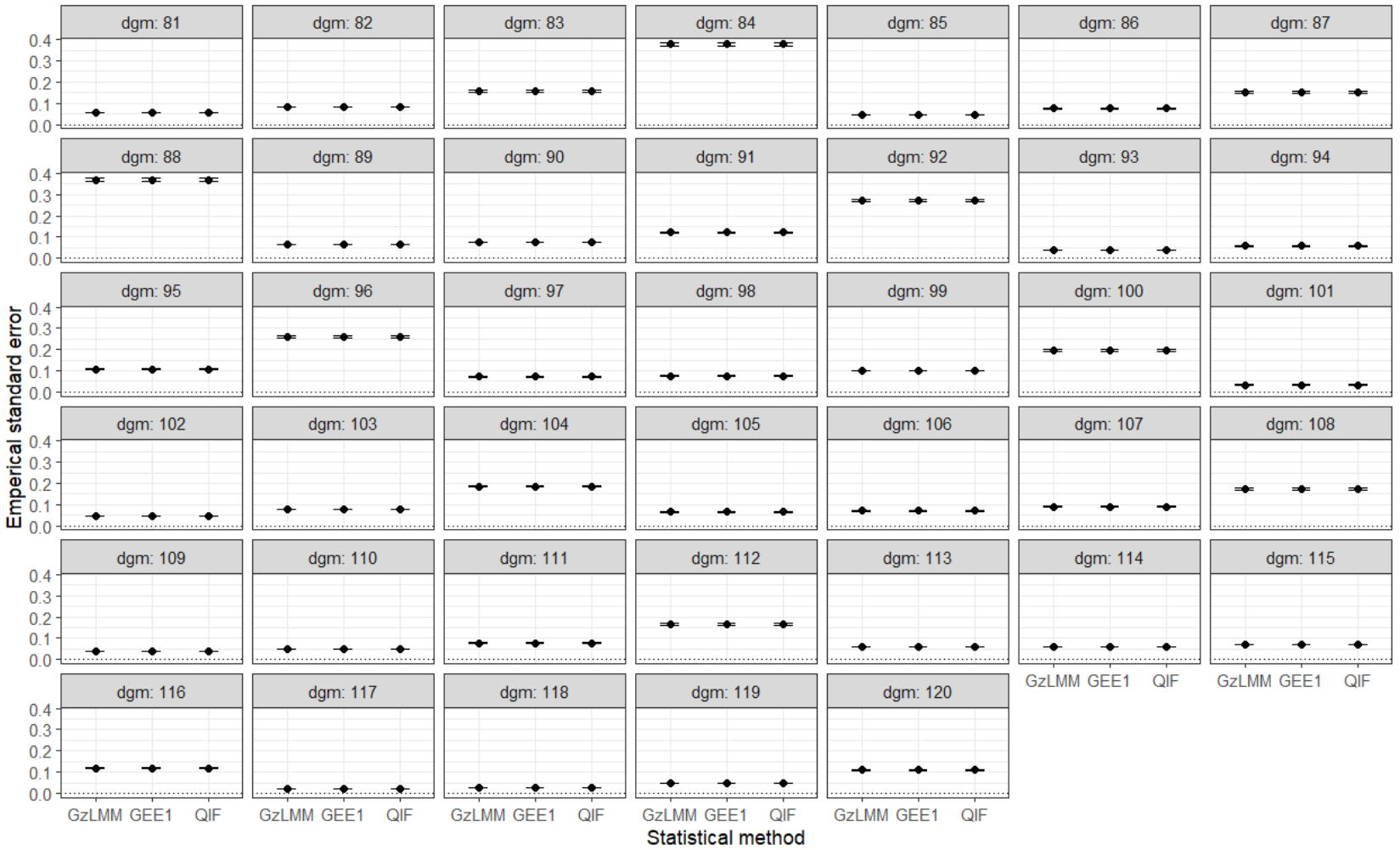

dgm: 81
dgm: 82
dgm: 83
dgm: 84
dgm: 85
dgm: 86
dgm: 87
dgm: 88
dgm: 89
dgm: 90
dgm: 91
dgm: 92
dgm: 93
dgm: 94
dgm: 95
dgm: 96
dgm: 97
dgm: 98
dgm: 99
dgm: 100
dgm: 101
dgm: 102
dgm: 103
dgm: 104
dgm: 105
dgm: 106
dgm: 107
dgm: 108
dgm: 109
dgm: 110
dgm: 111
dgm: 112
dgm: 113
dgm: 114
dgm: 115
dgm: 116
dgm: 117
dgm: 118
dgm: 119
dgm: 120
Bias
0.01
0.00
-0.01
-0.02
GzLMM GEE1 QIF
Statistical method

Appendix 6

Box plot for each of the 120 cRCT scenarios, based on 4000 empirical standard errors (ESEs) from the three methods, GzLMM, GEE1, and QIF

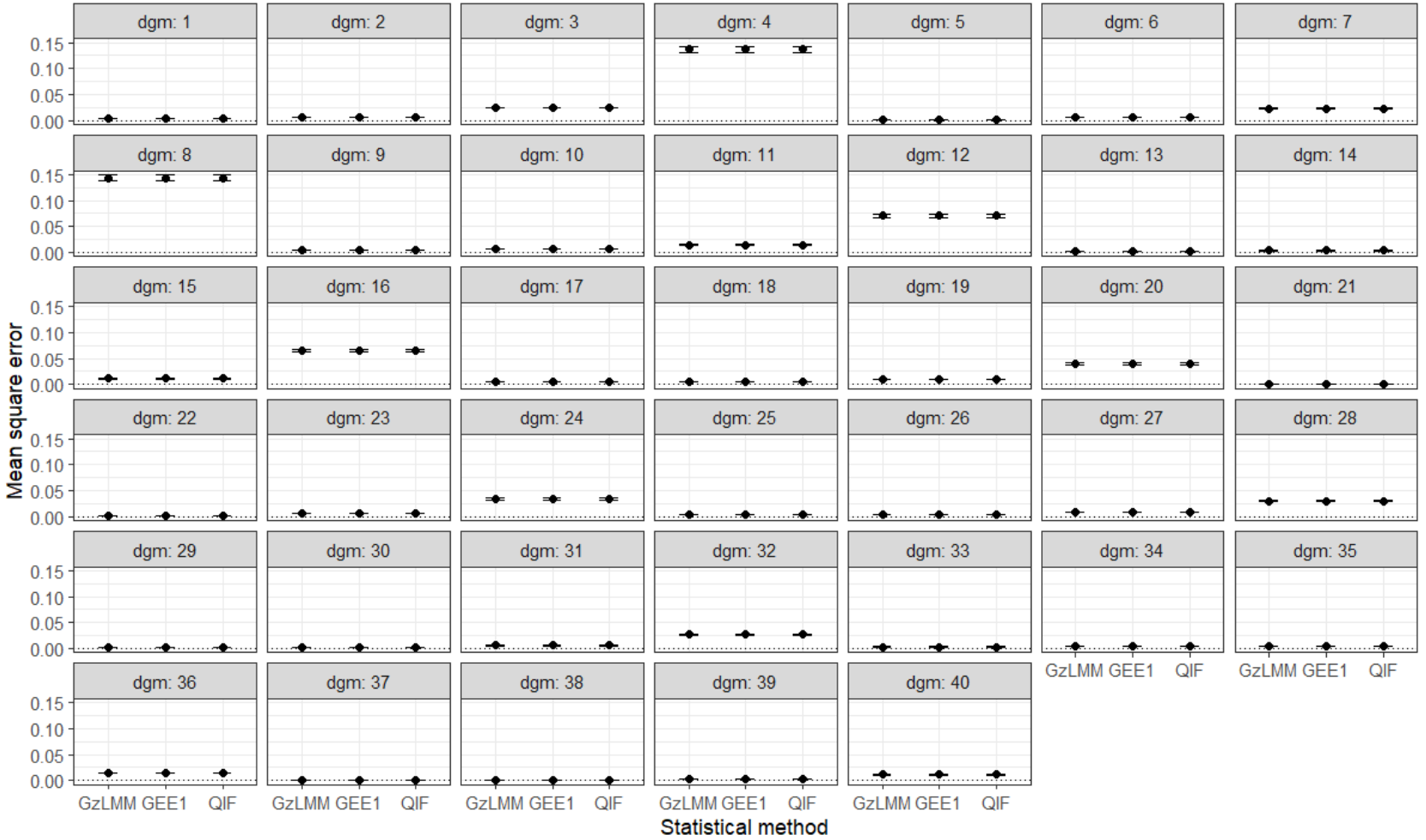


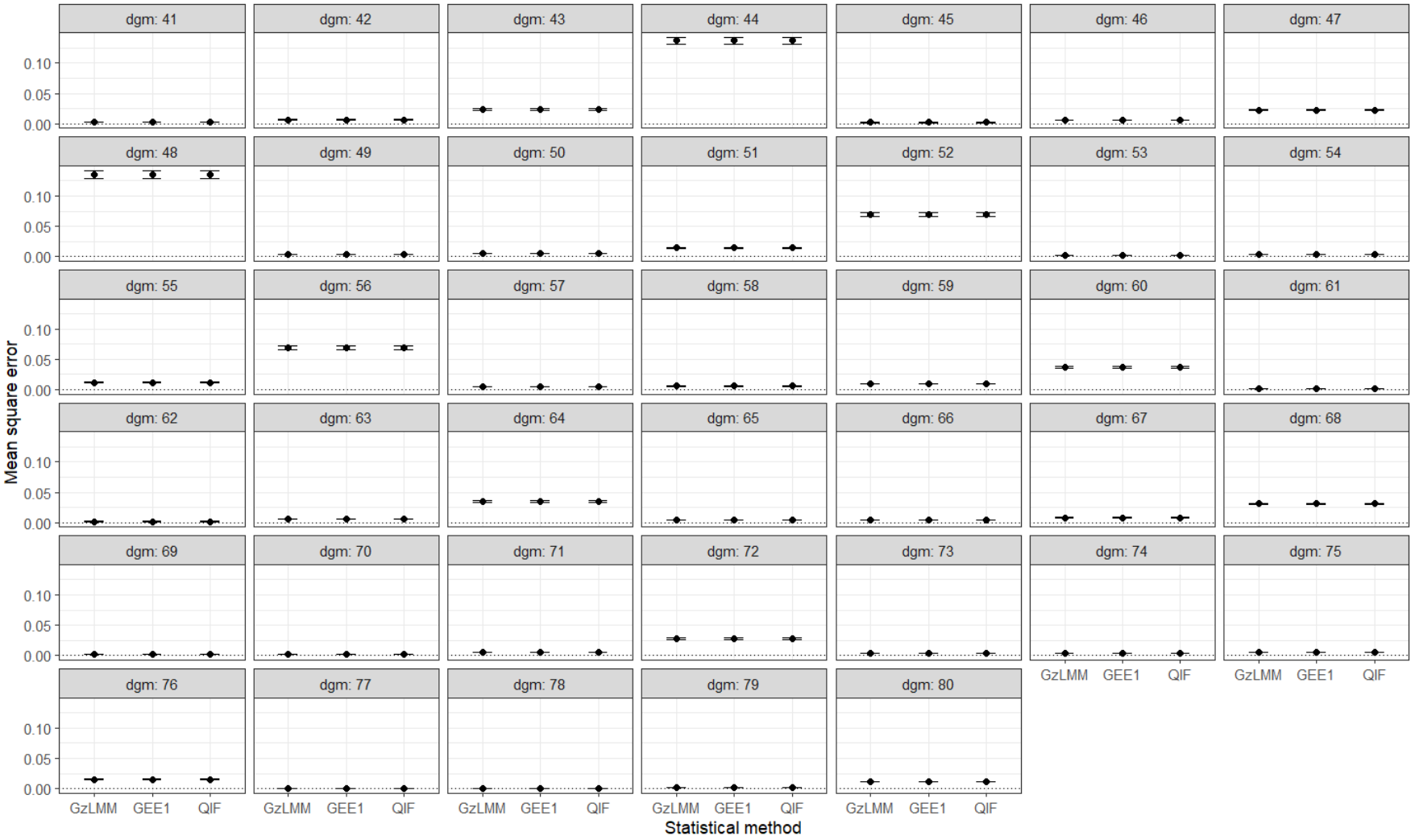

dgm: 41
dgm: 42
dgm: 43
dgm: 44
dgm: 45
dgm: 46
dgm: 47
dgm: 48
dgm: 49
dgm: 50
dgm: 51
dgm: 52
dgm: 53
dgm: 54
dgm: 55
dgm: 56
dgm: 57
dgm: 58
dgm: 59
dgm: 60
dgm: 61
dgm: 62
dgm: 63
dgm: 64
dgm: 65
dgm: 66
dgm: 67
dgm: 68
dgm: 69
dgm: 70
dgm: 71
dgm: 72
dgm: 73
dgm: 74
dgm: 75
dgm: 76
dgm: 77
dgm: 78
dgm: 79
dgm: 80
0.0
0.1
0.2
0.3
GzLMM
GEE1
QIF
Emperical standard error
Statistical method

