The University Of Sheffield.

Design and analysis of trials evaluating proportionate interventions and trials with intervention induced clustering

By: Jane Candlish

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

University of Sheffield

Faculty of Medicine, Dentistry and Health

School of Health and Related Research (ScHARR)

Supervised by: Dr Dawn Teare and Dr Judith Cohen

March 3, 2019

Acknowledgements

I would firstly like to thank my supervisors, Dawn Teare and Judith Cohen for their invaluable support and input into the direction of this research. They have helped me develop as a researcher and encouraged me to take on new opportunities throughout my PhD.

This PhD was made possible by the University of Sheffield Harry Worthington Scholarship.

Thank you to members of the ScHARR Medical Statistics Group for their helpful input into this work and stimulating discussions. In particular, Munyaradzi Dimairo, Laura Flight, Laura Mandefield, and Stephen Walters who provided external support and advice for the work in chapter 4. Thank you to Steven Julious and Jeremy Dawson for the teaching opportunities and to my colleagues at the University's Mathematics and Statistics Help (MASH) centre for the experiences provided there. I also gratefully acknowledge friends and family who offered their services for proof reading: Rebecca Simpson, Laura Flight, Daniel Hartmann and Helen Quirk.

I would like to acknowledge researchers who provided external advice which enhanced this research. The E-SEE trial team provided helpful feedback on my systematic review and Tracey Bywater input into the search strategy for chapter 3. I am also grateful to the responses I received from researchers I contacted about their trials or methodology.

The last three years have been made so much more enjoyable because of friends both within ScHARR and outside. To the friends with whom I have shared this PhD experience, thank you for the countless lunch breaks, write clubs, and many sporting and social activities. To my friends outside academia, thank you for providing me with distraction from my studies but also reminding me to appreciate this time as a postgraduate student. Thank you to my parents Malcolm and Jackie for giving me an excellent start in life and always encouraging my studies.

Lastly, to my partner Daniel Hartmann thank you for your constant support, understanding, and keeping me calm throughout with many adventures.

Research achievements

Publications

Publication arising directly from work for this PhD thesis:

Candlish J, Teare M D, Dimairo M, Flight L, Mandefield L, Walters S J. (2018) Appropriate statistical methods for analysing partially nested randomised controlled trials with continuous outcomes: a simulation study. BMC Medical Research Methodology, 18: 105.

Candlish J, Teare M D, Cohen, J, Bywater TJ. Statistical design and analysis in trials of proportionate interventions: a systematic review. Trials, (In press).

Publications during my PhD arising from both my previous and continued collaboration with the University of Manchester and collaboration with peers at the University of Sheffield:

Candlish J, Pate A, Sperrin M, & Staa T. (2017). Evaluation of biases present in the cohort multiple randomised controlled trial design: a simulation study. BMC medical research methodology, 17: 17.

Hoo Z H, **Candlish J**, & Teare D. (2017). What is an ROC curve? Emergency Medicine Journal, 34: 357-359.

Pate A, **Candlish J**, Sperrin M, & Van Staa T P. (2016). Cohort Multiple Randomised Controlled Trials (cmRCT) design: efficient but biased? A simulation study to evaluate the feasibility of the Cluster cmRCT design. BMC Medical Research Methodology, 16: 109.

Sperrin, M, **Candlish J**, Badrick, E, Renehan, A, & Buchan, I. (2016). Collider bias is only a partial explanation for the obesity paradox. Epidemiology, 27(4): 525-530.

Conference presentations

Oral presentations

Candlish J. Analysis methods for individually randomised trials with partial clustering of outcomes. Society for Clinical Trials, Portland, USA 2018.

Candlish J. Statistical design and analysis of proportionate intervention trials: a systematic review. International Society for Clinical Biostatistics, Birmingham 2016.

Candlish J. Challenges in undertaking methodological systematic reviews. International Society for Clinical Biostatistics Students' Day, Birmingham 2016.

Poster presentations

Candlish J. Analysis methods for partially nested randomised controlled trials. International Society for Clinical Biostatistics, Vigo Spain 2017.

Candlish J. Methods to analyse partially nested randomised controlled trials. Society for Clinical Trials, Liverpool 2017. *Finalist for the SCT Best Poster Prize.*

Other research contributions

Supervised a Sheffield Undergraduate Research Experience student (Kirsten Thomas) project titled 'Why do we randomise in clinical trials? Explaining the benefits and requirements to improve recruitment' during summer 2017.

Abstract

Introduction: Individually randomised controlled trials (iRCTs) of complex interventions commonly induce clustered outcomes in the intervention arm only, termed partially nested trials (pnRCTs). In addition, iRCTs are increasingly used to evaluate interventions delivered proportionate to individual need. This can result in only some of the intervention arm having clustered outcomes due to post randomisation allocation to clusters, termed within-arm pnRCTs.

Research question: What elements need to be considered in the design, analysis and reporting of complex intervention trials with continuous outcomes, with a particular focus on proportionate interventions and intervention induced clustering in one trial arm?

Methods: Firstly, a systematic review of trials of proportionate interventions was performed. Simulation of pnRCTs and within-arm pnRCTs were used to investigate appropriate analysis methods. Sample size formulae for such RCTs were identified and summarised. Finally, a review of publicly funded iRCTs with clustering was undertaken.

Results: Proportionate interventions commonly induced within-arm partial nesting. Appropriate analysis methods were identified and demonstrated for pnRCTs, although with few clusters, small cluster sizes, and small intracluster correlation coefficient (ICC) there was no optimal method. Accounting for non-random clustering in within-arm pnRCTs was not possible; however, under realistic scenarios ignoring clustering can provide valid statistical inference. Sample size formulae for pnRCTs require an ICC estimate. From 15 publicly funded iRCTs the median healthcare provider ICC was 0.009. To improve transparency an additional Consolidated Standards of Reporting Trials-nonpharmacologic treatments item related to reporting ICC is suggested.

Conclusions: This thesis demonstrates the extent of clustering in both proportionate intervention trials and publicly funded iRCTs. Appropriate analysis methods are demonstrated for pnRCTs. For within-arm pnRCTs it is typically recommended to ignore the clustering. Sam-

ple size methods are summarised and empirical ICCs provided. This work provides practical guidance for design, analysis and reporting for continuous outcomes in RCTs with intervention induced clustering in one trial arm.

Contents

Abbreviations

ANOVA	Analysis of Variance
CI	Confidence Interval
CONSORT	Consolidated Standards of Reporting Trials
cRCT	Cluster Randomised Controlled Trial
CReDECI2	Criteria for Reporting the Development and Evaluation of Complex Interventions
EMA	European Medicines Agency
DT	Dawn Teare
GEE	Generalised Estimating Equations
HTA	Health Technology Assesment
ICC	Intracluster Correlation Coefficient
IQR	Interquartile Range
iRCT	Individually Randomised Controlled Trial
ISRCTN	International Standardised Randomised Controlled Trial Number
ITT	Intention To Treat
MLE	Maximum Likelihood Estimation
MRC	Medical Research Council
MSE	Mean Square Error
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NPT	Non-pharmacological treatments
nRCT	Nested Randomised Controlled Trial
pnRCT	Partially Nested Randomised Controlled Trial
PP	Per Protocol
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QoL	Quality of Life
RCT	Randomised Controlled Trial
REML	Restricted Maximum Likelihood Estimation
ScHARR	School of Health and Related Research
SMART	Sequential Multiple Assignment Randomised Trial
TIDieR	Template for Intervention Description and Replication

Notations

Subscripts

i	Represents individuals $i = 1, 2, \ldots, N$
j	Represents clusters $j = 1, 2, \ldots, k$

Hypothesis testing

H_0, H_A	Null and alternative hypothesis
α	Type I error rate
β	Type II error rate

Statistics, parameters and outcomes

δ	Minimally clinical important difference
δ_s	Standardised effect
y	Primary continuous outcome
σ^2	Total variance in outcome
s^2	Sample variance in outcome
$beta_0$	Intercept
θ	Overall intervention effect
θ_1	Intervention effect of stage one
θ_2	Intervention effect of stage two
t	Indicator for intervention arm
t_1	Indicator for stage one
t_2	Indicator for stage two
σ^2_ϵ	Individual level variance
σ^2_r	Individual level variance
π_2	Proportion who receive stage two intervention

Clusters

m	Cluster size
$\bar{m}$	Mean cluster size
c/k	Number of clusters
cv	Coefficient of variation in cluster size

Measures of between-cluster variability

ρ	Intracluster correlation coefficient
σ^2_u	Between-cluster variance

Distribution parameters

Z_x	The x^{th}% point of the standard Normal distribution
t_x	The x^{th}% point of the t distribution

List of Figures

List of Tables

Chapter 1

Introduction

In recent decades there has been an increasing acceptance and emphasis on evidence based healthcare, the aim of delivering and making informed healthcare decisions through the use of sound research and evidence evaluation [1]. Resources for any healthcare system are limited, consequently rational policies and decision making will ideally use these limited resources to provide and implement interventions that have been proven to be both effective and cost-effective through well designed research and evaluation.

Randomised controlled trials (RCTs) play a central role in building this evidence as they are generally considered the gold standard for evaluating the effectiveness of interventions in healthcare. The term intervention defines an activity undertaken with the purpose of improving, assessing, promoting or modifying health or health conditions. RCTs have been widely used to evaluate the effectiveness of pharmacological interventions. There has also been a move in public health, primary care and educational research towards evaluating the effectiveness of non-pharmacological complex interventions. Complex interventions are conventionally those that are made up of a variety of interacting components [2]. For example, evaluating weight management groups for obesity, surgical interventions, and delivery methods of a psychotherapy intervention.

Individually randomised controlled trials (iRCTs) of complex interventions commonly induce clustered outcomes in the intervention arm only, termed partially nested trials (pnRCTs). RCTs are also being used to evaluate the effectiveness of complex interventions delivered proportionate to individual need. This can result in only some of the intervention arm having clustered outcomes due to post randomisation allocation to clusters, termed within-arm pnRCTs.

Appropriate statistical design and analysis of RCTs are required to enable valid and useful

conclusions to be made. The design and analysis of iRCTs evaluating complex interventions are improving; however, there is need for improved understanding. The overarching research question of this thesis is: what elements need to be considered in the design, analysis and reporting of complex intervention trials with continuous outcomes, with a particular focus on proportionate interventions and intervention induced clustering in one trial arm.

1.1 Chapter aims

This chapter aims to introduce a motivating example, the E-SEE trial design, providing a brief overview of the trial in order to present the various complexities that can arise. The thesis research question and research aims are then introduced. The chapter will conclude with an overview of the thesis structure.