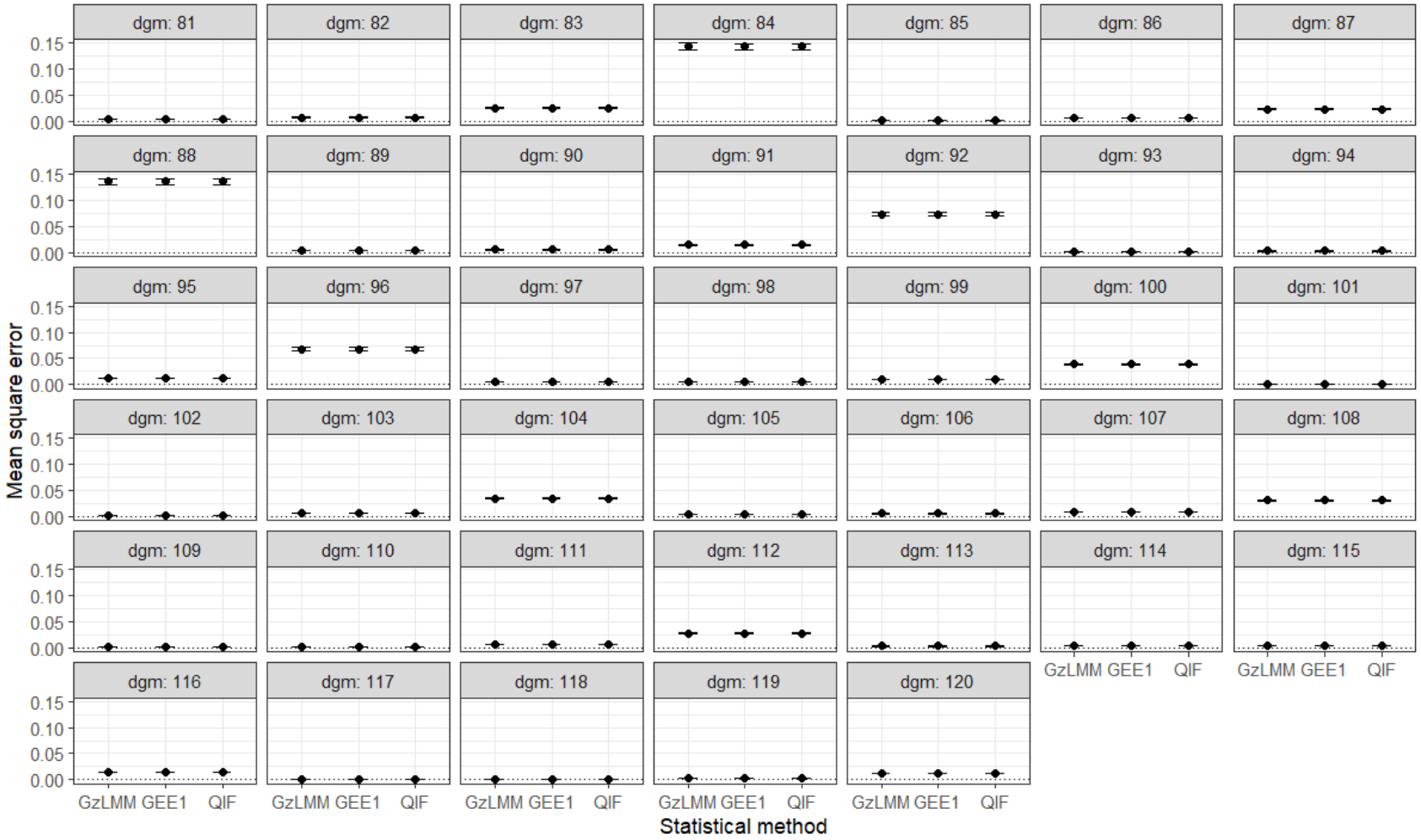

dgm: 81
dgm: 82
dgm: 83
dgm: 84
dgm: 85
dgm: 86
dgm: 87
dgm: 88
dgm: 89
dgm: 90
dgm: 91
dgm: 92
dgm: 93
dgm: 94
dgm: 95
dgm: 96
dgm: 97
dgm: 98
dgm: 99
dgm: 100
dgm: 101
dgm: 102
dgm: 103
dgm: 104
dgm: 105
dgm: 106
dgm: 107
dgm: 108
dgm: 109
dgm: 110
dgm: 111
dgm: 112
dgm: 113
dgm: 114
dgm: 115
dgm: 116
dgm: 117
dgm: 118
dgm: 119
dgm: 120
Emperical standard error
0.0
0.1
0.2
0.3
0.4
GzLMM
GEE1
QIF
Statistical method

Appendix 9

Box plot for each of the 120 cRCT scenarios, based on 4000 mean square error (MSE) from the three methods, GzLMM, GEE1, and QIF

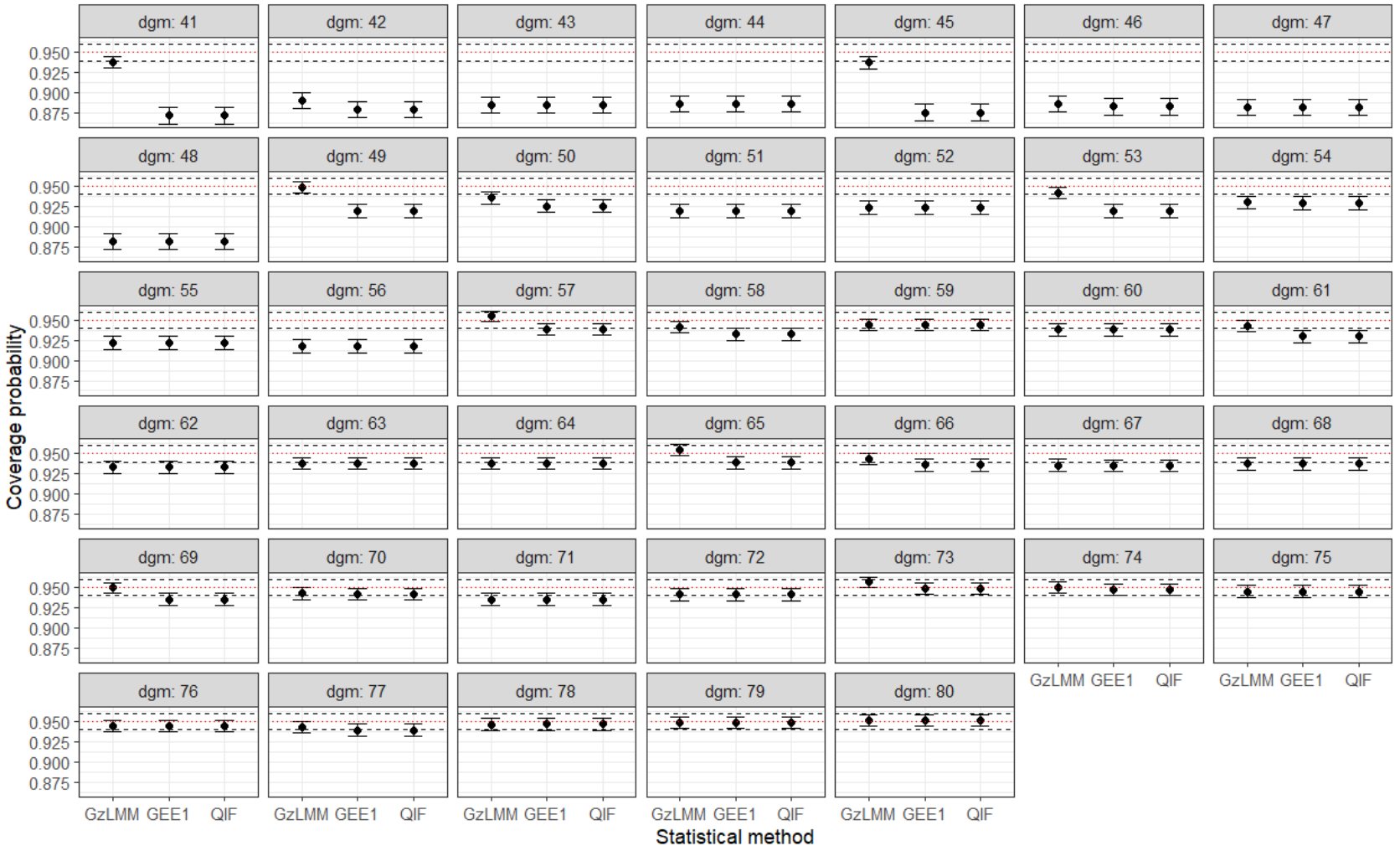



dgm: 41
dgm: 42
dgm: 43
dgm: 44
dgm: 45
dgm: 46
dgm: 47
dgm: 48
dgm: 49
dgm: 50
dgm: 51
dgm: 52
dgm: 53
dgm: 54
dgm: 55
dgm: 56
dgm: 57
dgm: 58
dgm: 59
dgm: 60
dgm: 61
dgm: 62
dgm: 63
dgm: 64
dgm: 65
dgm: 66
dgm: 67
dgm: 68
dgm: 69
dgm: 70
dgm: 71
dgm: 72
dgm: 73
dgm: 74
dgm: 75
dgm: 76
dgm: 77
dgm: 78
dgm: 79
dgm: 80
Mean square error
0.00
0.05
0.10
GzLMM
GEE1
QIF
Statistical method


dgm: 81
dgm: 82
dgm: 83
dgm: 84
dgm: 85
dgm: 86
dgm: 87
dgm: 88
dgm: 89
dgm: 90
dgm: 91
dgm: 92
dgm: 93
dgm: 94
dgm: 95
dgm: 96
dgm: 97
dgm: 98
dgm: 99
dgm: 100
dgm: 101
dgm: 102
dgm: 103
dgm: 104
dgm: 105
dgm: 106
dgm: 107
dgm: 108
dgm: 109
dgm: 110
dgm: 111
dgm: 112
dgm: 113
dgm: 114
dgm: 115
dgm: 116
dgm: 117
dgm: 118
dgm: 119
dgm: 120
0.15
0.10
0.05
0.00
Mean square error
GzLMM GEE1 QIF
Statistical method

Appendix 10

Forest plot for each of the 120 cRCT scenarios, based on 4000 coverage probabilities for $\widehat{\theta}$ from GzLMM, GEE1, and QIF


dgm: 41
dgm: 42
dgm: 43
dgm: 44
dgm: 45
dgm: 46
dgm: 47
dgm: 48
dgm: 49
dgm: 50
dgm: 51
dgm: 52
dgm: 53
dgm: 54
dgm: 55
dgm: 56
dgm: 57
dgm: 58
dgm: 59
dgm: 60
dgm: 61
dgm: 62
dgm: 63
dgm: 64
dgm: 65
dgm: 66
dgm: 67
dgm: 68
dgm: 69
dgm: 70
dgm: 71
dgm: 72
dgm: 73
dgm: 74
dgm: 75
dgm: 76
dgm: 77
dgm: 78
dgm: 79
dgm: 80
0.950
0.925
0.900
0.875
Coverage probability
GzLMM GEE1 QIF
Statistical method


dgm: 81
dgm: 82
dgm: 83
dgm: 84
dgm: 85
dgm: 86
dgm: 87
dgm: 88
dgm: 89
dgm: 90
dgm: 91
dgm: 92
dgm: 93
dgm: 94
dgm: 95
dgm: 96
dgm: 97
dgm: 98
dgm: 99
dgm: 100
dgm: 101
dgm: 102
dgm: 103
dgm: 104
dgm: 105
dgm: 106
dgm: 107
dgm: 108
dgm: 109
dgm: 110
dgm: 111
dgm: 112
dgm: 113
dgm: 114
dgm: 115
dgm: 116
dgm: 117
dgm: 118
dgm: 119
dgm: 120
0.96
0.94
0.92
0.90
0.88
Coverage probability
GzLMM
GEE1
QIF
Statistical method