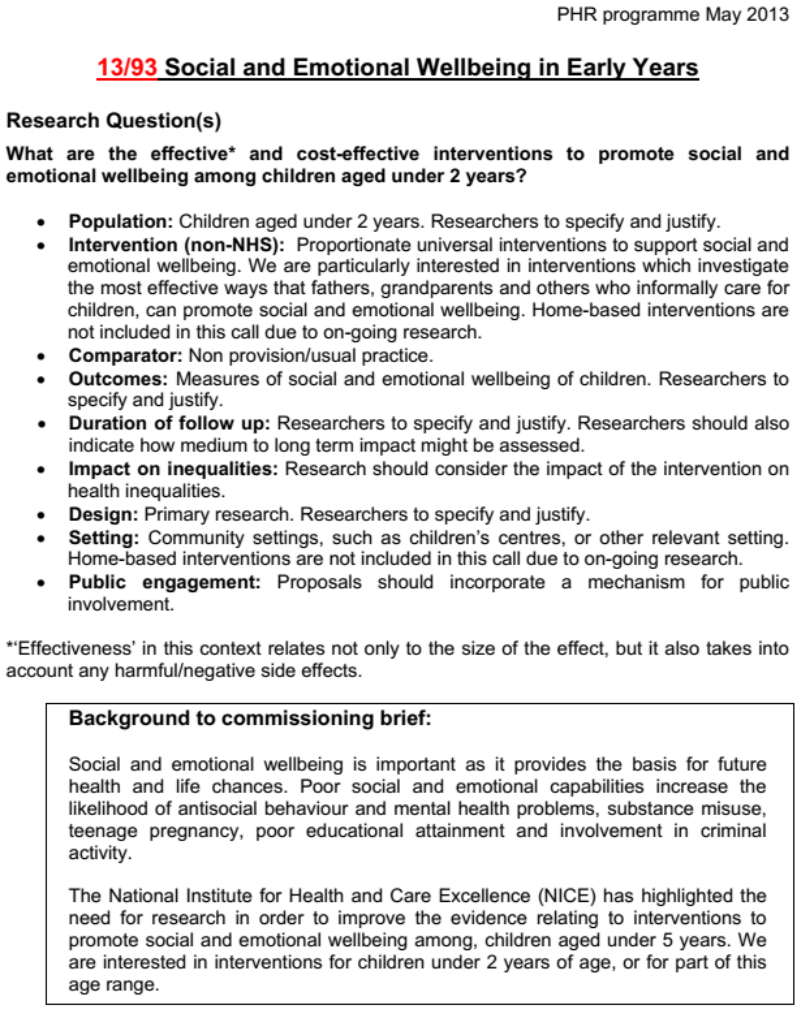
1.2 Motivating example: the E-SEE trial

Much of the motivation for this thesis originated from discussion regarding the design and analysis of the multi-centre E-SEE (Enhancing Social-Emotional health and wellbeing in the Early years) trial. The E-SEE trial aims to evaluate the effectiveness of a community based implementation of the Incredible Years (IY) group-based parenting programme, delivered proportionate to need [3]. The trial is funded through the National Institute for Health Research (NIHR) Public Health Research programme call asking the research question "What are the effective and cost-effective interventions to promote social and emotional wellbeing among children aged under 2 years?". Figure 1.1 shows the NIHR call [4], focussing on proportionate universal interventions [5] considering the impact on health inequalities.

The E-SEE trial aims to deliver IY as a proportionate universal intervention based on assessment of need. The research question is: "Are the IY-Infant (IY-I) and IY-Toddler (IY-T) programmes, when delivered in a dose proportionate to need, and when compared to services as usual, effective and cost-effective in enhancing child social and emotional well-being at 20 months of age?". At the time of writing this thesis the E-SEE trial is still ongoing and this PhD has been running in parallel with the trial.

Participants were randomised at the individual level to one of two trial arms: control arm (care as usual) or IY intervention arm. The IY intervention arm consists of a proportionate delivery of

Figure 1.1: NIHR PHR programme May 2013: 13/93 Social and Emotional Wellbeing in Early Years



PHR programme May 2013

13/93 Social and Emotional Wellbeing in Early Years

Research Question(s)

What are the effective* and cost-effective interventions to promote social and emotional wellbeing among children aged under 2 years?

- **Population:** Children aged under 2 years. Researchers to specify and justify.
- **Intervention (non-NHS):** Proportionate universal interventions to support social and emotional wellbeing. We are particularly interested in interventions which investigate the most effective ways that fathers, grandparents and others who informally care for children, can promote social and emotional wellbeing. Home-based interventions are not included in this call due to on-going research.
- **Comparator:** Non provision/usual practice.
- **Outcomes:** Measures of social and emotional wellbeing of children. Researchers to specify and justify.
- **Duration of follow up:** Researchers to specify and justify. Researchers should also indicate how medium to long term impact might be assessed.
- **Impact on inequalities:** Research should consider the impact of the intervention on health inequalities.
- **Design:** Primary research. Researchers to specify and justify.
- **Setting:** Community settings, such as children's centres, or other relevant setting. Home-based interventions are not included in this call due to on-going research.
- **Public engagement:** Proposals should incorporate a mechanism for public involvement.

*'Effectiveness' in this context relates not only to the size of the effect, but it also takes into account any harmful/negative side effects.

Background to commissioning brief:

Social and emotional wellbeing is important as it provides the basis for future health and life chances. Poor social and emotional capabilities increase the likelihood of antisocial behaviour and mental health problems, substance misuse, teenage pregnancy, poor educational attainment and involvement in criminal activity.

The National Institute for Health and Care Excellence (NICE) has highlighted the need for research in order to improve the evidence relating to interventions to promote social and emotional wellbeing among, children aged under 5 years. We are interested in interventions for children under 2 years of age, or for part of this age range.

three levels of the IY intervention program. Within the IY intervention arm IY-Baby is issued at the universal level to all. At two months, parents are offered the group based IY-Infant if they are showing levels of stress or mild depression (PHQ-9 score of $\geq$5), comprised of 10 weekly parenting group sessions (2 -9 months). Parents are assessed again at nine months and offered IY-Toddler if they show signs of stress or mild depression, comprised of weekly group parenting sessions (9-18 months). The E-SEE trial is an individually randomised controlled trial, with treatment induced clustering in one arm of the trial. Group sessions are delivered to groups of parents in local centres by trained facilitators, delivered in a number of authorities, with different facilitators at local centres. The clustering only occurs at certain stages of the intervention (clustering by group based intervention at IY-Infant and IY-Toddler). Clustering occurs in only one arm of the trial (partial nesting) and for only some individuals in that trial arm (within-arm partial nesting).

The original trial design aimed to evaluate the overall effectiveness of all three levels of the IY programme and the individual effectiveness of each level being investigated. The E-SEE design presents particular challenges for both the design and analysis of the trial and led to the initial idea of exploring this in depth as a PhD. Further exploration of the wider topic informed the generation and refining of the research objectives.

1.3 Research question

As highlighted above the focus of this thesis is to answer the research question: what elements need to be considered in the design, analysis and reporting of complex intervention trials with continuous outcomes, with a particular focus on proportionate interventions and intervention induced clustering in one trial arm?

1.4 Research aims

The specific thesis aims to address the research question are:

1. To review current practice of how randomised trials of proportionate interventions are designed and analysed and the extent of clustering in such trials.
2. To evaluate commonly used analysis methods for partially nested randomised trials and within-arm partially nested randomised trials to establish which methods are most appropriate and why.
3. To identify and collate a comprehensive summary and resource for sample size methods available for partially nested randomised trials.
4. To determine the extent and quality of reporting of clustering in iRCTs with intervention induced clustering and provide empirical estimates of ICCs.

1.5 Thesis Structure

In this chapter the E-SEE trial has been presented as an example of the various complexities that arise from randomised trials of complex interventions, focusing on the two key themes of proportionate interventions and treatment induced clustering. The research objectives of this

thesis are introduced. A description of the thesis structure and what each chapter includes follows.

Chapter 2 provides a conceptual framework and background for this thesis reviewing the public health and statistical literatures. Key terminology and concepts are defined, including complex intervention, proportionate universalism and proportionate interventions. Randomised controlled trial (RCT) designs are introduced including individual, cluster, partially nested, and within-arm partially-nested RCTs. Implications for the precision of treatment effects are given for statistical analyses and sample size, describing how the statistical analysis and reporting of a trial needs to correspond to the features of the design. Aspects of clustering including, types of clustering, quantifying clustering, and defining and calculating the intracluster-correlation coefficient (ICC) are introduced.

The framework is used in chapter 3 to structure a systematic review to address the aim of reviewing how randomised trials of proportionate interventions are designed and analysed in practice and the extent of clustering in such trials. The systematic review identifies trials of proportionate interventions, categorised as either stepped-care or optimal intervention strategy studies. The search strategy and eligibility criteria for selecting the trials are provided. Data is extracted and summarised on the therapeutic area, stages and decision rules of the intervention, statistical analysis used, whether different intervention stages were analysed and the frequency and treatment induced clustering. The most common therapeutic area is mental health and treatment induced clustering is present in the majority of trials identified.

Chapters 4 and 5 address the aim of evaluating appropriate statistical methods for analysing trials with nested outcomes in one trial arm. Chapter 4 summarises analysis models commonly employed for pnRCTs and evaluates these through a simulation study. The literature to date is considered, the issue of degrees of freedom, and the lack of clear guidance highlighted. The simulation study emulates pnRCTs over various scenarios of cluster size, number of clusters, ICC and both homoscedastic and heteroscedastic individual variances. Six analysis methods are compared in terms of bias, Type I error, ICC estimation, and power. Recommendations regarding the appropriate analysis method to use are presented in relation to the design of the study. Model fitting code is presented for R, Stata and SAS. The simulation code is given in an appendix.

Chapter 5 extends the case of pnRCTs to those similar to E-SEE and seen in the systematic review in chapter 3. Randomised trials which introduce within-arm partial nesting are further

introduced and potentially appropriate statistical analysis methods are presented and reviewed using a simulation study. Again, the simulation study emulates within-arm pnRCTs over various scenarios of cluster size, and number of clusters. Recommendations regarding the appropriate analysis method to use are presented in relation to the design of the study, the expected proportion of participants with clustered outcomes and expected ICC. The simulation study only considers one clustered intervention stage, trials which evaluate interventions with numerous clustered intervention stages are discussed and the limitations this may have on accounting for clustering in the analysis.

Chapter 6 reviews and collates sample size formulae and relevant statistical software for the design of trials with clustering in one arm. The sample size formulae are outlined, building up from individually randomised trials to cluster trials and finally to pnRCTs. The issue of within-arm clustering is discussed in the context of sample size calculations and guidance provided in relation to the simulation results in chapter 5. Practicalities of obtaining the values required for these sample size formulas are discussed.

Chapter 7 investigates the extent, reporting, and evidence of clustering and ICCs in individually randomised trials published in the Health Technology Assessment (HTA) Journal, as a representative source of publicly funded and reported trials in the UK. Trials with potential for treatment induced clustering are selected. Extent of clustering is reported alongside how it has been recognised in the design, analysis, and reporting. Reporting is assessed with adherence to key items of the CONSORT-non-pharmacological treatments checklist related to clustering. Evidence of ICCs used in sample size and empirical estimates of ICCs from the trials are presented, with the aim of raising awareness and informing current trial design and required sample size calculations.

A discussion of the thesis is found in chapter 8. This is structured according to the aims, placing the thesis in the context of the existing literature. Areas for further work are identified.

Chapter 2

Background

2.1 Introduction

RCTs of complex interventions raise numerous challenges in the design and analysis stages. The E-SEE trial was presented in chapter 1 as a means to introduce the key themes of this thesis: proportionate interventions and intervention induced clustering. RCTs are essential for evaluating the effectiveness of health interventions, however, clustering and proportionate delivery introduce additional complexity to the analysis of the trial.

This chapter provides background for the understanding and definition of complex interventions in health, introducing the motivation for proportionate universalism, proportionate interventions and focusing on the complexities that arise from clustered of outcome data and proportionate interventions.

2.2 Chapter aims

The chapter introduces key terminology and concepts, including complex intervention, proportionate universalism and proportionate interventions. Various types of RCT designs are introduced. Implications for the precision of treatment effects are given for statistical analyses and sample size, describing how the statistical analysis and reporting of a trial needs to correspond to the features of the design. Aspects of clustering including, types of clustering, measuring clustering, and defining and calculating the intracluster correlation coefficient (ICC) are introduced.