Appendix 11

Coverage probabilities and powers for θ =0.3 from GzLMM, GEE1 and QIF

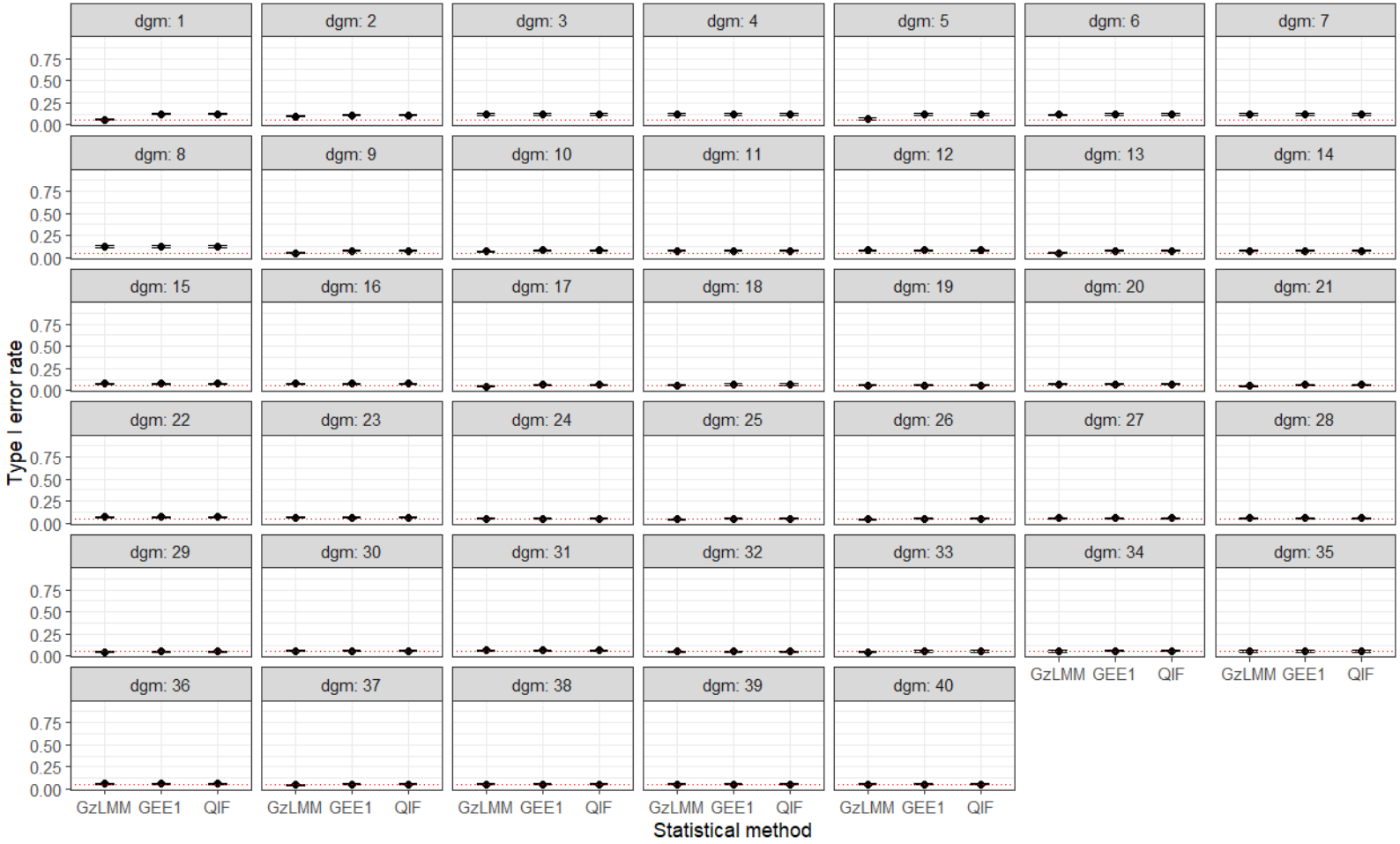
Table S9.1 Empirical coverage probability and power for each method for several cRCT scenarios specified by the combinations of N, n_i, ICC, and $\theta = 0.3$

Parameters			$\theta = 0.3$ Coverage probability (MCSE)			$\theta = 0.3$ Power (MCSE)		
N	n_i	ICC	GzLMM	GEE1	QIF	GzLMM	GEE1	QIF
10	150	0.001	0.94 (0.0037)	0.88 (0.0051)	0.88 (0.0051)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.01	0.89 (0.0050)	0.88 (0.0052)	0.88 (0.0052)	0.97 (0.0029)	0.97 (0.0029)	0.97 (0.0029)
		0.05	0.88 (0.0052)	0.88 (0.0052)	0.88 (0.0052)	0.60 (0.0078)	0.60 (0.0078)	0.60 (0.0078)
		0.25	0.88 (0.0052)	0.88 (0.0052)	0.88 (0.0052)	0.23 (0.0066)	0.23 (0.0066)	0.23 (0.0066)
	250	0.001	0.94 (0.0038)	0.88 (0.0051)	0.88 (0.0051)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.01	0.88 (0.0051)	0.88 (0.0052)	0.88 (0.0052)	0.98 (0.0021)	0.98 (0.0021)	0.98 (0.0021)
		0.05	0.88 (0.0051)	0.88 (0.0051)	0.88 (0.0051)	0.62 (0.0077)	0.62 (0.0077)	0.62 (0.0077)
		0.25	0.89 (0.0050)	0.89 (0.0050)	0.89 (0.0050)	0.22 (0.0065)	0.22 (0.0065)	0.22 (0.0065)
20	50	0.001	0.96 (0.0033)	0.93 (0.0041)	0.93 (0.0041)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.01	0.93 (0.0040)	0.92 (0.0043)	0.92 (0.0043)	0.98 (0.0025)	0.98 (0.0025)	0.98 (0.0025)
		0.05	0.92 (0.0044)	0.92 (0.0044)	0.92 (0.0044)	0.73 (0.0070)	0.73 (0.0070)	0.73 (0.0070)
		0.25	0.91 (0.0045)	0.91 (0.0045)	0.91 (0.0045)	0.26 (0.0070)	0.26 (0.0070)	0.26 (0.0070)
	150	0.001	0.95 (0.0036)	0.92 (0.0043)	0.92 (0.0043)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.01	0.92 (0.0043)	0.92 (0.0043)	0.92 (0.0043)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.05	0.92 (0.0042)	0.92 (0.0042)	0.92 (0.0042)	0.81 (0.0062)	0.81 (0.0062)	0.81 (0.0062)
		0.25	0.92 (0.0043)	0.92 (0.0043)	0.92 (0.0043)	0.26 (0.0069)	0.26 (0.0069)	0.26 (0.0069)
40	20	0.001	0.95 (0.0035)	0.93 (0.0040)	0.93 (0.0040)	0.99 (0.0018)	0.99 (0.0018)	0.99 (0.0018)
		0.01	0.95 (0.0035)	0.94 (0.0038)	0.94 (0.0038)	0.98 (0.0024)	0.98 (0.0024)	0.98 (0.0024)
		0.05	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0036)	0.84 (0.0058)	0.84 (0.0058)	0.84 (0.0058)
		0.25	0.93 (0.0040)	0.93 (0.0040)	0.93 (0.0040)	0.36 (0.0076)	0.36 (0.0076)	0.36 (0.0076)
	100	0.001	0.95 (0.0040)	0.93 (0.0040)	0.93 (0.0040)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.01	0.94 (0.0039)	0.94 (0.0039)	0.94 (0.0039)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.05	0.94 (0.0039)	0.94 (0.0039)	0.94 (0.0039)	0.97 (0.0027)	0.97 (0.0027)	0.97 (0.0027)
		0.25	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)	0.40 (0.0078)	0.40 (0.0078)	0.40 (0.0078)
50	20	0.001	0.95 (0.0034)	0.94 (0.0040)	0.94 (0.0040)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.01	0.95 (0.0035)	0.94 (0.0037)	0.94 (0.0037)	0.99 (0.0016)	0.99 (0.0016)	0.99 (0.0016)
		0.05	0.95 (0.0036)	0.95 (0.0036)	0.95 (0.0036)	0.92 (0.0043)	0.92 (0.0043)	0.92 (0.0043)
		0.25	0.94 (0.0038)	0.94 (0.0038)	0.94 (0.0038)	0.41 (0.0078)	0.41 (0.0078)	0.41 (0.0078)
	50	0.001	0.96 (0.0032)	0.95 (0.0035)	0.95 (0.0035)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.01	0.94 (0.0037)	0.94 (0.0038)	0.94 (0.0038)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.05	0.94 (0.0039)	0.94 (0.0039)	0.94 (0.0039)	0.97 (0.0026)	0.97 (0.0026)	0.97 (0.0026)
		0.25	0.94 (0.0036)	0.94 (0.0036)	0.94 (0.0036)	0.47 (0.0079)	0.47 (0.0079)	0.47 (0.0079)
120	10	0.001	0.95 (0.0034)	0.95 (0.0036)	0.95 (0.0036)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.01	0.95 (0.0034)	0.95 (0.0035)	0.95 (0.0035)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.05	0.95 (0.0035)	0.95 (0.0035)	0.95 (0.0035)	0.99 (0.0017)	0.99 (0.0017)	0.99 (0.0017)
		0.25	0.95 (0.0034)	0.95 (0.0034)	0.95 (0.0034)	0.71 (0.0072)	0.71 (0.0072)	0.71 (0.0072)
	80	0.001	0.95 (0.0034)	0.95 (0.0036)	0.95 (0.0036)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.01	0.95 (0.0034)	0.95 (0.0034)	0.95 (0.0034)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.05	0.94 (0.0037)	0.94 (0.0037)	0.94 (0.0037)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)
		0.25	0.94 (0.0037)	0.94 (0.0037)	0.94 (0.0037)	0.79 (0.0064)	0.79 (0.0064)	0.79 (0.0064)

Note: N is the number of clusters; n_i is the i^{th} cluster size; ICC is the intracluster correlation coefficient; θ is the true treatment effect; MCSE is the Monte Carlo standard error. Values shaded blue are equal or within the range of the expected nominal value (2 decimal places), while the ones shaded orange are greater than the expected nominal value (2 decimal places), and the ones shaded green are less than the expected nominal value. Note each cell in the table is based on 4,000 simulated datasets, except cells corresponding to GzLMM where some simulations failed to converge. The electronic version is in colour.

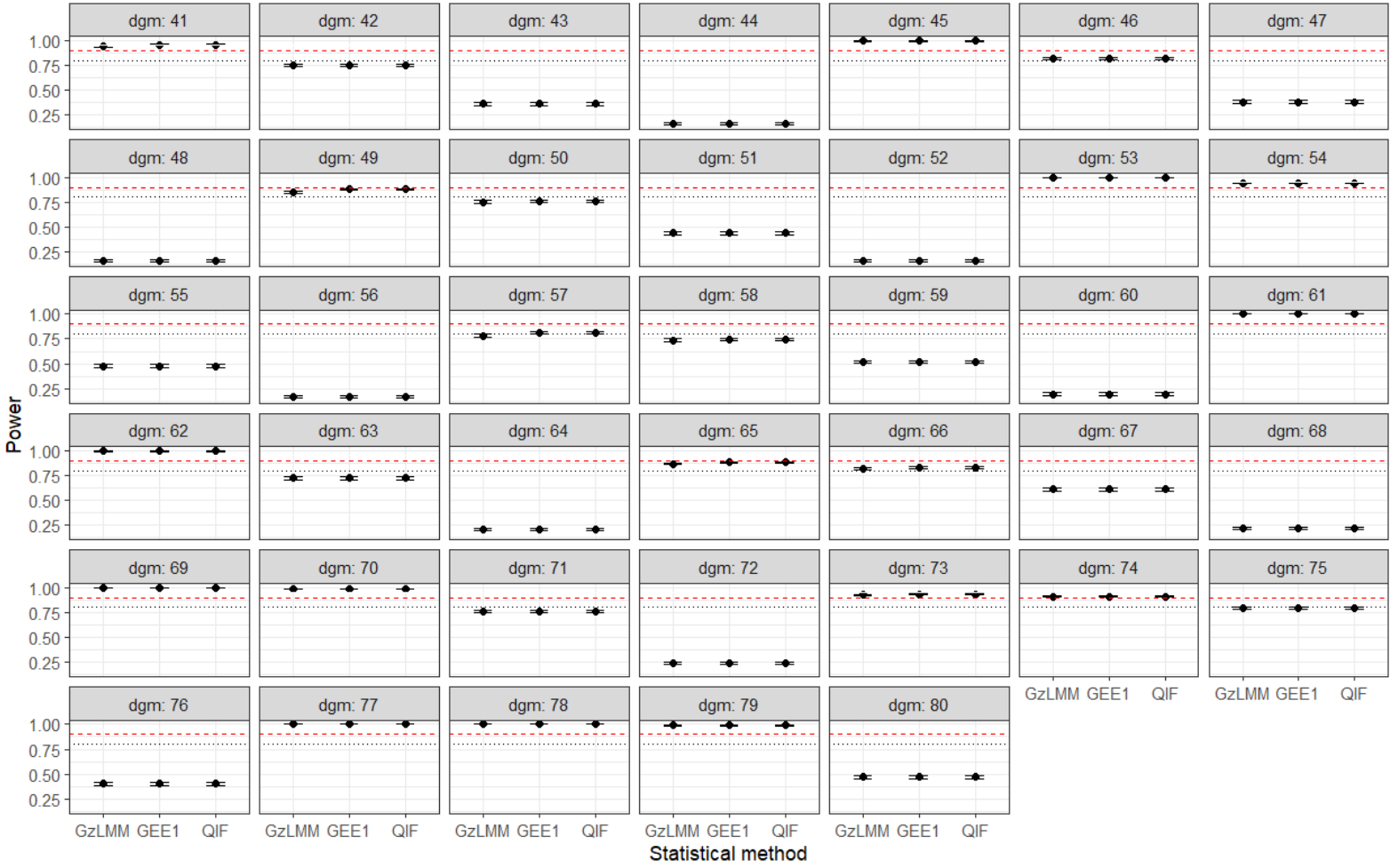
Appendix 12

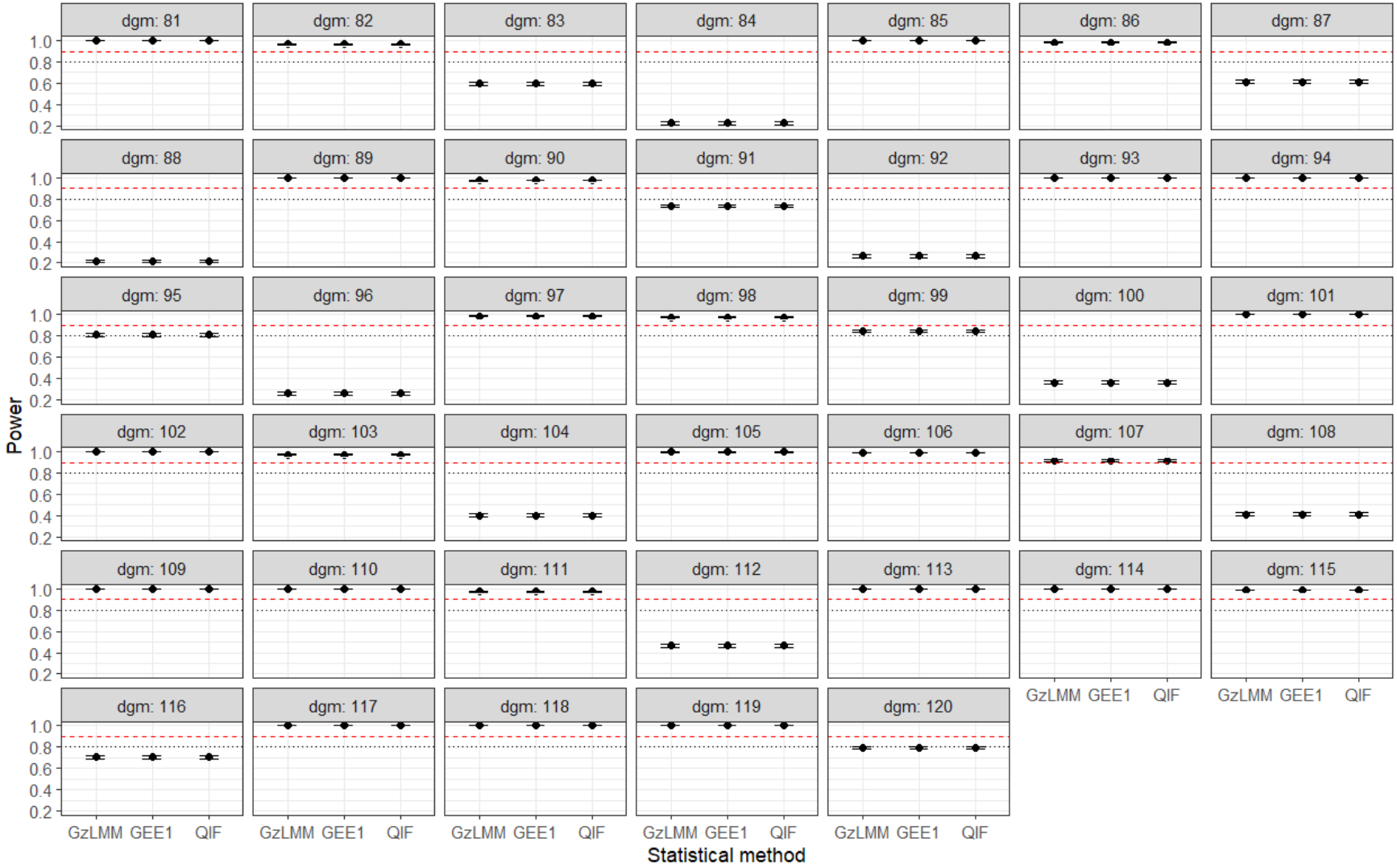
Forest plots of Type I error rates for each of the 40 scenarios based on 4000 estimates from GzLMM, GEE1 and QIF when $\theta = 0$



Appendix 73

forest plot for each of the 80 scenarios based on 4000 estimates of power from GzLMM, GEE1 and QIF under the alternative hypothesis (i.e., $\theta = 0.2\ or\ 0.3$)




dgm: 81
dgm: 82
dgm: 83
dgm: 84
dgm: 85
dgm: 86
dgm: 87
dgm: 88
dgm: 89
dgm: 90
dgm: 91
dgm: 92
dgm: 93
dgm: 94
dgm: 95
dgm: 96
dgm: 97
dgm: 98
dgm: 99
dgm: 100
dgm: 101
dgm: 102
dgm: 103
dgm: 104
dgm: 105
dgm: 106
dgm: 107
dgm: 108
dgm: 109
dgm: 110
dgm: 111
dgm: 112
dgm: 113
dgm: 114
dgm: 115
dgm: 116
dgm: 117
dgm: 118
dgm: 119
dgm: 120
Power
1.0
0.8
0.6
0.4
0.2
GzLMM GEE1 QIF
Statistical method

References

Adams, G. *et al.* (2004) 'Patterns of intra-cluster correlation from primary care research to inform study design and analysis', *Journal of Clinical Epidemiology*, 57(8), pp. 785–794. Available at: https://doi.org/10.1016/j.jclinepi.2003.12.013.

Agbla, S.C. and DiazOrdaz, K. (2018) 'Reporting non-adherence in cluster randomised trials: A systematic review', *Clinical Trials*, 15(3), pp. 294–304. Available at: https://doi.org/10.1177/1740774518761666.

Agresti, A., Caffo, B. and Ohman-Strickland, P. (2004) 'Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies', *Computational Statistics and Data Analysis*, 47(3), pp. 639–653. Available at: https://doi.org/10.1016/j.csda.2003.12.009.

Allan, Donner ; Neil, K. (2001) 'Current and future challenges in the design and analysis of cluster randomization trials', *Statistics in Medicine*, 20(24), pp. 3729–3740. Available at: https://doi.org/10.1002/sim.1115.

Arksey, H. and O'Malley, L. (2005) 'Scoping studies: Towards a methodological framework', *International Journal of Social Research Methodology: Theory and Practice*, 8(1), pp. 19–32. Available at: https://doi.org/10.1080/1364557032000119616.

Arnup, S.J. *et al.* (2016) 'Appropriate statistical methods were infrequently used in cluster-randomized crossover trials', *Journal of Clinical Epidemiology*. Elsevier USA, pp. 40–50. Available at: https://doi.org/10.1016/j.jclinepi.2015.11.013.

Austin, P.C. (2007) 'A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes', *Statistics in Medicine*, 26(19), pp. 3550–3565. Available at: https://doi.org/10.1002/sim.2813.

Austin, P.C. (2010) 'A comparison of the statistical power of different methods for the analysis of repeated cross-sectional cluster randomization trials with binary outcomes', *International Journal of Biostatistics*, 6(1). Available at: https://doi.org/10.2202/1557-4679.1179.

Austin, P.C. *et al.* (2021) 'Missing Data in Clinical Research: A Tutorial on Multiple Imputation', *Canadian Journal of Cardiology*. Elsevier Inc., pp. 1322–1331. Available at: https://doi.org/10.1016/j.cjca.2020.11.010.

Balemi, A. and Lee, A. (2009) 'Comparison of GEE1 and GEE2 estimation applied to clustered logistic regression', *Journal of Statistical Computation and Simulation*, 79(4), pp. 361–378. Available at: https://doi.org/10.1080/00949650701786085.

Ballard, C. *et al.* (2020) 'Improving mental health and reducing antipsychotic use in people with dementia in care homes: the WHELD research programme including two RCTs', *Programme Grants Appl Res*, 8(6). Available at: https://doi.org/10.3310/pgfar08060.

Balzer, L.B. *et al.* (2016) 'Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching', *Statistics in Medicine*, 35(21), pp. 3717–3732. Available at: https://doi.org/10.1002/sim.6965.

Balzer, L.B., Petersen, M.L. and J, van der L.M. (2016) 'Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching', *Statistics in Medicine*, 35(21), pp. 3717–3732. Available at: https://doi.org/http://dx.doi.org/10.1002/sim.6965.

Barker, D. *et al.* (2017) 'Minimum number of clusters and comparison of analysis methods for cross sectional stepped wedge cluster randomised trials with binary outcomes: A simulation study', *Trials*, 18(1). Available at: https://doi.org/10.1186/s13063-017-1862-2.

Barnhart, H.X. and Williamson, J.M. (1998) *Goodness-of-Fit Tests for GEE Modeling with Binary Responses*. Available at: https://about.jstor.org/terms.

Begg, C. *et al.* (1996) 'Improving the quality of reporting of randomized controlled trials: The CONSORT statement', *Journal of the American Medical Association*, 276(8), pp. 637–639. Available at: https://doi.org/10.1001/jama.276.8.637.

Bellamy, S.L. *et al.* (2000) 'Analysis of dichotomous outcome data for community intervention studies', *Statistical Methods in Medical Research*, 9(2), pp. 135–159. Available at: https://doi.org/10.1191/096228000672549488.

Bland, J.M. (2004) 'Cluster randomised trials in the medical literature: Two bibliometric surveys', *BMC Medical Research Methodology*, 4, pp. 2–7. Available at: https://doi.org/10.1186/1471-2288-4-21.

Borhan, S. *et al.* (2019) 'Sensitivity of methods for analyzing continuous outcome from stratified cluster randomized trials – an empirical comparison study', *Contemporary Clinical Trials Communications*, 15. Available at: https://doi.org/10.1016/j.conctc.2019.100405.

Borhan, S. *et al.* (2020) 'An empirical comparison of methods for analyzing over-dispersed zero-inflated count data from stratified cluster randomized trials', *Contemporary Clinical Trials Communications*, 17. Available at: https://doi.org/10.1016/j.conctc.2020.100539.

Bossoli, D. and Bottai, M. (2018) 'Marginal quantile regression for dependent data with a working odds-ratio matrix', *Biostatistics*, 19(4), pp. 529–545. Available at: https://doi.org/10.1093/biostatistics/kxx052.

Brown, R.L. (2013) 'Modeling impure clusters in a cluster randomized controlled trial', *Research in Nursing and Health*, 36(2), pp. 216–223. Available at: https://doi.org/10.1002/nur.21523.

Burton, A. *et al.* (2006) 'The design of simulation studies in medical statistics', *Statistics in Medicine*, 25(24), pp. 4279–4292. Available at: https://doi.org/10.1002/sim.2673.

Button, K.S. *et al.* (2013) 'Power failure: Why small sample size undermines the reliability of neuroscience', *Nature Reviews Neuroscience*, 14(5), pp. 365–376. Available at: https://doi.org/10.1038/nrn3475.

Cai, J. and Kim, J. (2003) 'Nonparametric quantile estimation with correlated failure time data', *Lifetime Data Analysis*, 9(4), pp. 357–371. Available at: https://doi.org/10.1023/B:LIDA.0000012422.30514.c7.

Caille, A., Leyrat, C. and Giraudeau, B. (2016) 'A comparison of imputation strategies in cluster randomized trials with missing binary outcomes', *Statistical Methods in Medical Research*, 25(6), pp. 2650–2669. Available at: https://doi.org/10.1177/0962280214530030.

Cameron, S.T. *et al.* (2020) 'Use of effective contraception following provision of the progestogen-only pill for women presenting to community pharmacies for emergency contraception (Bridge-It): a pragmatic cluster-randomised crossover trial', *The Lancet*, 396(10262), pp. 1585–1594. Available at: https://doi.org/10.1016/S0140-6736(20)31785-2.

Cameron, S.T. *et al.* (2021) 'Provision of the progestogen-only pill by community pharmacies as bridging contraception for women receiving emergency contraception: the Bridge-it RCT', *Health technology assessment (Winchester, England)*, 25(27), pp. 1–92. Available at: https://doi.org/10.3310/hta25270.

Campbell, M. and Walters, S. (2014) *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research*. New York, UNITED KINGDOM: John Wiley & Sons, Incorporated. Available at: http://ebookcentral.proquest.com/lib/sheffield/detail.action?docID=1662762.

Campbell, M.J. (2014) 'Challenges of cluster randomized trials', *Journal of Comparative Effectiveness Research*, 3(3), pp. 271–281. Available at: https://doi.org/10.2217/cer.14.21.

Campbell, Michael J. and Walters, S.J. (2014) 'Regression methods of analysis for binary, count and time-to-event outcomes for a cluster randomised controlled trial', in *How to design, analyse and report cluster randomised trials in medicine and health related research*. John Wiley & Sons, pp. 126–142.

Campbell, Michael J; and Walters, S.J. (2014) 'Regression methods of analysis for binary, count and time-to-event outcomes for a cluster randomised controlled trial', in *How to design, analyse and report cluster randomised trials in medicine and health related research*. John Wiley & Sons, pp. 126–142.

Campbell, M.K. *et al.* (2012) 'Consort 2010 statement: Extension to cluster randomised trials', *BMJ (Online)*, 345(7881), pp. 1–21. Available at: https://doi.org/10.1136/bmj.e5661.

Campbell, M.K., Elbourne, D.R. and Altman, D.G. (2004) 'CONSORT statement: extension to cluster randomised trials', *BMJ*, 328(7441), pp. 702-LP-708. Available at: https://doi.org/10.1136/bmj.328.7441.702.