2.3 Complex interventions

Many interventions aimed at improving healthcare can be defined as complex interventions, in that they are made up of multiple interacting components [6]. Trials of such interventions often present their own specific challenges [7]. Complex interventions are defined here based on the Medical Research Council (MRC) guidance [6] and the more recent work by Kühne et al. [8]. The role the researcher takes in defining an intervention simple or complex is also considered [9].

Complex interventions are defined as those that are made up of several interacting components, they are generally non-pharmacological interventions. The complexity may arise through [6]:

- Number of components and interactions between components;
- Number of different groups/organisational levels targeted by the intervention;
- Variability of outcomes;
- The degree of flexibility/tailoring of intervention permitted.

An intervention component is any aspect of the intervention that could potentially have an effect on the efficacy of the intervention. A component can be part of the intervention content, features that promote adherence or fidelity of delivery. The multiple components may target different levels of the social-ecological model, designed to affect change at these different levels [10]. For instance, an intervention can be delivered to the individual patient such as the OCTET trial of low intensity cognitive behavioural therapy and guided self-help for obsessive compulsive disorder [11]. They can be delivered at the community level such as the PLEASANT trial of a letter sent from General Practice (GP) aimed at reducing childhood asthma exacerbation levels [12]. Some interventions are delivered to the healthcare professional. The intervention can also be delivered at the family level, for example, the 'Families for Health' trial of family based childhood obesity treatment delivered to the whole family [13]. To add to the complexity, a combination of these different levels can also occur within one trial. The multiple components can affect the efficacy, effectiveness, and cost-effectiveness of the intervention.

Evaluation of complex interventions can raise difficulties due to the inherent features of the intervention itself. Datta and Petticrew [14, p.16] state that "the literature on complex interventions is thick with descriptions of complex, challenging interventions, but thin on practical advice on

how these should be dealt with". RCTs of complex interventions are sometimes criticised as being 'black box'. It can be challenging to evaluate and understand both the effectiveness of an intervention and how it works without examining underlying processes. This is something the MRCs guidance [6, 15] on developing and evaluating complex interventions emphasises, the importance of evaluating both the effectiveness of an intervention and how it works. These guidelines breakdown the key elements of complex intervention framework into: development; feasibility and piloting; evaluation; and implementation. Guidelines and frameworks for the reporting of trials of complex interventions also exist [16, 17].

Petticrew [9] argues that there is not an easy divide between simple and complex interventions, they are on a continuum and simple questions can be asked of complex interventions and vice versa. These are considered pragmatic perspectives used by researchers and it is actually the researchers perspective and research question that defines the simple and complex explanations of an intervention. With this viewpoint a staged proportionate intervention (each stage a different component), such as the IY intervention in the E-SEE trial, could be both simple and complex dependent upon the analysis. A complex analysis may want to examine the effects of how and whether the component parts work alone and together as well as the synergies between them. A simpler analysis may focus on whether the intervention works as a whole package. It can be difficult or impossible to identify whether it is a particular component of a complex intervention that is causing the effect. In addition, some components may induce clustering in the outcome due to the nature of the delivery of the intervention, for example, therapist led treatment or group based treatment (this will be expanded upon in section 2.7). This is where the importance of study design is key, to design a trial to answer the key research questions and to understand how and for whom an intervention is working.

Multi-component interventions can be costly and resource intensive. Patient tailored or stratified healthcare is becoming more common, and this principle can be applied to complex interventions. The different components may not all be required or appropriate to individual need. One approach to address this is the use of proportionate or adaptive interventions. Proportionate or adaptive interventions are multi-component interventions given in a staged manner. In such an approach participants receive a low intensity intervention, then individuals who do not respond receive a more intense intervention component. Before explaining proportionate interventions, the following section will discuss the origins and motivations of proportionate universalism which was breifly introduced in chapter 1 through the E-SEE trial and the NIHR call [4].

2.4 Proportionate universalism

There are limits on healthcare funds, hence, proportionate universalism [5] is focussed on providing care for those who need it when they need it with the aim of reducing health inequalities. The term proportion universalism originated from the Strategic Review of Health Inequalities in England Marmot Review [5]. There is a strong case for local authorities to invest in tackling health inequalities, benefits being both social and economic. Marmot et al. [5] writes:

> "Focusing solely on the most disadvantaged will not reduce health inequalities sufficiently. To reduce the steepness of the social gradient in health, actions must be universal, but with a scale and intensity that is proportionate to the level of disadvantage. We call this proportionate universalism." [5, p.15].

Marmot et al. [5] argued that universal healthcare has done little to reduce health inequalities. Suggesting that consideration should be given to exploring different ways complex interventions could be administered in order to reach those most in need, improve health, and reduce inequalities. In 2012 the National Institute for Health and Care Excellence (NICE) stated that only 0.4% of public health research had been focussed on interventions aimed to improve health inequalities [18]. Since the Marmot Review [5] there has been an increased interest in the importance the social gradient in health can have throughout society, it does not just affect those at the very low socio-economic statuses but is graded throughout the population [19].

Many standard interventions can actually increase health inequalities. The affluent often access interventions with higher frequency than the more deprived, thus, though the interventions may improve the health of those who receive the interventions they are actually leading to an increase in health inequalities due to access to services and interventions [5]. The universal level of a proportionate intervention is aimed at bringing the general health of the whole population of interest up and then targeting additional stages of the proportionate intervention to those in need.

Proportionate universalism may be interpreted in varying manners, from the development of direct health interventions for those in most need to dose-response interventions with tailoring of the intensity of interventions proportionate to need. A recent framework for the application of proportionate universalism [20] argued that interventions in a proportionate universalist set-up need to be applied to some degree to the whole population (the universal level), rather than only to those most disadvantaged. Targeting specific groups risks labelling these groups and

thus the associated stigmas that come with labelling. However, proportionate universalism will incorporate a level of selectivism based on an individuals needs (the proportionate level) and in turn this will require some targeting. Few systems or policies are truly universal; some argue that the need for judgements restricts universalism through the decisions of who gets what service [20].

The principle and terminology of proportionate universalism has been taken up largely in the fields addressing parent [21, 22] and child health [19, 23–26], including mental health [26], all focal in the Marmot Review [5]. The literature varies from narrative articles, qualitative studies, to quantitative analyses of observational data identifying health inequalities.

A key focus of studies which mention proportionate universalism has been on promoting the best start for a child in the early years, significant in developmental processes that shape the rest of life [19]. Unequal access to services is central to the inequalities in child health and development, these inequalities increase as a child grows up [19]. The studies pointed to a lack of evidence on how to promote child health and development in an equitable way even though there is a large amount of evidence and discourse of its importance [19, 27].

A number of reviews exist aiming to identify interventions that promote health equality for children [23, 26]. The scoping review by Welsh et al. [26] focussed on interventions to promote mental health well-being and reduce inequalities in children in high-income countries. Out of more than 1000 potentially relevant interventions, none were understood to follow a proportionate universalism framework, they were either targeted or universal. Some of the universal programmes somewhat emphasised inequalities by benefiting the advantaged children more, whereas others found stronger effect sizes for the most disadvantaged. This emphasizes the importance of a proportionate universalism framework, providing appropriate healthcare for all, not merely those at the top or bottom of the socio-economic scale. A systematic review by Morrison et al. [23] identified 23 universal, targeted and proportionate parenting interventions in European countries (1999-2013) which aimed to reduce inequalities in child health and development. However, only two of the 23 interventions followed a proportionate universalism principle, whilst the rest were targeted at those at higher risk or who were already showing signs of a developmental problem. These reviews have identified a gap between the recommendations of proportionate universalism and the available interventions.

2.5 Proportionate interventions

One consequence of proportionate universalism as a motivation for intervention delivery is the development and evaluation of proportionate complex interventions in RCTs. Proportionate interventions comprise of various stages or different component parts. For example, everyone receives the first stage of the intervention and their response to this stage determines if they progress onto the next stage; it is delivered proportionate to the need of the individual. The E-SEE trial aims to evaluate a proportionate delivery model of the IY parenting programme [3]. Proportionate universalism has two levels: the universal and the proportionate.

Proportionate universalism is a relatively new term, aiming not to increase inequalities by implementing health interventions. However, the principle of adapting interventions proportionate to need has been present for much longer. Proportionate universal interventions are closely linked to other terminology, including:

- dynamic treatment regimens;
- adaptive treatment regimens;
- adaptive interventions;
- adaptive treatment strategies;
- stepped-care interventions;
- multi-level interventions.

The following defines what is meant by a proportionate intervention, using the literature on adaptive interventions to motivate this work. A proportionate intervention or adaptive intervention entails an individualised approach with numerous treatment sequences dependent upon individual need as treatment is adapted dependent on need (for example, continue, augment, switch, step-down) [28]. This may mimic how decisions are made in practice, thus guiding the intervention process. These interventions have two main components: (a) individualised treatment based on participants needs and (b) a time varying intervention that adapts in response to participants changing need over time [29]. There are often a variety of questions to answer when developing a proportionate intervention, including: when is the optimal time to assess responsiveness to treatment?; should non-responsive individuals be offered an augmented treatment or an alternative treatment?; should the initial period of treatment be based on an individuals

baseline characteristics?; could the decision to augment treatment be individualised based on other outcomes?.

A proportionate intervention includes: (i) decision stages and at each decision stage: (ii) intervention options; (iii) tailoring variable(s); (iv) a decision rule; (v) outcomes. The decision rule utilises the tailoring variable(s) to decide upon the staged intervention an individual receives. The tailoring variable is patient information. The decision rule cut-off should ideally be based on previous research. The intervention may be individualised based on decision rules using dynamic information about the individual that is likely to change due to intervention such as response or adherence either singularly (only one decision) or sequentially (multiple decision stages).

Proportionate interventions are useful as individuals who require a step-up/down or switch in treatment receive that and for those that are responding to the current treatment there are no increased burdens such as side effects or invested time. Increased burden can lead to non-adherence, which may in turn reduce positive intervention effects. In addition, all interventions incur costs and as healthcare resources are limited, the ability to reduce costs of receiving unnecessary further interventions whilst treating those in greatest need is an important goal. Proportionate interventions may be particularly relevant for interventions with heterogeneous responses. For example, interventions developed for mental health disorders often produce heterogeneous responses due to the within person (over time) and between person differential responses to intervention [30].

Developing and evaluating proportionate interventions in trials raises specific issues in the design and analysis. Teams developing such complex intervention packages may want to evaluate the effectiveness of the individual stages or the incremental benefits of each stage in addition to the overall intervention. This presents fresh challenges for the design and statistical analysis of such interventions. In general, trials randomise individuals or clusters to a whole treatment pathway to assess effectiveness. However, a proportionate universalist design creates multiple treatment pathways, each dependent upon outcomes at the previous stage of treatment.

In an RCT, an intention-to-treat (ITT) analysis provides an estimate of the average effect size for those randomised to the intervention of interest. A per-protocol (PP) provides an estimate of the average effect size for those who adhere to the protocol fully. In a staged or proportionate intervention the intervention is delivered dependent upon the need of an individual, thus the estimates of an average effect size may not be as relevant as they are in more standardised interventions.