Campbell, M.K., Grimshaw, J.M. and Elbourne, D.R. (2004) 'Intracluster correlation coefficients in cluster randomized trials : empirical insights into how should they be reported', *BMC Medical Research Methodology*, 5, pp. 1–5.

Carey, V., Zeger, S. and Diggle, P. (1993) 'Modelling Multivariate Binary Data with Alternating Logistic Regressions Author (s): Vincent Carey , Scott L . Zeger and Peter Diggle Published by : Oxford University Press on behalf of Biometrika Trust Stable URL : https://www.jstor.org/stable/2337173', *Biometrika*, 80(3), pp. 517–526.

Charvat, H. *et al.* (2016) 'A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates', *Statistics in Medicine*, 35(18), pp. 3066–3084. Available at: https://doi.org/10.1002/sim.6881.

Chatterjee, S. and Bandyopadhyay, U. (2017) 'Three-period, two-treatment crossover design under long-term carryover effect', *Statistica Neerlandica*, 71(4), pp. 263–285. Available at: https://doi.org/https://doi.org/10.1111/stan.12110.

Chatterjee, S. and Bandyopadhyay, U. (2019) 'Nonparametric approaches for comparing three-period, two-treatment, four-sequence crossover designs', *Journal of Statistical Computation and Simulation*, 89(7), pp. 1153–1182. Available at: https://doi.org/10.1080/00949655.2019.1575381.

Chen, B.E. and Wang, J. (2019) 'Joint modeling of binary response and survival for clustered data in clinical trials', *Statistics in Medicine*, (August), pp. 1–14. Available at: https://doi.org/10.1002/sim.8403.

Chen, B.E. and Wang, J. (2020) 'Joint modeling of binary response and survival for clustered data in clinical trials', *Statistics in Medicine*, 39(3), pp. 326–339. Available at: https://doi.org/10.1002/sim.8403.

Chen, C.M. and Yu, C.Y. (2012) 'A two-stage estimation in the Clayton-Oakes model with marginal linear transformation models for multivariate failure time data', *Lifetime Data Analysis*, 18(1), pp. 94–115. Available at: https://doi.org/10.1007/s10985-011-9205-1.

Chondros, P. *et al.* (2021) 'When should matching be used in the design of cluster randomized trials?', *Statistics in Medicine*, 40(26), pp. 5765–5778. Available at: https://doi.org/10.1002/sim.9152.

Christian, M.S., Evans, C. EL and Cade, J.E. (2014) 'Does the Royal Horticultural Society Campaign for School Gardening increase intake of fruit and vegetables in children? Results from two randomised controlled trials', *Public Health Research*, 2(4), pp. 1–162. Available at: https://doi.org/10.3310/phr02040.

Christian, N.J., Ha, I. Do and Jeong, J.H. (2016) 'Hierarchical likelihood inference on clustered competing risks data', *Statistics in Medicine*, 35(2), pp. 251–267. Available at: https://doi.org/10.1002/sim.6628.

Christie, J., O'Halloran, P. and Stevenson, M. (2009) 'Planning a cluster randomized controlled trial', *Nursing Research*, 58(2), pp. 128–134. Available at: https://doi.org/10.1097/NNR.0b013e3181900cb5.

Chuang, J.H., Hripcsak, G. and Jenders, R.A. (2000) 'Considering clustering: a methodological review of clinical decision support system studies.', *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp. 146–150.

Clark, A.B. *et al.* (2010) 'Bayesian methods of analysis for cluster randomized trials with count outcome data', *Statistics in Medicine*, 29(2), pp. 199–209. Available at: https://doi.org/10.1002/sim.3747.

Coens, C. *et al.* (2020) 'International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium', *The Lancet Oncology*, 21(2), pp. e83–e96. Available at: https://doi.org/10.1016/S1470-2045(19)30790-9.

Connolly, P. *et al.* (2018) 'A cluster randomised controlled trial and evaluation and cost-effectiveness analysis of the Roots of Empathy schools-based programme for improving social and emotional well-being outcomes among 8- to 9-year-olds in Northern Ireland', *Public Health Research*, 6(4), pp. 1–108. Available at: https://doi.org/10.3310/phr06040.

Corder, K.L. *et al.* (2021) 'A school-based, peer-led programme to increase physical activity among 13- to 14-year-old adolescents: the GoActive cluster RCT', *Public Health Research*, 9(6), pp. 1–134. Available at: https://doi.org/10.3310/phr09060.

CORNFIELD, J. (1978) 'RANDOMIZATION BY GROUP: A FORMAL ANALYSIS', *American Journal of Epidemiology*, 108(2), pp. 2–4.

Crespi, C.M. (2016) 'Improved Designs for Cluster Randomized Trials', *Annu. Rev. Public Health*, 37, pp. 1–16. Available at: https://doi.org/10.1146/annurev-publhealth-032315-021702.

Crespi, C.M., Wong, W.K. and Mishra, S.I. (2009) 'Using second-order generalized estimating equations to model heterogeneous intraclass correlation in cluster-randomized trials', *Statistics in Medicine*, 28(5), pp. 814–827. Available at: https://doi.org/10.1002/sim.3518.

Daniel, R., Zhang, J. and Farewell, D. (2021) 'Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets', *Biometrical Journal*, 63(3), pp. 528–557. Available at: https://doi.org/10.1002/bimj.201900297.

Davies, M.J. *et al.* (2017) 'A community-based primary prevention programme for type 2 diabetes mellitus integrating identification and lifestyle intervention for prevention: a cluster randomised controlled trial', *Programme Grants for Applied Research*, 5(2), pp. 1–290. Available at: https://doi.org/10.3310/pgfar05020.

Díaz-Ordaz, K. *et al.* (2014) 'Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines', *Clinical Trials*, 11(5), pp. 590–600. Available at: https://doi.org/10.1177/1740774514537136.

Diehr, P. *et al.* (1995) 'Breaking the matches in a paired t-test for community interventions when the number of pairs is small', *Statistics in Medicine*, 14(13), pp. 1491–1504. Available at: https://doi.org/10.1002/sim.4780141309.

Donner, A. (1985) 'A regression approach to the analysis of data arising from cluster randomization', *International Journal of Epidemiology*, 14(2), pp. 322–326. Available at: https://doi.org/10.1093/ije/14.2.322.

Donner, A. (1986) *A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model*.

Donner, A. (1998) 'Some aspects of the design and analysis of cluster randomization trials', *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 47(1), pp. 95–113. Available at: https://doi.org/10.1111/1467-9876.00100.

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), pp. 795–800. Available at: https://doi.org/10.1093/ije/19.4.795.

Donner, A. and Klar, N. (1994) 'Methods for comparing event rates in intervention studies when the unit of allocation is a cluster', *American Journal of Epidemiology*, 140(3), pp. 279–289. Available at: https://doi.org/10.1093/oxfordjournals.aje.a117247.

Dormandy, E. *et al.* (2010) 'Antenatal screening for haemoglobinopathies in primary care: A cohort study and cluster randomised trial to inform a simulation model. The screening for haemoglobinopathies in first trimester (SHIFT) trial', *Health Technology Assessment*, 14(20), pp. 1–160. Available at: https://doi.org/10.3310/hta14200.

Du, R. and Lee, J.H. (2019) 'A weighted Jackknife method for clustered data', *Communications in Statistics - Theory and Methods*, 48(8), pp. 1963–1980. Available at: https://doi.org/10.1080/03610926.2018.1440597.

Dziura, J.D. *et al.* (2013) *Strategies for dealing with Missing data in clinical trials: From design to Analysis*, *YALE JOuRNAL OF BIOLOGY AND MEDICINE*.

Ebbutt, A.F. (1984) 'Three-Period Crossover Designs for Two Treatments', *Biometrics*, 40(1), pp. 219–224. Available at: https://doi.org/10.2307/2530762.

Elbourne, D.R., Campbell, M.K. and D.R., E. (2001) 'Extending the CONSORT statement to cluster randomized trials: For discussion', *Statistics in Medicine*, 20(3), pp. 489–496. Available at: https://doi.org/10.1002/1097-0258(20010215)20:3<489::AID-SIM806>3.0.CO;2-S.

Eldridge, S. *et al.* (2001) 'Sample size calculations for intervention trials in primary care randomizing by primary care group: an empirical illustration from one proposed intervention trial', *STATISTICS IN MEDICINE*, 20(3), pp. 367–376. Available at: https://doi.org/10.1002/1097-0258(20010215)20:3<367::AID-SIM798>3.0.CO;2-R.

Eldridge, S. and Kerry, S. (2012) 'Sample Size Calculations', in *A Practical Guide to Cluster Randomised Trials in Health Services Research*, pp. 137–171. Available at: https://doi.org/10.1002/9781119966241.ch7.

Eldridge, S.M., Ukoumunne, O.C. and Carlin, J.B. (2009) 'The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions', *International Statistical Review*, 77(3), pp. 378–394. Available at: https://doi.org/10.1111/j.1751-5823.2009.00092.x.

Fay, M.P. and Graubard, B.I. (2001) 'Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators', *Biometrics*, 57(4), pp. 1198–1206. Available at: https://doi.org/10.1111/j.0006-341X.2001.01198.x.

Fiero, M.H. *et al.* (2016) 'Statistical analysis and handling of missing data in cluster randomized trials: A systematic review', *Trials*, 17(1). Available at: https://doi.org/10.1186/s13063-016-1201-z.

Fitzmaurice, G. *et al.* (2008) *Longitudinal Data Analysis*. 0 edn. Chapman and Hall/CRC. Available at: https://www.taylorfrancis.com/books/9781420011579.

Forbes, A.B. *et al.* (2015) 'Cluster randomised crossover trials with binary data and unbalanced cluster sizes: Application to studies of near-universal interventions in intensive care', *Clinical Trials*, 12(1), pp. 34–44. Available at: https://doi.org/10.1177/1740774514559610.

Foy, R. *et al.* (2020) 'Developing and evaluating packages to support implementation of quality indicators in general practice: the ASPIRE research programme, including two cluster RCTs', *Programme Grants Appl Res*, 8(4). Available at: https://doi.org/10.3310/pgfar08040.

Gail, M.H. *et al.* (1992) 'Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT)', *Controlled Clinical Trials*, 13(1), pp. 6–21. Available at: https://doi.org/10.1016/0197-2456(92)90026-V.

Gao, F. *et al.* (2015) 'Sample size calculations for the design of cluster randomized trials: A summary of methodology', *Contemporary Clinical Trials*, 42, pp. 41–50. Available at: https://doi.org/10.1016/j.cct.2015.02.011.

Gates, S. *et al.* (2017) 'Prehospital randomised assessment of a mechanical compression device in out-of-hospital cardiac arrest (PARAMEDIC): A pragmatic, cluster randomised trial and economic evaluation', *Health Technology Assessment*, 21(11), pp. 1–175. Available at: https://doi.org/10.3310/hta21110.

Gaughran, F. *et al.* (2020) 'A health promotion intervention to improve lifestyle choices and health outcomes in people with psychosis: a research programme including the IMPaCT RCT', *Programme Grants Appl Res*, 8(1). Available at: https://doi.org/10.3310/pgfar08010.

Gogtay, N. (2019) 'Reporting of randomized controlled trials: Will it ever improve?', *Perspectives in Clinical Research*, 10(2), p. 49. Available at: https://doi.org/10.4103/picr.PICR_11_19.

Grant, M.J. and Booth, A. (2009) 'A typology of reviews: an analysis of 14 review types and associated methodologies: A typology of reviews, *Maria J. Grant & Andrew Booth*', *Health Information & Libraries Journal*, 26(2), pp. 91–108. Available at: https://doi.org/10.1111/j.1471-1842.2009.00848.x.

Group, T.G.H.S. (1987) 'The Gambia Hepatitis Intervention Study', *Cancer Res*, 47(21), pp. 5782–5787.

Gulliford, M.C. *et al.* (2019) 'Electronically delivered interventions to reduce antibiotic prescribing for respiratory infections in primary care: cluster RCT using electronic health records and cohort study', *Health Technol Assess*, 23, p. 11. Available at: https://doi.org/10.3310/hta23110.

Hall, D.B. and Severini, T.A. (1998) 'Extended Generalized Estimating Equations for Clustered Data', *Journal of the American Statistical Association*, 93(444), pp. 1365–1375. Available at: https://doi.org/10.1080/01621459.1998.10473798.

Handayani, D. *et al.* (2017) 'A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (GLMM)', in *AIP Conference Proceedings*. American Institute of Physics Inc. Available at: https://doi.org/10.1063/1.4979449.

Hansen, L.P. (2010) 'Generalized method of moments estimation', in S.N. Durlauf and L.E. Blume (eds) *Macroeconometrics and Time Series Analysis*. London: Palgrave Macmillan UK, pp. 105–118. Available at: http://link.springer.com/10.1057/9780230280830_13.

Harrington, D.M. *et al.* (2019) 'A school-based intervention ("Girls Active") to increase physical activity levels among 11- to 14-year-old girls: cluster RCT', *Public Health Research*, 7(5), pp. 1–162. Available at: https://doi.org/10.3310/phr07050.