Further complexities also arise from the clustering of outcomes in one or more stages of proportionate interventions. For instance, in the E-SEE trial the first component of the IY intervention is delivered at the individual level and the second and third components of the intervention are delivered to groups of individuals. This will potentially induce treatment induced clustering from both the group dynamics and the role of the group facilitator. This clustering is induced by the nature of the intervention rather than randomisation to a cluster. Clustering in iRCTs will be discussed in more detail in sections 2.6 and 2.7.

2.6 Randomised controlled trials

An RCT is a controlled experiment designed to evaluate the effectiveness of one or more interventions to an appropriate comparator. The intervention may be: a drug or other form of medical intervention such as surgery or therapy; or a method of organising or delivering healthcare such as a new system or training for care providers. This section provides a background to some of the key types of RCTs used in health research and the corresponding requirements for the design and analysis of such trials. More detailed information on power calculations and sample sizes are given in chapter 6 and comparison of analysis methods given in chapters 4 and 5.

The main aim of an RCT is typically to obtain an unbiased and reliable estimate of the effectiveness of the intervention being tested. In a parallel two arm trial individuals or units known as clusters are randomised to either receive a control condition, often the standard of care, or an experimental intervention condition. By using randomisation, all known and unknown factors that could possibly confound the estimate of the intervention effect should be balanced on average between the two trial arms; this allows any difference between the two arms to be attributed to the intervention. It is then possible to make statistical inference about the causal effectiveness of an intervention by comparing the outcomes of the trial arms after a predetermined follow-up period.

Drug trials are classified by phase, with four main stages and well defined guidelines for each of phase I/II/III/IV purposes and scope. The same phases of development are not used for intervention trials. However, there has been an increasing use and acceptance that clinical treatments and public health decisions should be based on a comprehensive review of the evidence. This evidence should be based on rigorously conducted studies, which evaluate both benefits and ad-

verse events. The use of such evidence will ideally enable policy makers and healthcare services to allocate resources accordingly to interventions which have been proven to be both effective and cost-effective [31].

2.6.1 Superiority trials and hypothesis testing

A two-arm parallel RCT is commonly used to show one intervention is superior to another. Individuals are randomly allocated to one of two arms and a statistical test is undertaken to make inference about the intervention effect. This is referred to as a superiority trial.

Let's define a null hypothesis (H_0) and alternative hypothesis (H_A). In a two-arm superiority trial comparing intervention treatment $t = 1$ to control treatment $t = 0$ these are

- H_0: $\bar{y_0} = \bar{y_1}$, two treatments are the same
- H_A: $\bar{y_0} \neq \bar{y_1}$, two treatments are different

where $\bar{y}$ is the average outcome measure in the corresponding treatment arm.

An RCT only investigates a sample of the population; the outcome is an estimate of the true population outcome. A hypothesis test can be used to test the significance of the difference in outcome between the control and intervention group. A 0.05 cut-off level for statistical significance is typically chosen, however, this is arbitrary and confidence intervals should be reported alongside any p-value.

The use of a p-value cut-off results in a dichotomous decision, giving two possible errors, Type I and Type II error summarised in Table 2.1. Type I error occurs when we reject the null-hypothesis when it is true, a false positive (α = probability of Type I error) and is determined in advance of a study. Type II error occurs when we fail to reject the null-hypothesis when it is false, a false negative, and is dependent upon the sample size and the effect size of interest (β = probability of Type II error). We more commonly report the power of a study $(1 - \beta)$ to detect an effect size.

Table 2.1: Summary of Type I and Type II errors

	H_0 True	H_0 False
Do not reject H_0	Correct decision	Type II error
	$(1-\alpha)$	(β)
Reject H_0	Type I error	Correct decision
	(α)	$(1-\beta)$

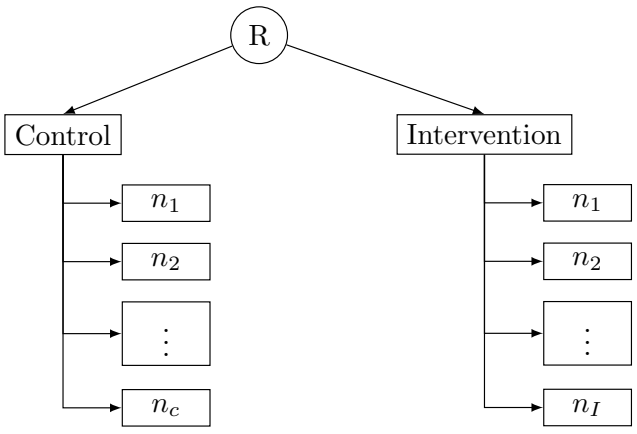
2.6.2 Continuous outcomes

This thesis considers continuous outcomes, these are commonly used in complex intervention and public health research for example through the use of outcomes such as body mass index (BMI) and Patient Health Questionnaire-9 (PHQ-9) and Quality of Life (QoL) outcomes such SF-36, EQ5D, PHQ-9. Continuous outcomes were the most common primary endpoint in a recent review of UK publicly funded trials (NIHR funded), 45.8% of trials published between 2006 and 2016 (49 of 107 RCTs published in the Health Technology Assessment journal) [32].

2.6.3 Individually randomised trial

RCTs can employ an individually randomised controlled trial (iRCT) design, where participants are individually randomised to receive one of the investigative treatments. Figure 2.1 presents a schematic of a two-arm parallel iRCT, they are commonly used in drug trials and other types of individual therapies. In iRCTs we often assume that the outcomes from participants are independent of one another and thus usual assumptions of independence of outcomes in statistical analysis are met. The corresponding sample size calculations for such trials can use standard calculations which assume independent outcomes [33].

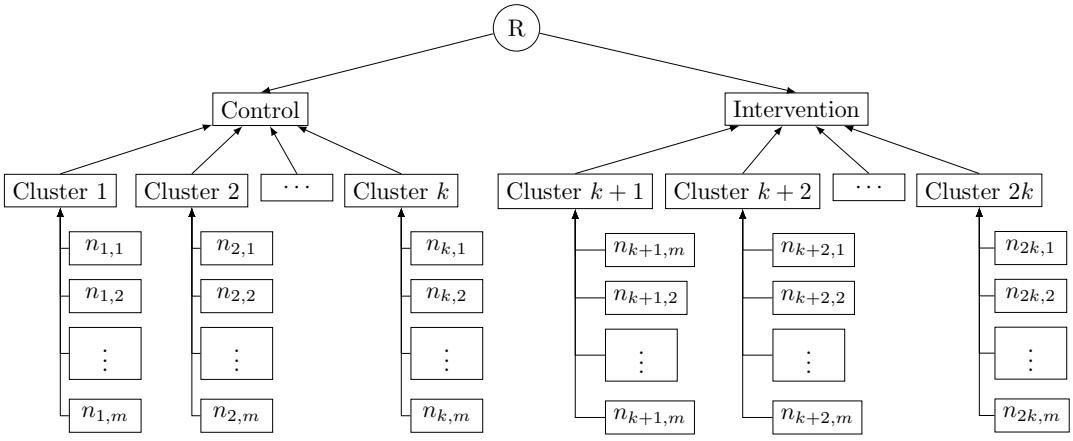
Figure 2.1: Schematic of a two-arm parallel individually randomised trial where R represents randomisation and there are n_c and n_I patients in the control and intervention arms, respectively.



2.6.4 Cluster randomised trial

Using an iRCT is not always appropriate or feasible. Cluster randomised trials (cRCTs) may be more suitable. Cluster randomisation occurs when an entire unit (for example GP practice, school or hospital) known as clusters, are randomised to interventions [31]. Figure 2.2 presents a schematic of a cluster trial with two arms, a control and an intervention arm. They are employed for a number of reasons. The intervention may need to be applied to whole communities or it is easier to administer it this way, for example, implementation of new procedures in a GP practice or hospital where the intervention applies to the whole unit. Cluster randomisation helps reduce the likelihood of contamination between intervention arms; it may not be realistic to offer the intervention to one individual without others at the same cluster (hospital or GP practice) being exposed to the intervention as well.

Figure 2.2: Schematic of a cluster trial where R represents randomisation (randomisation is at the cluster level), k clusters in each treatment arm, and m number of patients in each cluster.



Clustering is defined in general for an RCT to be when outcomes are grouped together based upon a common property, such as GP practice, school or hospital. Outcomes for subjects within the same cluster are expected to be more similar than those from different clusters.

The implications of clustering in cRCTs are widely acknowledged [34]. Cluster trials generally require larger sample sizes than individual trials designed to answer the same research question (if an iRCT is possible) and should account for the correlation of outcomes in the analysis method. Analysis methods are further discussed in section 2.8.2.

2.6.5 Individually randomised controlled trials with clustering

Where randomisation occurs at the individual level clustering may also occur, consequently, clustering of outcomes can be present in iRCTs. For instance, clustering of participants' out-

comes due to receiving treatment as part of a group-based parenting intervention [35], treatment in specialist clinics for the treatment of venous leg ulcers [36], or participants under the care of a surgeon for comparison for hemostasis in elective benign thyroid surgery [37]. The clusters in iRCTs are not necessarily the organisational unit, they are the care provider or intervention group which may play a role in the causal pathway of the intervention effect. We might expect a correlation of outcomes between individuals either in the same group or receiving treatment from the same care provider. This clustering can be caused by care provider characteristics such as level of experience, level of training, competence, or in a group trial through group dynamics [38, 39].

When designing and analysing iRCTs with clustering we need to consider implications of the potential lack of independence. Ignoring clustering in the analysis can lead to misleadingly precise results and consequently incorrect conclusions [40].

There is increasing acknowledgement of clustering present in iRCTs, with a growing awareness of the need to account for this clustering [41–46]. Consolidated Standards of Reporting Trials (CONSORT) statement consists of a minimum set of reccomendations for reporting randomised trials [16]. Extended CONSORT guidelines for the reporting of RCTs of nonpharmalogical interventions (CONSORT-NPT) have drawn attention to the need, when applicable, to address clustering by care provider and/or centre [17, 44]. However, studies still fail to account for potential treatment induced clustering and a recent CONSORT extension for social and psychological interventions (published July 2018) does not address the issue of potential clustering though it is common in psychological interventions [47].

2.6.6 Nested and partially nested randomised controlled trials

Treatment induced clustering in iRCTs has been termed a nested randomised controlled trial (nRCT) [48] and can occur in both arms of the trial as presented in Figure 2.3. In a similar vain, an increasingly applied design in healthcare and education research is a partially nested randomised controlled trial (pnRCT), where participants are individually randomised to trial arms and clustering of outcomes occurs in only one arm of the trial [48, 49] (sometimes termed partially clustered trials). A schematic of a pnRCT is presented in Figure 2.4. The STEPWISE trial is an example of a pnRCT, assessing a structured lifestyle education programme aimed at supporting weight loss for adults with schizophrenia and first episode psychosis in a community mental health setting. Individuals were randomised to either an intervention arm of group-

Figure 2.3: Schematic of a nested trial where R represents randomisation, there are k clusters in each treatment arm, and m number of patients in each cluster.

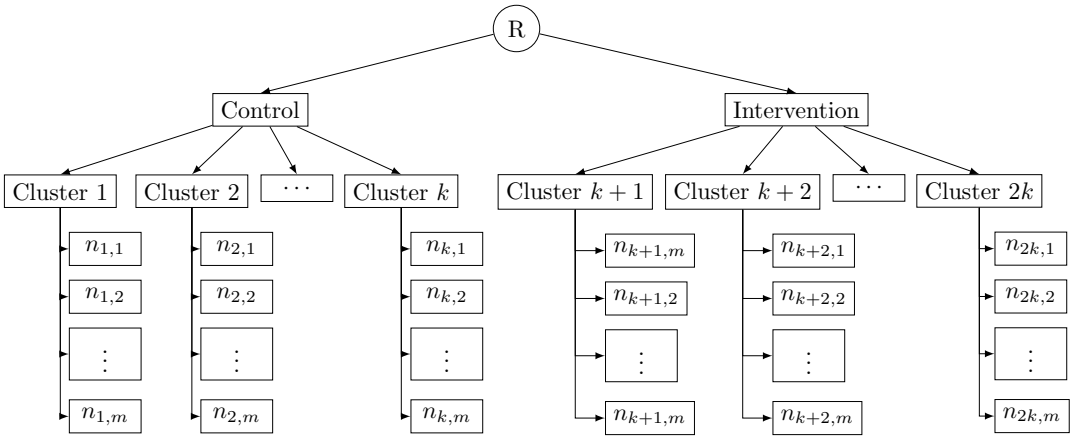
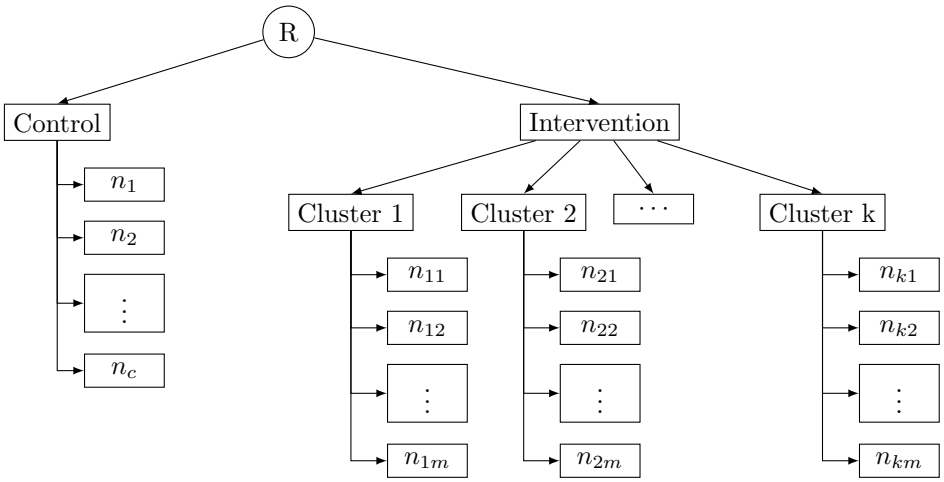


Figure 2.4: Schematic of a partially nested trial where R represents randomisation, there are n_c patients in the control arm, k clusters in the intervention arm, and m number of patients in each cluster.



based lifestyle education sessions or a control arm receiving usual care at the individual level [50]. In iRCTs with treatment induced clustering where clustering does not occur for all the trial participants, analysis must be at the individual level.

The cluster sizes in nRCTs and pnRCTs are typically smaller than those in cRCTs. Cluster size in a cRCT is often externally fixed, such as the number of individuals attending a clinic or number of individuals in a GP practice. In contrast to cRCTs, there may be more researcher control over cluster sizes in pnRCTs such as the number of patients treated by a care provider or the group sizes. Although group sizes are typically decided by intervention developers; they need to form a viable group for the intervention delivery and develop group dynamics. Limits on care providers capacity will also lead to cluster size restrictions.

2.6.7 Further complicated designs

The above sections have described some of the most commonly used RCT designs in complex intervention research, though many others exist. One obvious extension from the above include what Walwyn and Roberts [38] term crossed designs, where care providers treat patients in both trial arms. A crossed design may result in contamination, for example, asking the same therapist or physician to present two very different therapies to their patients could be impractical, the therapist may get confused, and patients of the same therapist or physician might be acquainted [51].

A second extension from the pnRCT design is seen in the E-SEE trial and trials of other proportionate interventions. These are a more complex version of the pnRCT in which only one arm has clustered outcomes and within that arm only some of the patients have clustered outcomes, as not all receive the clustered intervention. For instance, in E-SEE only those who step-up to the parenting groups will be clustered by group, the other participants receive only the parenting book (their outcomes would not be expected to be clustered). For the remainder of this thesis these types of trials are referred to as 'within-arm partially nested trials' (within-arm pnRCTs) in which only some of the intervention arm have clustered outcomes.

Landau and Chalder [52] recommend "where feasible randomise participants to clusters that are related to the delivery of treatment (therapists, groups)". However, in trials such as E-SEE group interventions are offered only to a subset of the intervention arm (those who meet some criteria after the previous intervention stage). In addition, only one or two groups are to be delivered in each location and group allocation is based on nearest available group. This type of trial design will result in non-random allocation to groups, consequently it is likely that a potential clustering effect occurs. For example, participants offered the group intervention are those with more similar responses and people from the same catchment area may have similarities, such as socio-economic class and health, adding to the potential clustering effect (if it is not possible to fully adjust for the predictive characteristics using baseline covariates).

2.6.8 Multi-centre trials

All of the above designs have only considered either one or two-levels of hierarchy in the data, the individual and a single cluster level variability. There are often more than two-levels, for instance large public health research iRCTs are commonly run across multiple sites, such as geographical

regions, NHS hospitals or mental health clinics. These trials are termed multi-centre studies, participants are recruited across multiple centres, in order to achieve the required sample size and to improve generalisability of findings. Participants from the same centre may be expected to have similar outcomes implying a positive correlation and possibly the need to account for the centre based clustering. Details of different types of clustering are presented in more detail in section 2.7.2. A multi-centre iRCT design can result in three or more levels of hierarchy in the data. For example, the SARAH: Stretching and Strengthening for Rheumatoid Arthritis of the Hand study was a multi-centre iRCT evaluating the clinical and cost-effectiveness of an exercise programme over and above usual care for Rheumatoid Arthritis. The trial included four levels of data hierarchy: seventeen NHS trusts in England, comprising 21 rheumatology and therapy departments, 48 hand therapists, and finally individual level variance [53].

2.7 Clustering

In the previous section different types of RCTs were introduced. This section goes into more detail regarding the clustering of outcomes, provides a formal definition of clustering and summarises the types of clustering that may occur in trials.

2.7.1 Defining clustering

Clustering in the context of RCTs can be defined as "when observations are grouped together based upon common attributes" [54, p.2]. Consequently, outcomes from individuals within the same cluster may be expected to be correlated to one another. The correlation of outcomes of individuals from the same cluster results in a lack of independence of outcomes.

Two key reasons for correlation of outcomes within a cluster exist. Firstly, patients within the same cluster may have similar characteristics, for example, patients from the same hospital may have similar socio-economic status. Secondly, clusters themselves can influence the patients outcome, for example, patients within the same hospital may have more similar outcomes due to the quality of the hospital staff or hospital procedures or patients being treated by the same therapist may have more similar outcomes due to the therapists' experience [54].

The 'clustering effect' is commonly quantified using the intracluster correlation coefficient (ICC). The ICC measures the extent to which outcomes from participants within the same cluster are correlated to one another [40]. This correlation violates usual assumptions for sample size

calculations and analysis methods of independent observations. If clustering is ignored and we analyse results as if independent we assume we have more information than we actually do. An estimate of the ICC is commonly used to calculate the variance inflation factor [55], also known as the design effect. The design effect is used to adjust sample sizes to allow for clustering. Further details regarding how we define and estimate the ICC are discussed in section 2.8.2 and how to calculate sample sizes in chapter 6.

Clustering is defined by Kahan and Morris [54] as either pre- or post-randomisation. Pre-randomisation clustering relates to when patients are grouped into clusters and then randomised, for example when patients present to different hospitals and then are randomised upon presentation. Post-randomisation clustering occurs when patients are randomised and subsequently assigned to clusters, for example when patients are randomised to a type of therapeutic intervention and then assigned a therapist. Whereas, if therapist were used as a stratification variable in the randomisation procedure (patients are assigned to therapist and then randomised) it would be defined as pre-randomisation clustering.

2.7.2 Types of clustering

Table 2.2 draws together and summarises the different clusters and associated clustering that may be present in RCTs. These are centre in a multi-centre RCT, cluster in a cluster RCT, care provider, and group treatment.

Table 2.2: Types of clustering in RCTs and associated clustering

Cluster	**Example**	**Possible associated clustering**
Centre in a multi-centre RCT	Hospital, NHS trust, site	Pre- and post-intervention outcomes due to being from the same centre, possibly similar characteristics. Post-intervention outcomes if delivery or implementation of the intervention varies by centre.
Cluster in cRCT	GP practice, school	Pre- and post-treatment outcomes due to being from the same cluster, possibly similar characteristics. Post-intervention outcomes if delivery or implementation of the intervention varies by cluster.
Care provider	Therapist, GP, facilitator	Post-intervention outcomes due to variability in care provider delivering the intervention being tested.
Group treatment	Parenting group, weight loss group	Post-intervention outcome may be clustered due to group treatment effects/dynamics.

2.7.3 Clustering and randomised trials

Table 2.3 draws together section 2.6 and section 2.7.2 providing a summary of differences between individual, cluster, care provider, and group treatment RCTs (adapted from Roberts and Roberts [48]). Many individual, care provider, and group treatment RCTs are also multi-centre studies and thus need to consider the possibility of clustering by centre.

Table 2.3: Summary of differences between individual, cluster, care provider, and group treatment RCTs

	iRCT*	**cRCT**	**Care provider RCT**	**Group treatment RCT**
Randomisation	Individual	Individual/Cluster	Individual/Cluster	Individual/Cluster
Cluster size	-	Mean cluster size expected to be same across intervention arms	Depends on interventions being compared, may be different across intervention arms	Depends on interventions being compared compared, may be different across intervention arms
Variance in cluster size	-	Equal between arms	Variable, based on care providers capacity	Likely to be small within an intervention
Cluster membership	-	Defined at randomisation	Defined by intervention and may be more than one care provider - ideally recorded in measurement protocol	Defined by intervention - ideally recorded in measurement protocol
ICC	-	Considered equal between arms under null hypothesis	Care provider effects	Group dynamics

*Clustering may be present due to centre in multi-centre iRCT

2.8 Analysing and measuring clustered outcomes

In addition to obtaining sufficient power and accurate results, accounting for clustering enables us to estimate the ICC. The following sections provide a description of the ICC, briefly introduce analysis methods for clustered outcomes, and how to estimate the ICC. ICCs are important for the interpretation of trial results where we may be directly interested in the group or therapist effects. ICCs are also required when calculating sample sizes for RCTs with clustering to maintain power and control Type I error rates [40]. Therefore, better reporting of ICCs in trial results papers is vital for providing an evidence base of ICCs and improving the assumptions used in trial design.

2.8.1 Defining the ICC

The ICC (introduced in section 2.7.1) explains the extent of similarity between individual outcomes within the same cluster.

The definition of the ICC comes from the expression for the correlation between outcomes from two individuals [56]. The assumptions which underlie this definition are given by Eldridge et al. [56] as:

1. any two responses from different clusters are independent, but pairs of responses within clusters are correlated, and
2. the correlation is the same for all pairs of individuals from the same cluster.

Figure 2.5 presents formal proportion of variance definitions of the ICC, it is the ratio of the between-cluster variance to the total variance (the sum of between and within cluster variance). In a pnRCT the total variance refers to the variance in the clustered trial arm. There are circumstances in which we may be interested in modelling the dependence of the correlation on cluster characteristics. However, when considering a trial with one level of clustering, for example care provider or parenting group, it is commonly assumed that there is a a single common correlation ρ presented in equation 2.2, Figure 2.5.