Harris, T. *et al.* (2018) 'A pedometer-based walking intervention in 45- to 75-year-olds, with and without practice nurse support: The PACE-UP three-arm cluster RCT', *Health Technology Assessment*, 22(37), pp. 1–273. Available at: https://doi.org/10.3310/hta22370.

Hartinger, S.M. *et al.* (2020) 'A factorial cluster-randomised controlled trial combining home-environmental and early child development interventions to improve child health and development: Rationale, trial design and baseline findings', *BMC Medical Research Methodology*, 20(1). Available at: https://doi.org/10.1186/s12874-020-00950-y.

Hassiotis, A. *et al.* (2018) 'Positive behaviour support training for staff for treating challenging behaviour in people with intellectual disabilities: A cluster RCT', *Health Technology Assessment*, 22(15), pp. 1–110. Available at: https://doi.org/10.3310/hta22150.

Hauck, W.W., Anderson, S. and Marcus, S.M. (1998) 'Should We Adjust for Covariates in Nonlinear Regression Analyses of Randomized Trials?', *Controlled Clinical Trials*, 19(3), pp. 249–256. Available at: https://doi.org/10.1016/S0197-2456(97)00147-5.

Hayes, R.J. and Moulton, L.H. (2009) *Cluster Randomised Trials*. 0 edn. Chapman and Hall/CRC. Available at: https://www.taylorfrancis.com/books/9781584888178.

Heller, S. *et al.* (2014) 'Improving management of type 1 diabetes in the UK: the Dose Adjustment For Normal Eating (DAFNE) programme as a research test-bed. A mixed-method analysis of the barriers to and facilitators of successful diabetes self-management, a health economic analys', *Programme Grants for Applied Research*, 2(5), pp. 1–188. Available at: https://doi.org/10.3310/pgfar02050.

Heller, S. *et al.* (2017) 'A cluster randomised trial, cost-effectiveness analysis and psychosocial evaluation of insulin pump therapy compared with multiple injections during flexible intensive insulin therapy for type 1 diabetes: The REPOSE Trial', *Health Technology Assessment*, 21(20), pp. 1–277. Available at: https://doi.org/10.3310/hta21200.

Hemming, K. *et al.* (2015) 'The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting', *BMJ (Online)*, 350. Available at: https://doi.org/10.1136/bmj.h391.

Hemming, K. *et al.* (2021) 'Contamination: How much can an individually randomized trial tolerate?', *Statistics in Medicine*, 40(14), pp. 3329–3351. Available at: https://doi.org/10.1002/sim.8958.

Hemming, K. and Taljaard, M. (2023) 'Key considerations for designing, conducting and analysing a cluster randomized trial', *International Journal of Epidemiology*, 52(5), pp. 1648–1658. Available at: https://doi.org/10.1093/ije/dyad064.

Heo, M. and Leon, A.C. (2005) 'Performance of a mixed effects logistic regression model for binary outcomes with unequal cluster size', *Journal of Biopharmaceutical Statistics*, 15(3), pp. 513–526. Available at: https://doi.org/10.1081/BIP-200056554.

Heymans, M.W. and Twisk, J.W.R. (2022) 'Handling missing data in clinical research', *Journal of Clinical Epidemiology*, 151, pp. 185–188. Available at: https://doi.org/10.1016/j.jclinepi.2022.08.016.

Ho, M.W. *et al.* (2013) 'A nested Dirichlet process analysis of cluster randomized trial data with application in geriatric care assessment', *Journal of the American Statistical Association*, 108(501), pp. 48–68. Available at: https://doi.org/10.1080/01621459.2012.734164.

Højsgaard, S., Halekoh, U. and Yan, J. (2005) 'The R Package geepack for Generalized Estimating Equations', *Journal of Statistical Software*, 15(2), pp. 1–11. Available at: https://doi.org/10.18637/jss.v015.i02.

Horton, N.J. *et al.* (1999) 'Goodness-of-fit for GEE: An example with mental health service utilization', *Statistics in Medicine*, 18(2), pp. 213–222. Available at: https://doi.org/10.1002/(SICI)1097-0258(19990130)18:2<213::AID-SIM999>3.0.CO;2-E.

Hosmer, D.W. *et al.* (1997) 'A COMPARISON OF GOODNESS-OF-FIT TESTS FOR THE LOGISTIC REGRESSION MODEL', *Statistics in Medicine*, 16(9), pp. 965–980. Available at: https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O.

Hosmer, D.W. and Lemesbow, S. (1980) 'Goodness of fit tests for the multiple logistic regression model', *Communications in Statistics - Theory and Methods*, 9(10), pp. 1043–1069. Available at: https://doi.org/10.1080/03610928008827941.

Hossain, A. *et al.* (2017) 'Missing continuous outcomes under covariate dependent missingness in cluster randomised trials', *Statistical Methods in Medical Research*, 26(3), pp. 1543–1562. Available at: https://doi.org/10.1177/0962280216648357.

Hossain, A. and Bartlett, J.W. (2017) 'Missing binary outcomes under covariate-dependent missingness in', *Statistical Methods in Medical Research*, 36(19), pp. 3092–3109. Available at: https://doi.org/10.1002/sim.7334.

Hubbard, A.E. *et al.* (2010) 'To GEE or Not to GEE', *Epidemiology.*, 21(4), pp. 467–474.

Humphrey, N. *et al.* (2018) 'The PATHS curriculum for promoting social and emotional well-being among children aged 7–9 years: a cluster RCT', *Public Health Research*, 6(10), pp. 1–116. Available at: https://doi.org/10.3310/phr06100.

Hussey, M.A. and Hughes, J.P. (2007) 'Design and analysis of stepped wedge cluster randomized trials', *Contemporary Clinical Trials*, 28(2), pp. 182–191. Available at: https://doi.org/10.1016/j.cct.2006.05.007.

Iliffe, S. *et al.* (2014) 'Multicentre cluster randomised trial comparing a community group exercise programme and home-based exercise with usual care for people aged 65 years and over in primary care', *Health Technology Assessment*, 18(49), pp. 1–105. Available at: https://doi.org/10.3310/hta18490.

Ivers, N.M. *et al.* (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 (Online)*, 343(7827). Available at: https://doi.org/10.1136/bmj.d5886.

Johnson, J.L. *et al.* (2015) 'Recommendations for choosing an analysis method that controls Type I error for unbalanced cluster sample designs with Gaussian outcomes', *Statistics in Medicine*, 34(27), pp. 3531–3545. Available at: https://doi.org/10.1002/sim.6565.

Julious, S.A. *et al.* (2016) 'PLEASANT: Preventing and lessening exacerbations of Asthma in school-age children associated with a new term-a cluster randomised controlled trial and economic evaluation', *Health Technology Assessment*, 20(93), pp. 1–154. Available at: https://doi.org/10.3310/hta20930.

Julious, S.A. (2023) *Sample Sizes for Clinical Trials*. 2nd edn. Chapman & Hall/CRC.

Kahan, B.C. *et al.* (2014) 'The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies', *Trials*, 15(1). Available at: https://doi.org/10.1186/1745-6215-15-139.

Kahan, B.C. *et al.* (2023) 'Estimands in cluster-randomized trials: Choosing analyses that answer the right question', *International Journal of Epidemiology*, 52(1), pp. 107–118. Available at: https://doi.org/10.1093/ije/dyac131.

Kang, W., Lee, M.-S. and Lee, Y. (2005) 'HGLM versus conditional estimators for the analysis of clustered binary data', *Statistics in Medicine*, 24(5), pp. 741–752. Available at: https://doi.org/10.1002/sim.1772.

Killaspy, H. *et al.* (2017) 'The Rehabilitation Effectiveness for Activities for Life (REAL) study: a national programme of research into NHS inpatient mental health rehabilitation services across England', *Programme Grants for Applied Research*, 5(7), pp. 1–284. Available at: https://doi.org/10.3310/pgfar05070.

Kim, H.-Y. *et al.* (2006) 'Multilevel analysis of group-randomized trials with binary outcomes', *Community Dentistry and Oral Epidemiology*, 34(4), pp. 241–251. Available at: https://doi.org/10.1111/j.1600-0528.2006.00307.x.

Kitchener, H.C. *et al.* (2016) 'A cluster randomised trial of strategies to increase cervical screening uptake at first invitation (STRATEGIC)', *Health Technol Assess*, 20(68). Available at: https://doi.org/10.3310/hta20680.

Lam, K.F. and Ip, D. (2003) 'REML and ML estimation for clustered grouped survival data', *Statistics in Medicine*, 22(12), pp. 2025–2034. Available at: https://doi.org/10.1002/sim.1323.

Lamb, S.E. *et al.* (2012) 'Managing injuries of the neck trial (mint): A randomised controlled trial of treatments for whiplash injuries', *Health Technology Assessment*, 16(49), pp. 1–141. Available at: https://doi.org/10.3310/hta16490.

Lawlor, D.A. *et al.* (2016) 'Active for Life Year 5: a cluster randomised controlled trial of a primary school-based intervention to increase levels of physical activity, decrease sedentary behaviour and improve diet', *Public Health Research*, 4(7), pp. 1–156. Available at: https://doi.org/10.3310/phr04070.

Lewis, J. and Julious, S.A. (2021) 'Sample sizes for cluster-randomised trials with continuous outcomes: Accounting for uncertainty in a single intra-cluster correlation estimate', *Statistical Methods in Medical Research*, 30(11), pp. 2459–2470. Available at: https://doi.org/10.1177/09622802211037073.

Leyrat, C. *et al.* (2018) 'Cluster randomized trials with a small number of clusters: Which analyses should be used?', *International Journal of Epidemiology*, 47(1), pp. 321–331. Available at: https://doi.org/10.1093/ije/dyx169.

Li, J. and Jung, S. (2020) 'Sample size calculation for cluster randomization trials with a time-to-event endpoint', *Statistics in Medicine*, 39(25), pp. 3608–3623. Available at: https://doi.org/10.1002/sim.8683.

Li, Z., Xu, X. and Shen, J. (2017) 'Semiparametric Bayesian analysis of accelerated failure time models with cluster structures', *Statistics in Medicine*, 36(25), pp. 3976–3989. Available at: https://doi.org/10.1002/sim.7406.

Liang, K.-Y., Zeger, S.L. and Qaqish, B. (1992) 'Multivariate Regression Analyses for Categorical Data', *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1), pp. 3–40. Available at: http://www.jstor.org/stable/2345947.

Liang and Zeger (1986) 'Longitudinal Data Analysis Using Generalized Linear Models', *Biometrics*, 42(1), pp. 121–130.

Litière, S., Alonso, A. and Molenberghs, G. (2008) 'The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models', *Statistics in Medicine*, 27(16), pp. 3125–3144. Available at: https://doi.org/10.1002/sim.3157.

Lu, S.-E. and Wang, M.-C. (2005) 'Marginal analysis for clustered failure time data', *Lifetime Data Analysis*, 11(1), pp. 61–79. Available at: https://doi.org/10.1007/s10985-004-5640-6.

Ma, J. *et al.* (2009) '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 medical research methodology*, 9(1), p. 37. Available at: https://doi.org/10.1186/1471-2288-9-37.

Ma, J. *et al.* (2013) 'Comparison of population-averaged and cluster-specific models for the analysis of cluster randomized trials with missing binary outcomes: a simulation study.', *BMC medical research methodology*, 13, p. 9. Available at: https://doi.org/10.1186/1471-2288-13-9.

Macarthur, C. *et al.* (2003) 'Redesigning postnatal care: a randomised controlled trial of protocol-based midwifery-led care focused on individual women's physical and psychological health needs', *Health Technology Assessment 2*, 7(37), pp. 1–98.

Machin, D. and Campbell, M.J. (2005) *The Design of Studies for Medical Research*. Wiley. Available at: https://books.google.co.uk/books?id=rrUIRz1Y9nEC.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. 2nd edn. Routledge. Available at: https://www.taylorfrancis.com/books/9781351445856.

McCulloch, C.E. (1997) 'Maximum Likelihood Algorithms for Generalized Linear Mixed Models', *Journal of the American Statistical Association*, 92(437), pp. 162–170. Available at: https://doi.org/10.1080/01621459.1997.10473613.

McCulloch, C.E. and Neuhaus, J.M. (2011) 'Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter', *Statistical Science*, 26(3), pp. 388–402. Available at: https://doi.org/10.1214/11-STS361.

McCulloch, C.E. and Searle, S.R. (2000) *Generalized, Linear, and Mixed Models*. 1st edn, *Wiley Series in Probability and Statistics*. 1st edn. Wiley. Available at: https://onlinelibrary.wiley.com/doi/book/10.1002/0471722073.

McDonald, A.M. *et al.* (2006) 'What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies', *Trials*, 7, pp. 1–8. Available at: https://doi.org/10.1186/1745-6215-7-9.

McNeish, D.M. (2017). Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction. *Multivariate Behavioral Research, 52*, 661 - 670.

McNeish, D. and Stapleton, L.M. (2016) 'Modeling Clustered Data with Very Few Clusters', *Multivariate behavioral research*, 51(4), pp. 495–518. Available at: https://doi.org/http://dx.doi.org/10.1080/00273171.2016.1167008.