Figure 2.5: Definitions of the ICC

For an outcome Y, the ICC between the i^{th} and l^th responses in the j^th cluster can be defined as

$$\rho_j = \frac{\text{cov}(Y_{ij}, Y_{lj})}{\sqrt{\text{var}(Y_{ij})\text{var}(Y_{lj})}} \forall i \neq l \tag{2.1}$$

It is common to assume there is a single common correlation ρ which applies to all clusters in a trial, also known as compound symmetry, or exchangeability assumption [56]. The assumption of common correlation enables us to define a general ICC definition of

$$\rho = \frac{\text{cov}(Y_{ij}, Y_{lj})}{\sqrt{\text{var}(Y_{ij})\text{var}(Y_{lj})}} \forall i \neq l \text{ and } \forall j. \tag{2.2}$$

The true ICC will lie in the interval $[-1/(m-1), 1]$, where m is the cluster size. The lower bound of the ICC is $-1/(m-1)$ if all the clusters are the same size. When cluster sizes vary the lower bound becomes $-1/(m_{max}-1)$, where n_{max} is the largest cluster size [56]. The closer to 1 the more correlated the outcomes within the same cluster are, and vice versa. However, it is generally believed that ICCs in trials with clustering are unlikely to be negative [51] and under the proportion of variance definition of the ICC the ICC is always positive. It is plausible that

individuals treated by the same care provider, in the same group or from the same GP practice are positively correlated to one another and unlikely that they will be negatively correlated. However, negative correlation may arise in some instances. Individuals may respond to the same therapist differently, some responding positively, some with no change and some deteriorate. A negative ICC could occur if this increased variability occurs more within therapists than between therapists. Negative ICCs could also occur if there is competition among individuals being treated by the same therapist leading to an unequal distribution of limited therapist resources; if lots of resources are used for one patient in a cluster then less will be available for another patient in that cluster. For example, competition for attention from a therapist in a group administered intervention or a therapist burning out toward the end of a trial [57].

2.8.2 Analysing clustered outcomes and calculating the ICC

Different analysis methods are available to analyse clustered outcomes and estimate ICCs from both cRCTs and iRCTs with clustering. Model choice depends upon research goals, design and type of outcome.

Two standard approaches for analysing clustered data exist: analysis at the cluster level and analysis at the individual level. If statistical inference is aimed at the cluster level such as GP practice then cluster level analysis may be appropriate. If statistical inference is aimed at the individual level then individual level analysis is more suitable, and CONSORT [34] recommend the use of models that analyse individual level data whilst controlling for clustering effect.

There are four common approaches used to analyse RCT data with a continuous outcome, whilst adjusting for clustering [40]:

1. Cluster level analysis - analysis carried at the cluster or care provider level
2. Linear regression with cluster robust standard errors - analysis carried at the individual level
3. Mixed effects models - analysis carried at the individual level
4. Marginal models - analysis carried at the individual level

We can account for clustering by including the cluster as a fixed effect in a regression model. Though this method is simple to implement and has been used in practice in both cRCTs

and iRCTs with clustering it is not recommended. This strategy can result in a Type I error rate inflated above what would have occurred if we ignored the clustering altogether [43]. In addition, using the fixed effects method has been argued to limit the results of the analysis to the specific clusters used in a study (section 7.3.1.2 includes more discussion on this) and will generally produce low estimates of the treatment effect variability as the cluster level variability is removed [51].

For the following models define y_{ij} as a continuous outcome for individual i in cluster j, $i = 1, \ldots, N$, $j = 1, \ldots, k$, t_{ij} is the intervention indicator (0 = for control, 1 = for intervention), θ is the treatment effect, β_0 is an intercept term, and ϵ_{ij} errors represents individual level variation and u_j represent cluster level varaition.

2.8.2.1 Cluster level analysis

Analysis at the cluster level can be conducted using a two-stage process in which we create a summary measure of the individual outcomes for each cluster (for example proportion of individuals who quit smoking or the mean of a continuous outcome). The cluster level summaries are then analysed using an appropriate statistical test, commonly an independent two-samples t-test or a non-parametric test such as the Wilcoxon's rank sum test [31]. Randomisation ensures the cluster summary measures are statistically independent.

Cluster level analysis is often not the most efficient analysis approach. Firstly, when clusters are of differing sizes this can violate the assumptions of the independent samples t-test (that cluster summary measures are Normally distributed within each arm and that the there are equal variances across arms) [51]. Secondly, cluster level analysis is not suitable for trials in which clustering of outcomes only occurs for some individuals and not others, for example in pnRCTs only those in one trial arm belong to clusters and those in the other trial arm are independent of one another. Finally, cluster-level analysis does not provide an estimate of the ICC.

2.8.2.2 Individual level analysis

This section explains the three individual level analysis methods.

Linear regression with cluster robust standard errors

It is possible to use linear regression with robust cluster variance estimators. Robust standard errors gives regression coefficients that are more robust to violations of the underlying assumptions. The concept of robust variance estimates have been extended to cover the situation of clustering of outcomes [58]. These will be discussed in more detail in chapter 5.

Mixed effects model

Mixed effects models can be used to analyse individual level outcome data whilst accounting for both the between- and within- cluster variation. They represent the different levels in data (cluster, individual, repeated measures level) and the residual variance constitutes variance components for the different levels. They also allow multiple levels of clustering and nested data to be accounted for by adding additional random effects, such as therapists nested within sites.

Mixed effects models for clustered data typically comprise: a constant, the fixed effects (which include the intervention effect), individual residuals and a random effect representing cluster-specific effects. These models estimate the cluster specific effect of the intervention on the endpoint, the variance of the distribution of cluster means is estimated (between-cluster variance) (σ_u^2) and within-cluster variance (σ_ϵ^2). The random effects refer to u_j and are assumed to be taken at random from a population of clusters. This is also referred to as the random intercept model. The fixed part of the model states an overall regression line representing the population average outcome and the random effect u_j moves this regression line up or down according to each cluster. It is typically assumed that the cluster residuals u_j are Normally distributed.

If we consider the simple case of a randomised trial with two levels of data hierarchy for all individuals: everyone belongs to a cluster resulting in cluster level variability and there is individual level variability. The following mixed effects model can be defined for a continuous outcome for individual i in cluster j, $i = 1, \ldots, N$, $j = 1, \ldots, k$,

$$\begin{aligned} y_{ij} =& \beta_0 + \theta t_{ij} + u_j + \epsilon_{ij}, \\ u_j \sim& N(0, \sigma_u^2), \\ \epsilon_{ij} \sim& N(0, \sigma_\epsilon^2) \end{aligned} \tag{2.3}$$

where the random intercept term u_j represents between cluster variation and ϵ_{ij} the individual residuals. The above model assumes the effect of every cluster j is to add a random effect u_j to outcomes. The parameters of the mixed effects model are commonly estimated using maximum

likelihood estimation (MLE) or restricted maximum likelihood methods (REML), with the latter shown to produce less biased results particularly when there are a small number of clusters [40]. The statistical significance of the parameters are usually assessed using the likelihood-ratio, Score or Wald statistics.

In mixed effects models parameters of additional levels will be harder to estimate, for each additional level (or each additional random effect) more data is required, especially for the variance-covariance parameters of the higher levels. Convergence diagnostics may become an issue if there are a large number of levels. However, mixed effects models allow fitting of complex models such as the inclusion of covariates, stratification variables and longitudinal outcome data and their handling of missing data by incorporating all available data therefore, these are the focus of this work.

Estimation of the ICC is a by product of fitting a mixed effects models. The ICC is calculated using

$$\begin{aligned}\rho_h =& \frac{\text{cov}(Y_{ij}, Y_{lj})}{\sqrt{\text{var}(Y_{ij})}} \\ &= \frac{\text{cov}(\theta + u_j + \epsilon_{ij}, \theta + u_j + \epsilon_{lj})}{\text{var}(u_j) + \text{var}(\epsilon_{ij})} \\ &= \frac{\text{cov}(u_j, u_j)}{\text{var}(u_j) + \text{var}(\epsilon_{ij})} \\ &= \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}\end{aligned} \tag{2.4}$$

where σ_u^2 is the component of outcome variance related to differences between clusters, the between-cluster variation and σ_ϵ^2 is the component of outcome variance related to differences between individuals within clusters, the within cluster variation. The total variance of the clustered outcomes is $\sigma^2 = \sigma_u^2 + \sigma_\epsilon^2$, hence, ρ_h is the proportion of variance explained by the between-cluster variation [31]. The variance components cannot be negative, resulting in the positivity constraint of ρ_h (ρ_h lies between 0 and 1).

Marginal models using generalised estimating equations

Marginal models provide an alternative method of analysis for individual level data to estimate regression coefficients. They provide population averaged rather than cluster specific estimates [31]; the term population averaged comes because we can estimate θ by averaging over the clusters treating the correlation as nuisance parameters. The population averaged or marginal

model can be written as

$$y_{ij} = \beta_0 + \theta t_{ij} + v_{ij} \tag{2.5}$$

where the error term v_{ij} is regarded as random error, correlated within clusters. The error term is considered a nuisance term which is accounted for in the estimation procedure (there are no additional terms in the model that estimate cluster specific effects as in the mixed effects models).

The GEE method, developed by Liang and Zeger [59], can be used to estimate the regression coefficients in marginal models, estimating the intervention effect and its precision are carried out separately. The residuals are assumed to be correlated, with $\text{Corr}(v_{ij}, v_{lj}) = \rho(t_{ij}, t_{lj}; \Re)$. A working correlation matrix R is used to estimate the correlation matrix $\Re$ and different correlation structures can be assumed to model the clustering of individuals in the same cluster. An exchangeable correlation structure is commonly assumed, the same correlation within each cluster this common correlation ρ is the ICC.

GEEs typically use robust standard errors of the parameter estimates to allow for clustering which are derived from the observed variability in the data (as opposed to an underlying probability model). The significance of the intervention effect estimate is then determined using a Wald test statistic using the robust variance estimate [59]. The robust standard error estimate is consistent even when correlation structure is misspecified, however, when there are only a small number of clusters in the intervention arm it has been shown to be underestimated [31]. A number of methods have been proposed to address the limits of the robust standard error estimator.

2.8.2.3 Small numbers of clusters

The individual analysis methods are based on asymptotic theory, with the assumption that there are a large number of clusters. When there are only a small number of clusters the methods may be unreliable. Various minimum numbers of clusters for trials with continuous outcomes have been suggested in order to maintain Type I error rates at for example 5%. Hayes and Moulton [31] suggest that more than 30 clusters (15 clusters per arm) are required to use mixed effects models or GEEs. Additionally, a minimum of 30-40 clusters have been suggested for mixed effects models and 40-50 for GEEs [51]. However, performance of models is also affected by how other model assumptions are met, the size of clusters and variability of cluster sizes. Mixed

effects models have been shown to be less biased than GEEs when there are a small number of clusters [42, 51, 60]. The feasibility of running trials with large number of clusters is key to the problem. Partial nesting may add to this issue, there may be a smaller number of clusters due to clustering only occurring in one trial arm.

A number of small sample corrections have been developed to circumvent the problem of having only a small number of clusters. It is not always possible to design trials with a large number of clusters, particularly in iRCTs with clustering when the number of clusters may be limited by the number of care providers available to deliver the intervention. The corrections work in one of two ways: increasing the estimated standard error of the intervention effect or altering the degrees of freedom used to calculate confidence intervals and/or p-values for the intervention effect. Leyrat et al. [61] undertook a recent comparison of analysis methods for cRCTs with small numbers of clusters (40 or fewer). They found that unweighted and variance-weighted cluster-level analysis, mixed effects models with degree of freedom corrections (Satterthwaite), and GEE with a small-sample correction provided Type I error rate at or below 5% in most scenarios, down to as few as six or eight clusters, whereas uncorrected approaches lead to inflated Type I error rates. Consequently, where individual level analysis is required and there are a small number of clusters it is recommended to use small sample corrections for degrees of freedom (for example Satterthwaite) where possible. Small sample corrections relevant to this thesis are discussed in more detail in chapter 4 section 4.4.6.

2.8.2.4 Methods for ICC estimation

Analysis at the individual level allows estimation of the ICC as shown in equation 2.4. In addition to mixed effects models and GEEs, analysis of variance (ANOVA) can be and is commonly used to calculate the ICC. This can be done using a one-way ANOVA with the cluster level variable as a random factor. ANOVA estimates the mean squares between and within clusters, referred to as MSB and MSW, respectively. MSW estimates the within-cluster variance (estimating σ_u^2) and MSB estimates the mean square between clusters which varies due to between- and within-cluster variance (estimating $m\sigma_r^2 + \sigma_u^2$, $m$is number of individuals per cluster). The ANOVA proportion of variance estimation interpretation of the ICC is estimated using

$$\rho_A = \frac{MSB - MSW}{MSB + (m-1)MSW} \tag{2.6}$$

The ICC estimate from equation 2.6, ρ_A, can be negative. If cluster sizes are unequal the cluster size m is replaced with weighted average cluster size m_0 in equation 2.6 calculated as

$$m_0 = \frac{1}{k-1}\left(N_s - \frac{\sum m_j^2}{N_s}\right) \tag{2.7}$$

where k is the number of clusters, N_s is the total number of individuals, and m_j is the number of individuals in cluster j. Ukoumunne et al. [62] explain that the above weighted average cluster size is used in place of the arithmetic mean cluster size because the between-cluster variance can be underestimated when using the arithmetic mean.

Table 2.4 briefly summarises and compares the three different methods for estimating ICCs. Most methods will give similar ICC results unless the data are very extreme, for instance contain outliers and unbalanced cluster sizes [40].

Table 2.4: Methods to estimate ICCs

Method	**Advantages**	**Disadvantages**
One way ANOVA of between and within cluster variation	No positivity constraint, can give zero and negative ICC estimates.	Less suited with varying cluster sizes and when adjusting for covariates [63].
Mixed effects models: using MLE or REML	Can adjust for covariates and/or multiple levels of nesting.	Positivity constraint, for small ICCs the ICC estimate may have positive bias.
GEE marginal models	Can adjust for covariates.	Treats clustering as nuisance parameter, underestimates standard error of intervention effect when the number of clusters is small [64].

2.9 Summary

The motivation for this thesis originated from difficulties in the design and analysis of staged interventions such as those evaluated in the E-SEE trial. This prompted work to understand the aims of the proportionate universalism framework and proportionate interventions, how this framework fits into the wider context of complex intervention trials in public health and any particular challenges in such trials. The first aim of this thesis is to review current practice of how randomised trials of proportionate interventions are designed and analysed and the extent of clustering in such trials. This is addressed in the following chapter using a systematic review of trials. An initial scoping search and discussions around the E-SEE trial identified proportionate interventions as likely to present particular issues relating to the hierarchical data structures. Consequently this thesis investigates design, analysis and reporting of trials with

partial clustering of outcomes, starting with the simpler case of pnRCTs and building up to the more complex case of a within-arm pnRCTs. This chapter outlined the different trial designs, complexities that arise through clustered outcomes and relevant terminology which will be used throughout the thesis.

The next chapter presents a systematic review of trials of proportionate intervention trials addressing the aim to review current practice of how randomised trials of proportionate interventions are designed and analysed and the extent of clustering in such trials.

Chapter 3

Design and analysis of trials of proportionate interventions: systematic review

3.1 Introduction

In chapter 2, proportionate universalism was introduced and the move towards evalutating proportionate interventions in RCTs and a few of the subsequent challenges discussed. The proportionate universal framework has been discussed in NICE guidelines [18], National Health Service (NHS) documents [65], charities [66] and by public health authorities [67]. However, there is little written in academic literature on how to actually implement proportional universalism in practice or how to assess effectiveness of these interventions. A recent framework for the application of proportionate universalism has been published [20] with the aim of filling the gap between principle and practice. The framework provides an approach for governments and policy makers but does not extend to how to best evaluate which proportionate interventions are effective in practice. This chapter presents a systematic literature review of published trials of proportionate interventions.

The current review addresses the first research aim of this thesis, to review current practice of how trials of proportionate interventions are designed and analysed and the extent of intervention induced clustering in such trials. Particular interest is given to whether any trials evaluated the effectiveness of the different components of these interventions to understand the process or

treated them as a whole. The presence of intervention induced clustering in the trials included in this systematic review are also reviewed.

In addition to my supervisors, the chapter acknowledges the collaborative support of Tracey Bywater, researcher at the University of York. Tracey contributed to the search strategy design. The work was led and carried out by myself and has been submitted to Trials journal [68].

3.2 Chapter aims

This chapter reviews current practice of how randomised trials of proportionate interventions are designed and analysed and the extent of intervention induced clustering in such trials. A systematic review is conducted to address this aim the specific objectives of this systematic methodological review are to:

1. explore how trials evaluating proportionate interventions are being conducted and reported;

2. review the type of statistical design and analysis methods being implemented in randomised trials involving staged proportionate interventions;

3. review whether trials of proportionate interventions are being analysed differently to more conventional non-proportional intervention trials and if the component parts and clustering of outcomes are considered in the analysis.

3.3 Methods

Details of the protocol for this systematic review were registered on PROSPERO (`www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42016033781`).

3.3.1 Literature search

Proportionate interventions evaluated in a randomised trial between 2010 and 2015 were sought. A scoping study was undertaken to identify relevant search terms with guidance provided by a systematic reviewer regarding search terms, databases and making full use of truncation. Advice regarding search terms was also sought from PhD supervisors and external expert collaborator

Tracey Bywater. The objective related to whether intervention induced clustering of outcomes was considered in the analysis was added to the protocol after further consideration of the motivation for this review and findings from the scoping search identified many trials with potential clustering.

Electronic searches were undertaken using the databases: MEDLINE (OvidSP), Web of Science (Core Collection), and PsycINFO. The following search terms were used in the title or abstract: 'proportionate universalism', 'proportionate intervention', 'proportionate treatment', 'staged intervention', 'staged treatment', 'adaptive treatment regimen', 'adaptive intervention', 'adaptive treatment strategy', 'dynamic treatment regimen', 'multi-level intervention', or 'stepped care'. The final search combined search terms with the Boolean operator 'OR' and the Boolean operator 'AND' to combine them with the randomised trial search strategy. The randomised trial search strategy was based on the Cochrane Highly Sensitive Search Strategies for identifying randomized trials [69]. The start date of the six year time frame was chosen based on the 2010 publication date of the Marmot review [5] (it was anticipated no trials would use the term proportionate universalism prior to this).

The final search was conducted on 16th March 2016 (after piloting of the search strategy and refining). See Figure 3.1 for the MEDLINE search strategy. It was intended to provide a thorough overview of the types of trials evaluating proportionate interventions being used in practice and not be exhaustive, therefore, additional hand searching or searching of clinical trials registers was not incorporated.

Figure 3.1: MEDLINE search strategy

1. (randomi#ed controlled trial OR controlled clinical trial).pt. OR randomi#ed.ab. OR placebo.ab. OR clinical trial as topic.sh. OR randomly.ab. OR trial.ti.
2. limit 1 to yr="2010 -2015"
3. (proportionate universalism OR proportionate intervention$ OR proportionate treatment$ or staged intervention$ or staged treatment$ OR multi-level intervention$ OR multi-level program$ OR multi-level system$ OR multi-level treatment$ OR stepped care).ab,ti.
4. (adaptive treatment$ OR adaptive treatment regime$ OR adaptive intervention$ OR adaptive treatment strateg$ OR dynamic treatment regime$).ab,ti.
5. 3 OR 4
6. 1 AND 5
7. 2 AND 6
8. limit 7 to English language

3.3.2 Eligibility criteria

All results that were trials or pilot studies (including protocols) which evaluated interventions delivered proportionate to need were eligible. An intervention was defined as proportionate when there was a variation in the intervention dependent upon either an intermediate or primary outcome measured prior to the study endpoint. The intervention included decision stages and at each stage there were intervention options based on tailoring variables and pre-defined decision rules. Interventions which were tailored without decision rules were excluded from this review. Observational studies were excluded and the review was restricted to English language results only. Where more than one article for a single study was found, the main published results articles were included if present and superseded any protocol or cost-effectiveness study. All therapeutic areas were considered and no restrictions imposed on the types of participants/demographics.

3.3.3 Quality control

No quality assessment of the identified studies was used as the purpose of this review was to understand the extent of studies evaluating proportionate interventions and how they are being designed and analysed.

3.3.4 Study selection

Searches were conducted and all duplicates removed. Study selection based on the eligibility criteria was performed by myself to identify relevant results from the search strategy. At the initial screening stage, titles and abstracts were assessed to identify eligible studies. The full articles of studies meeting review criteria were inspected to identify relevant studies that fulfil the inclusion criteria.

3.3.5 Data extraction and analysis

A dedicated data extraction tool was developed for this review in an Excel spreadsheet. The data extraction tool was discussed and finalised amongst three reviewers, myself and two thesis supervisors, to agree on data extraction fields. Data were not double-extracted for the purpose of this review. However, to quality check agreement, clarity of eligibility criteria, and the

data extraction tool two second reviewers (thesis supervisors) reviewed a random sample of ten results each. After the quality check, a small clarification to the eligibility criteria was made but no changes to the data extraction were deemed necessary. For a small number of studies a second opinion was sought from thesis supervisors where inclusion was not clear. The review evaluated designs and methods used in proportionate intervention trials, therefore, there was no meta-analysis undertaken. The data extraction included: publication year, location of study (country), therapeutic area, type of study (trial results, protocol or secondary analysis), design type, aim, eligibility criteria, intervention, tailoring variable and decision rules, number of decision stages, control intervention, final study follow-up period, sample size, primary outcome, overall statistical model, whether analysis of different stages was undertaken, and intervention induced clustering.

PRISMA guidelines for reporting systematic reviews were followed where relevant [70]. A PRISMA checklist was completed to reflect the manuscript submitted to Trials based on this the work from this chapter.