Mdege, N.D. *et al.* (2014) 'The 2 × 2 cluster randomized controlled factorial trial design is mainly used for efficiency and to explore intervention interactions: A systematic review', *Journal of Clinical Epidemiology*. Elsevier USA, pp. 1083–1092. Available at: https://doi.org/10.1016/j.jclinepi.2014.06.004.

Moberg, J. and Kramer, M. (2015) 'A brief history of the cluster randomised trial design', *Journal of the Royal Society of Medicine*, 108(5), pp. 192–198. Available at: https://doi.org/10.1177/0141076815582303.

Moerbeek, M. and Van Schie, S. (2016) 'How large are the consequences of covariate imbalance in cluster randomized trials: A simulation study with a continuous outcome and a binary covariate at the cluster level', *BMC Medical Research Methodology*, 16(1). Available at: https://doi.org/10.1186/s12874-016-0182-7.

Molas, M. and Lesaffre, E. (2010) 'Hurdle models for multilevel zero-inflated data via h-likelihood', *Statistics in Medicine*, 29(30), pp. 3294–3310. Available at: https://doi.org/10.1002/sim.3852.

Mollison, J. *et al.* (2000a) 'Comparison of analytical methods for cluster randomised trials: an example from a primary care setting.', *Journal of epidemiology and biostatistics*, 5(6), pp. 339–348. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0034568262&partnerID=40&md5=565e82c603d0af5029a9a0ed2eb3fd29.

Mollison, J. *et al.* (2000b) 'Comparison of analytical methods for cluster randomised trials: an example from a primary care setting.', *Journal of epidemiology and biostatistics*, 5(6), pp. 339–348. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0034568262&partnerID=40&md5=565e82c603d0af5029a9a0ed2eb3fd29.

Moniz-Cook, E. *et al.* (2017) 'Challenge Demcare: management of challenging behaviour in dementia at home and in care homes – development, evaluation and implementation of an online individualised intervention for care homes; and a cohort study of specialist community mental health car', *Programme Grants for Applied Research*, 5(15), pp. 1–290. Available at: https://doi.org/10.3310/pgfar05150.

Morgan, K. *et al.* (2004) 'Psychological treatment for insomnia in the regulation of long-term hypnotic drug use', *Health Technology Assessment*, 8(8). Available at: https://doi.org/10.3310/hta8080.

Morgan, K.E. *et al.* (2016) 'Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome', *Statistics in Medicine*, 36(2), pp. 318–333. Available at: https://doi.org/10.1002/sim.7137.

Morgan, K.E. *et al.* (2017) 'Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome', *Statistics in Medicine*, 36(2), pp. 318–333. Available at: https://doi.org/10.1002/sim.7137.

Morrell, C.J. *et al.* (2009) 'Psychological interventions for postnatal depression: Cluster randomised trial and economic evaluation. The PoNDER trial', *Health Technology Assessment*. Available at: https://doi.org/10.3310/hta13300.

Morris, T.P., White, I.R. and Crowther, M.J. (2019) 'Using simulation studies to evaluate statistical methods', *Statistics in Medicine*, 38(11), pp. 2074–2102. Available at: https://doi.org/10.1002/sim.8086.

Mouncey, P.R. *et al.* (2019) 'A nurse-led, preventive, psychological intervention to reduce PTSD symptom severity in critically ill patients: the POPPI feasibility study and cluster RCT', *Health Services and Delivery Research*, 7(30), pp. 1–174. Available at: https://doi.org/10.3310/hsdr07300.

Müller, P., Quintana, F.A. and Rosner, G.L. (2007) 'Semiparametric Bayesian inference for multilevel repeated measurement data', *Biometrics*. Blackwell Publishing Inc., pp. 280–289. Available at: https://doi.org/10.1111/j.1541-0420.2006.00668.x.

Munn, Z. *et al.* (2018) 'Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach', *BMC Medical Research Methodology*, 18(1). Available at: https://doi.org/10.1186/s12874-018-0611-x.

Murray, D.M. *et al.* (2004) 'Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments', *American Journal of Public Health*, 94(3), pp. 423–432. Available at: https://doi.org/10.2105/AJPH.94.3.423.

Murray, D.M. *et al.* (2006) 'A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial', *Statistics in Medicine*, 25(3), pp. 375–388. Available at: https://doi.org/10.1002/sim.2233.

Murray, D.M. *et al.* (2008) 'Design and analysis of group-randomized trials in cancer: A review of current practices', *Journal of the National Cancer Institute*, 100(7), pp. 483–491. Available at: https://doi.org/10.1093/jnci/djn066.

Nelder, J.A. and Wedderburn, R.W.M. (1972) 'Generalized Linear Models', *Journal of the Royal Statistical Society. Series A (General)*, 135(3), p. 370. Available at: https://doi.org/10.2307/2344614.

Neuhaus, J.M. and McCulloch, C.E. (2011) 'Estimation of covariate effects in generalized linear mixed models with informative cluster sizes', *Biometrika*, 98(1), pp. 147–162. Available at: https://doi.org/10.1093/biomet/asq066.

O'Cathain, A. *et al.* (2002) 'Use of evidence based leaflets to promote informed choice in maternity care: Randomised controlled trial in everyday practice', *British Medical Journal*, 324(7338), pp. 643–646. Available at: https://doi.org/10.1136/bmj.324.7338.643.

Odueyungbo, A. *et al.* (2008) 'Comparison of generalized estimating equations and quadratic inference functions using data from the National Longitudinal Survey of Children and Youth (NLSCY) database', *BMC medical research methodology*, 8(28), pp. 1–10. Available at: https://doi.org/10.1186/1471-2288-8-28.

Offorha, B., Walters, S. and Jacques, R. (2022) 'Statistical analysis of publicly funded cluster randomised controlled trials: a review of the National Institute for Health Research Journals Library', *Trials*. BioMed Central Ltd. Available at: https://doi.org/10.1186/s13063-022-06025-1.

Offorha, B., Walters, S. and Jacques, R. (2023) 'Analysing cluster randomised controlled trials using GLMM, GEE1, GEE2, and QIF: results from four case studies', *BMC Medical Research Methodology*, 23(1), p. 293. Available at: https://doi.org/10.1186/s12874-023-02107-z.

Offorha, B.C., Walters, S.J. and Jacques, R.M. (2022) 'Statistical analysis of publicly funded cluster randomised controlled trials: a review of the National Institute for Health Research Journals Library', *Trials*, 23(1), p. 115. Available at: https://doi.org/10.1186/s13063-022-06025-1.

Olsen, M.K. *et al.* (2008) 'Strategies for analyzing multilevel cluster-randomized studies with binary outcomes collected at varying intervals of time', *Statistics in Medicine*, 27(29), pp. 6055–6071. Available at: https://doi.org/10.1002/sim.3446.

Omar, R.Z. *et al.* (2000) 'Analysis of a cluster randomized trial with binary outcome data using a multi-level model', *Statistics in Medicine*, 19(19), pp. 2675–2688. Available at: https://doi.org/10.1002/1097-0258(20001015)19:19<2675::AID-SIM556>3.0.CO;2-A.

Pacheco, G.D. *et al.* (2009) 'Performance of analytical methods for overdispersed counts in cluster randomized trials: Sample size, degree of clustering and imbalance', *Statistics in Medicine*, 28(24), pp. 2989–3011. Available at: https://doi.org/10.1002/sim.3681.

Page, M.J. *et al.* (2021) 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', *BMJ*, p. n71. Available at: https://doi.org/10.1136/bmj.n71.

Pan, C., Cai, B. and Wang, L. (2017) 'Multiple frailty model for clustered interval-censored data with frailty selection', *Statistical Methods in Medical Research*, 26(3), pp. 1308–1322. Available at: https://doi.org/10.1177/0962280215576987.

Pan, W. (2001) *Akaike's Informat ion Criterion in Generalized Estimating Equations*.

Pan, W. (2002) *Goodness-of-Fit Tests for GEE with Correlated Binary Data*, *Scandinavian Journal of Statistics*.

Peden, C.J. *et al.* (2019) 'A national quality improvement programme to improve survival after emergency abdominal surgery: the EPOCH stepped-wedge cluster RCT', *Health Services and Delivery Research*, 7(32), pp. 1–96. Available at: https://doi.org/10.3310/hsdr07320.

Pedroza, C. and Truong, V.T.T. (2017) 'Estimating relative risks in multicenter studies with a small number of centers - which methods to use? A simulation study', *Trials*, 18(1). Available at: https://doi.org/10.1186/s13063-017-2248-1.

Peek, N., Goud, R. and De Keizer, N. (2013) 'Handling intra-cluster correlation when analyzing the effects of decision support on health care process measures', in *Studies in Health Technology and Informatics*. Dept. of Medical Informatics, University of Amsterdam, PO Box 22700, 1100 DD Amsterdam, Netherlands, pp. 22–27. Available at: https://doi.org/10.3233/978-1-61499-240-0-22.

Perez, J. *et al.* (2016) 'Understanding causes of and developing effective interventions for schizophrenia and other psychoses', *Programme Grants Appl Res*, 4(2). Available at: https://doi.org/10.3310/pgfar04020.

Perin, J. and Preisser, J.S. (2016) 'Alternating logistic regressions with improved finite sample properties', *Biometrics*, 73(2), pp. 696–705. Available at: https://doi.org/10.1111/biom.12614.

Peters, M. *et al.* (2020) *JBI Manual for Evidence Synthesis*, *JBI Manual for Evidence Synthesis*. Available at: https://doi.org/10.46658/JBIMES-20-01.

Peters, T.J. *et al.* (2003) 'Comparison of methods for analysing cluster randomized trials: An example involving a factorial design', *International Journal of Epidemiology*, 32(5), pp. 840–846. Available at: https://doi.org/10.1093/ije/dyg228.

Pladevall, M. *et al.* (2014) 'Designing multicenter cluster randomized trials: an introductory toolkit'.

Prague, M. *et al.* (2016) 'Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes', *Biometrics*, 72(4), pp. 1066–1077. Available at: https://doi.org/10.1111/biom.12519.

Prentice, R.L. (1988) 'Correlated Binary Regression with Covariates Specific to Each Binary Observation', *Biometrics*, 44(4), p. 1033. Available at: https://doi.org/10.2307/2531733.

Prentice, R.L. and Zhao, L.P. (1991) *Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses*. Available at: https://about.jstor.org/terms.

Qu, A., Lindsay, B.G. and Bing, L.I. (2000) 'Improving generalised estimating equations using quadratic inference functions', *Biometrika*, 87(4), pp. 823–836. Available at: https://doi.org/10.1093/biomet/87.4.823.

Qu, A. and Song, P.X.K. (2004) 'Assessing robustness of generalised estimating equations and quadratic inference functions', *Biometrika*, 91(2), pp. 447–459. Available at: https://doi.org/10.1093/biomet/91.2.447.

Raine, R. *et al.* (2017) 'Testing innovative strategies to reduce the social gradient in the uptake of bowel cancer screening: a programme of four qualitatively enhanced randomised controlled trials', *Programme Grants Appl Res*, 5(8), p. 338. Available at: https://doi.org/10.3310/pgfar05080.

Ramsay, C.R. *et al.* (2018) 'Improving the quality of dentistry (IQuaD): A cluster factorial randomised controlled trial comparing the effectiveness and cost-benefit of oral hygiene advice and/or periodontal instrumentation with

routine care for the prevention and management of perio', *Health Technology Assessment*, 22(38), pp. vii–143. Available at: https://doi.org/10.3310/hta22380.

Relton, C. *et al.* (2018) 'Effect of Financial Incentives on Breastfeeding A Cluster Randomized Clinical Trial', *JAMA - Journal of the American Medical Association*, 172(2), pp. 1–7. Available at: https://doi.org/10.1001/jamapediatrics.2017.4523.

Ribeiro, D.C., Milosavljevic, S. and Abbott, J.H. (2018) 'Sample size estimation for cluster randomized controlled trials', *Musculoskeletal Science and Practice*, 34, pp. 108–111. Available at: https://doi.org/10.1016/j.msksp.2017.10.002.

Ridout, M.S., Demétrio, C.G.B. and Firth, D. (1999) 'Estimating intraclass correlation for binary data', *Biometrics*, 55(1), pp. 137–148. Available at: https://doi.org/10.1111/j.0006-341X.1999.00137.x.

Ritz, J. and Spiegelman, D. (2004) 'Equivalence of conditional and marginal regression models for clustered and longitudinal data', *Statistical Methods in Medical Research*. Arnold, pp. 309–323. Available at: https://doi.org/10.1191/0962280204sm368ra.

Rothwell, J.C., Julious, S.A. and Cooper, C.L. (2018) 'A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal', *Trials*, 19(1), p. 544. Available at: https://doi.org/10.1186/s13063-018-2886-y.

Rubin, D.B. (1976) 'Inference and missing data', *Biometrika*, 63(3), pp. 581–592.

Rutterford, C. *et al.* (2015) 'Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: A review', *Journal of Clinical Epidemiology*, 68(6), pp. 716–723. Available at: https://doi.org/10.1016/j.jclinepi.2014.10.006.

Salisbury, C. *et al.* (2019) 'A patient-centred intervention to improve the management of multimorbidity in general practice: the 3D RCT', *Health Services and Delivery Research*, 7(5), pp. 1–238. Available at: https://doi.org/10.3310/hsdr07050.

Samsa, G. and Neely, M. (2018) 'Two questions about the analysis and interpretation of randomised trials', *International Journal of Hyperthermia*, 34(8), pp. 1396–1399. Available at: https://doi.org/10.1080/02656736.2017.1385861.

Sarkodie, S.K., Wason, J.M.S. and Grayling, M.J. (2023) 'A hybrid approach to comparing parallel-group and stepped-wedge cluster-randomized trials with a continuous primary outcome when there is uncertainty in the intra-cluster correlation', *Clinical Trials*, 20(1), pp. 59–70. Available at: https://doi.org/10.1177/17407745221123507.

Sauzet, O. *et al.* (2013) 'Modelling the hierarchical structure in datasets with very small clusters: A simulation study to explore the effect of the proportion of clusters when the outcome is continuous', *Statistics in Medicine*, 32(8), pp. 1429–1438. Available at: https://doi.org/10.1002/sim.5638.

Schulz, K.F. *et al.* (2010) 'CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials', *BMC Medicine*, 8(1), p. 18. Available at: https://doi.org/10.1186/1741-7015-8-18.

Senn, S.S. (2002) *Cross-over Trials in Clinical Research*. Wiley (Statistics in Practice). Available at: https://books.google.co.uk/books?id=MWbHzwpPTgEC.

Simmons, R.K. *et al.* (2016) 'A randomised trial of the effect and cost-effectiveness of early intensive multifactorial therapy on 5-year cardiovascular outcomes in individuals with screen-detected type 2 diabetes: The Anglo–Danish–Dutch Study of Intensive treatment in people with scr', *Health Technology Assessment*, 20(64), pp. 1–86. Available at: https://doi.org/10.3310/hta20640.

Snooks, H. *et al.* (2018) 'Predictive risk stratification model: a randomised stepped-wedge trial in primary care (PRISMATIC)', *Health Services and Delivery Research*, 6(1), pp. 1–164. Available at: https://doi.org/10.3310/hsdr06010.

Snooks, H.A. *et al.* (2017) 'Support and assessment for fall emergency referrals (SAFER) 2: A cluster randomised trial and systematic review of clinical effectiveness and cost-effectiveness of new protocols for emergency ambulance paramedics to assess older people following a fall wi', *Health Technology Assessment*, 21(13), pp. 1–218. Available at: https://doi.org/10.3310/hta21130.

Song, P. (2007) *Correlated Data Analysis: Modeling, Analytics, and Applications*. 1st edn, *Springer Series in Statistics*. 1st edn. New York, NY: Springer New York. Available at: http://link.springer.com/10.1007/978-0-387-71393-9.

Song, P.X.-K. *et al.* (2009) 'Quadratic inference functions in marginal models for longitudinal data', *Statistics in Medicine*, 28(29), pp. 3683–3696. Available at: https://doi.org/10.1002/sim.3719.

Speed, C. *et al.* (2010) 'LIFELAX – diet and LIFEstyle versus LAXatives in the management of chronic constipation in older people : randomised controlled trial', 14(52).

Sully, B.G.O., Julious, S.A. and Nicholl, J. (2013) 'A reinvestigation of recruitment to randomised, controlled, multicenter trials: A review of trials funded by two UK funding agencies', *Trials*, 14(1), pp. 1–9. Available at: https://doi.org/10.1186/1745-6215-14-166.

Sumnall, H. *et al.* (2017) 'Steps Towards Alcohol Misuse Prevention Programme (STAMPP): a school- and community-based cluster randomised controlled trial', *Public Health Research*, 5(2), pp. 1–154. Available at: https://doi.org/10.3310/phr05020.

Surr, C.A. *et al.* (2020) 'Dementia care mapping™ to reduce agitation in care home residents with dementia: The epic cluster rct', *Health Technology Assessment*, 24(16), pp. 1–174. Available at: https://doi.org/10.3310/hta24160.

Tawiah, R. *et al.* (2019) 'Multilevel model with random effects for clustered survival data with multiple failure outcomes', *Statistics in Medicine*, 38(6), pp. 1036–1055. Available at: https://doi.org/10.1002/sim.8041.

The COMMIT Research Group (1995) 'Community Intervention Trial for Smoking Cessation (COMMIT): I. cohort results from a four-year community intervention.', *American Journal of Public Health*, 85(2), pp. 183–192. Available at: https://doi.org/10.2105/AJPH.85.2.183.

Thompson, D.G. *et al.* (2018) 'A randomised controlled trial, cost-effectiveness and process evaluation of the implementation of self-management for chronic gastrointestinal disorders in primary care, and linked projects on identification and risk assessment', *Programme Grants for Applied Research*, 6(1), pp. 1–154. Available at: https://doi.org/10.3310/pgfar06010.

Thompson, J.A. *et al.* (2022) 'Cluster randomised trials with a binary outcome and a small number of clusters: comparison of individual and cluster level analysis method', *BMC Medical Research Methodology*, 22(1). Available at: https://doi.org/10.1186/s12874-022-01699-2.

Thompson, S.G., Warn, D.E. and Turner, R.M. (2004) 'Bayesian methods for analysis of binary outcome data in cluster randomized trials on the absolute risk scale', *Statistics in Medicine*, 23(3), pp. 389–410. Available at: https://doi.org/10.1002/sim.1567.

Tricco Erin; Zarin Wasifa; O'Brien Kelly K; Colquhoun Heather; Levac Danielle; Moher David; Peters Micah DJ; Horsley Tanya; Weeks Laura; Hempel Susanne; Akl Elie A; Chang Christine; McGowan Jessie; Stewart Lesley;

Hartling Simon, A.C.L. (2018) 'PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation', *Annals of internal medicine*, 169(7), pp. 467–473. Available at: https://doi.org/10.7326/M18-0850.

Tsiatis, A.A. (1980) 'A note on a goodness-of-fit test for the logistic regression model', *Biometrika*, 67(1), pp. 250–251. Available at: https://doi.org/10.1093/biomet/67.1.250.

Turner, E.L. (2017) 'GROUP-RANDOMIZED TRIALS : PART 2 - ANALYSIS', *Am J Public Health*, 107(7), pp. 1078–1086. Available at: https://doi.org/10.2105/AJPH.2017.303707.REVIEW.

Turner, J. *et al.* (2000) 'A randomised controlled trial of prehospital intravenous fluid', *Health technology assessment (Winchester, England)*, 4(31), pp. 1–57.

Twardella, D., Bruckner, T. and Blettner, M. (2005) '[Statistical analysis of community-based studies -- presentation and comparison of possible solutions with reference to statistical meta-analytic methods].', *Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany))*, 67(1), pp. 48–55. Available at: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med6&NEWS=N&AN=15672306.

Ukoumunne, O.C. *et al.* (1999) *Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review Review*, *HTA Health Technology Assessment NHS R&D HTA Programme Health Technology Assessment*. Available at: www.hta.ac.uk/htacd.htm.

Ukoumunne, O.C. (2002) 'A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials', *Statistics in Medicine*, 21(24), pp. 3757–3774. Available at: https://doi.org/10.1002/sim.1330.

Ukoumunne, O.C. *et al.* (2008) 'Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials', *Statistics in Medicine*, 27(25), pp. 5143–5155. Available at: https://doi.org/10.1002/sim.3359.

Ukoumunne, O.C., Carlin, J.B. and Gulliford, M.C. (2007) 'A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials', *Statistics in Medicine*, 26(18), pp. 3415–3428. Available at: https://doi.org/10.1002/sim.2769.

Walters, S.J. *et al.* (2017) 'Recruitment and retention of participants in randomised controlled trials: A review of trials funded and published by the United Kingdom Health Technology Assessment Programme', *BMJ Open*, 7(3), pp. 1–10. Available at: https://doi.org/10.1136/bmjopen-2016-015276.

Walters, S.J., Morrell, C.J. and Slade, P. (2011) 'Analysing data from a cluster randomized trial (cRCT) in primary care: A case study', *Journal of Applied Statistics*, 38(10), pp. 2253–2269. Available at: https://doi.org/10.1080/02664763.2010.545375.

Wang, R. *et al.* (2017) 'The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials', *Statistics in Medicine*, 36(18), pp. 2831–2843. Available at: https://doi.org/10.1002/sim.7329.

Westgate, P.M. (2012) 'A bias-corrected covariance estimate for improved inference with quadratic inference functions', *Statistics in Medicine*, 31(29), pp. 4003–4022. Available at: https://doi.org/10.1002/sim.5479.

Westgate, P.M. and Braun, T.M. (2012) 'The effect of cluster size imbalance and covariates on the estimation performance of quadratic inference functions', *Statistics in Medicine*, 31(20), pp. 2209–2222. Available at: https://doi.org/10.1002/sim.5329.

Westgate, P.M. and Braun, T.M. (2013) 'An improved quadratic inference function for parameter estimation in the analysis of correlated data', *Statistics in Medicine*, 32(19), pp. 3260–3273. Available at: https://doi.org/10.1002/sim.5715.

Wright, B. *et al.* (2016) 'Social stories™ to alleviate challenging behaviour and social difficulties exhibited by children with autism spectrum disorder in mainstream schools: Design of a manualised training toolkit and feasibility study for a cluster randomised controlled trial w', *Health Technology Assessment*, 20(6). Available at: https://doi.org/10.3310/hta20060.

Wykes, T. *et al.* (2018) 'Patient involvement in improving the evidence base on mental health inpatient care: the PERCEIVE programme', *Programme Grants for Applied Research*, 6(7), pp. 1–182. Available at: https://doi.org/10.3310/pgfar06070.

Wyld, L. *et al.* (2021) 'Bridging the age gap in breast cancer: Cluster randomized trial of two decision support interventions for older women with operable breast cancer on quality of life, survival, decision quality, and treatment choices', *British Journal of Surgery*, 108(5), pp. 499–510. Available at: https://doi.org/10.1093/bjs/znab005.

Yan, J. (2002) 'geepack: Yet Another Package for Generalized Estimating Equations', *R-News*, 2, pp. 12–14.

Yan, J. and Fine, J. (2004) 'Estimating equations for association structures', *Statistics in Medicine*, 23(6), pp. 859–874. Available at: https://doi.org/10.1002/sim.1650.

Yang, W. and Liao, S. (2017) 'A study of quadratic inference functions with alternative weighting matrices', *Communications in Statistics---Simulation and Computation-2.75pt*, 46(2), pp. 994–1007. Available at: https://doi.org/10.1080/03610918.2014.988255.

Yelland, L.N. *et al.* (2015) 'Analysis of Randomised Trials Including Multiple Births When Birth Size Is Informative', *Paediatric and Perinatal Epidemiology*, 29(6), pp. 567–575. Available at: https://doi.org/10.1111/ppe.12228.

Yelland, L.N., Salter, A.B. and Ryan, P. (2011) 'Relative risk estimation in cluster randomized trials: A comparison of generalized estimating equation methods', *International Journal of Biostatistics*, 7(1). Available at: https://doi.org/10.2202/1557-4679.1323.

Young, M.L. *et al.* (2007) 'Comparison of subject-specific and population averaged models for count data from cluster-unit intervention trials', *Statistical Methods in Medical Research*, 16(2), pp. 167–184. Available at: https://doi.org/10.1177/0962280206071931.

Yu, H., Li, F. and Turner, E.L. (2020) 'An evaluation of quadratic inference functions for estimating intervention effects in cluster randomized trials', *Contemporary Clinical Trials Communications*, 19, p. 100605. Available at: https://doi.org/10.1016/j.conctc.2020.100605.

Zeger, L. and Liang, S. (1986) 'Longitudinal Data Analysis for Discrete and Continuous Outcomes Author (s): Scott L . Zeger and Kung-Yee Liang Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2531248', *Biometrics*, 42(1), pp. 121–130.

Zhang, P. *et al.* (2008) 'Efficient estimation for patient-specific rates of disease progression using nonnormal linear mixed models', *Biometrics*, 64(1), pp. 29–38. Available at: https://doi.org/10.1111/j.1541-0420.2007.00824.x.

Zhang, X. (2015) *A Tutorial on Restricted Maximum Likelihood Estimation in Linear Regression and Linear Mixed-Effects Model*. Available at: http://proofwiki.org/wiki/Hermitian_Matrix_has_Real_Eigenvalues.

Zhang, Y. *et al.* (2023) 'GEEMAEE: A SAS macro for the analysis of correlated outcomes based on GEE and finite-sample adjustments with application to cluster randomized trials', *Computer Methods and Programs in Biomedicine*, 230. Available at: https://doi.org/10.1016/j.cmpb.2023.107362.

Ziegler, A. *et al.* (2000) 'Familial associations of lipid profiles: a generalized estimating equations approach', *Statistics in Medicine*, 19(24), pp. 3345–3357. Available at: https://doi.org/10.1002/1097-0258(20001230)19:24<3345::AID-SIM829>3.0.CO;2-5.

Ziegler, A. (2011) *Generalized estimating equations*, *Lecture notes in statistics*. New York: Springer.