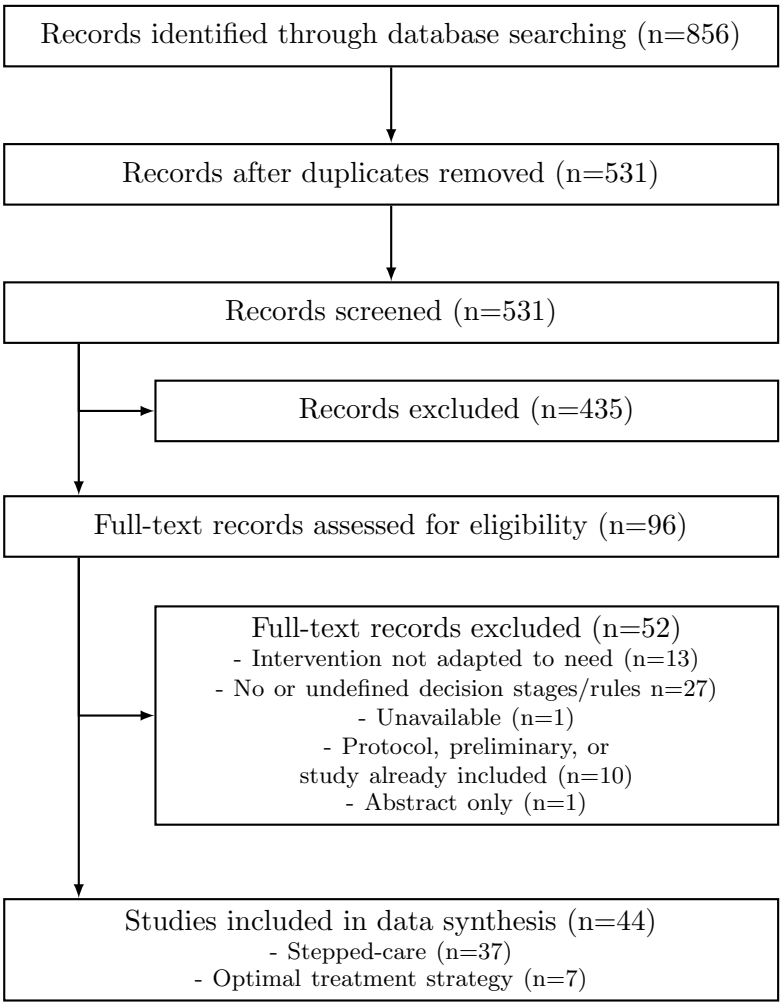
The review results were presented using summary statistics and a narrative synthesis, providing a description of any similarities and differences across the included studies. Studies were grouped by design type, with description of study characteristics tabulated to allow comparison of the main features.

3.4 Results

3.4.1 Study Selection

Figure 3.2 presents the selection of studies in this systematic review. Of the 531 unique records identified from the database search, 44 eligible studies were identified. The narrative synthesis is split into two subcategories by type of study design, stepped-care and optimal intervention strategy. Inclusion of a control was not required for eligibility. Due to the nature of assessing proportionate interventions some results did not include a control, for instance if the study objective was to identify an optimal intervention strategy.

Figure 3.2: PRISMA study flow diagram, representing the number of records identified, included and excluded during the literature search



3.4.2 Study characteristics

Table 3.1 presents an overview of included studies. There were 18 studies based in the United States, 16 in The Netherlands, two in each of Australia, England and Scotland, Norway, and Sweden, and three based in other countries (India, Nigeria and a multi-site study across France, Hungary, Romania and Slovakia).

The median number of decision stages (points at which the intervention was adapted according to need based on predefined decision rules) was 2 (interquartile range (IQR) 1-3). The median length of trial follow-up was 12 months (IQR 6-12 months) and the median sample size was 236 (IQR 150 to 387).

Measures were taken at baseline, at each decision stage and after the end of the final intervention stage. Generally, follow-up measures were also taken a number of months after completion of interventions. There were both individually and cluster randomised trials included in the results.

There were a variety of reasons given for adopting a proportionate intervention, including costs, resources, and providing interventions appropriate to individuals.

Table 3.1: Overview of studies included in the systematic review

First Author	Date[a]	Therapeutic area	Country	Follow-up[b]	N
Ell [71]	2010	Depression and anxiety	United States	12	387
Van't Veer-Tazelaar [72]	2010	Depression and anxiety	The Netherlands	12	170
Braamse [73]	2010	Distress after autologous stem cell transplantation	The Netherlands	10	286
Patel [74]	2010	Depression and Anxiety	India	12	2796
Gilliam [75]	2010	Obsessive compulsive disorder	United States	3	14
Kay-Lambkin [76]	2010	Depression among methamphetamine users	Australia	5	8
Richter [77]	2011	Blood pressure	France, Hungary, Romania, Slovakia	6	256
Weiss[d] [78]	2011	Prescription opioids dependence	United States	6	653
Mitchell [79]	2011	Bulimia nervosa	United States	12	293
Seekles [80]	2011	Depression and anxiety	The Netherlands	6	120
Tolin [81]	2011	Obsessive compulsive disorder	United States	3	34
van der Leeden [82]	2011	Anxiety in children	The Netherlands	6	133
Apil [83]	2012	Depression	The Netherlands	12	136
Karp [84]	2012	Depression and chronic pain	United States	12	250
Shortreed [d] [85]	2012	Schizophrenia	United States	18	1460
Dozeman [86]	2012	Depression and anxiety	The Netherlands	10	185
Nordin [87]	2012	Stress management of cancer patients	Sweden	12	300
Jakicic [88]	2012	Weight loss	United States	18	363
Wang[d] [89]	2012	Oncology	United States	7	150
Pommer [90]	2012	Depression and anxiety in patients with asthma or COPD	The Netherlands	24	160
Lamb [91]	2012	Whiplash injuries	England and Scotland	12	3851
Krebber [92]	2012	Distress in head and neck and lung cancer patients	The Netherlands	12	176
Borsari [93]	2012	Alcohol consumption	United States	9	598
Rose[d] [94]	2013	Smoking cessation	United States	6	606
Watson [95]	2013	Alcohol consumption	England and Scotland	12	529
Oosterbaan [96]	2013	Common mental disorders	The Netherlands	8	163
van Dijk [97]	2013	Depression among patients with diabetes and/or coronary heart disease	The Netherlands	12	236
Arving [98]	2013	Stress management of cancer patients	Norway	24	300
Mattsson [99]	2013	Depression and anxiety	Sweden	24	200
Carels [100]	2013	Weight loss	United States	4	52
van der Aa [101]	2013	Depression and anxiety	The Netherlands and Belgium	24	230
Kasari [d] [102]	2014	Communication for minimally verbal children with autism	United States	8	61
Muntingh [103]	2014	Panic and anxiety	The Netherlands	12	180
Kilbourne [d] [104]	2014	Mood disorder	United States	24	1600
Hamall [105]	2014	Families living with childhood chronic illness	Australia	6	1050
Gureje [106]	2015	Depression	Nigeria	12	1190
Stoop [107]	2015	Depression and anxiety in patients with diabetes, asthma or COPD	The Netherlands	18	46
Stam [108]	2015	Impairment in older dizzy people	The Netherlands	12	300
Lock [109]	2015	Anorexia nervosa	United States	6	45
Schuurhuizen [110]	2015	Distress in patients with metastatic colorectal cancer	The Netherlands	11	715
Haug [111]	2015	Panic and anxiety	Norway	12	173
Salloum [112]	2015	Post-traumatic stress in children	United States	3	53
Wu[d] [113]	2015	Bipolar disorder	United States	3	365
Painter [114]	2015	Depression in HIV patients	United States	12	249

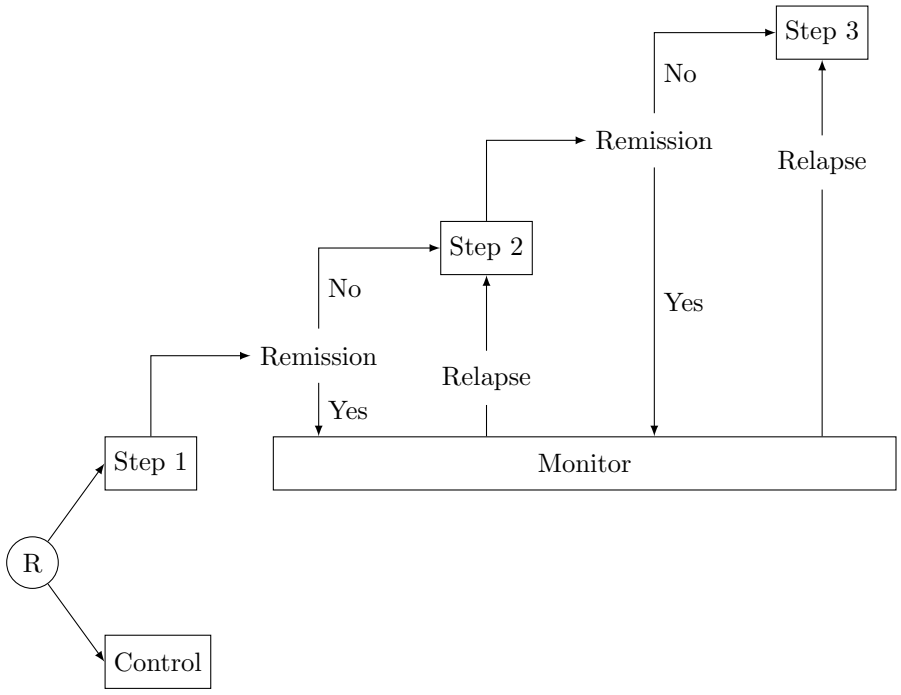
[a] Publication date, [b] Primary follow-up post baseline in months, N = Sample size, [d] Optimal-intervention strategy subcategory

3.4.3 Stepped-care

Table 3.3 presents a summary of the included studies categorised as stepped-care. A total of 84% (37 of 44) of the studies followed a stepped-care model for the intervention. The stepped-care model is recommended by NICE [115] for the provision of services for common mental health disorders. In a stepped-care model the least intensive intervention, or level of intervention, is delivered first to all patients, and patients step up or down the stepped-care pathway dependent upon their response to the previous intervention step.

Figure 3.3 represents the flow of patients through an example of a typical stepped-care trial with three intervention steps. The key principles of stepped care are: to provide the most appropriate and best intervention according to need; reduce the burden on patients by providing only the intervention required; and improve cost-effectiveness by providing the level of intervention required for a positive outcome [115]. The reduction of costs for those who respond to lower level intensity interventions can free up resources for those who require more intensive intervention [116].

Figure 3.3: Example of a stepped-care trial with three steps and the option to rejoin treatment if relapse occurs (R - Randomise)



The majority of stepped-care studies, 73% (27 of 37) were focussed on the therapeutic areas of depression, anxiety, stress, or some form of mental health disorder. Other therapeutic areas targeted included: weight loss [88, 100], alcohol consumption [95, 117], eating disorders [79, 109], whiplash injuries [91], blood pressure control [77], resilience and wellbeing of families living with

childhood chronic illness [105], and impairment in older dizzy people [108].

The intervention often involved some form of watchful waiting period for the first step followed by regular monitoring at pre-defined follow-up times of an outcome measure (either secondary or primary). Based on this outcome measure, decisions were made whether to progress to the next step or not, this process continued for however many steps were included in the intervention. Each decision stage comprised of either a choice of interventions, continued treatment, augmented treatment, or to discontinue the treatment all together for the following step based on an individuals' outcome. The progression of treatment steps for interventions aimed at mental health disorders commonly included a watchful waiting period or bibliotherapy/guided self-help, psychotherapy sessions (either individual or group based), with possible progression to medication (for example, antidepressants).

Control conditions were generally usual care or enhanced usual care with others being assessment only [105], waitlist control [93], or the active intervention delivered in a non-stepped model [81, 100, 118]. Four of the stepped-care trials included no explicit control [75, 77, 82, 87]. The lack of control was argued by [82, p.70] as "partially inherent to the stepped care design, since it would be unethical to assign children to a waiting list after a first treatment phase if they needed further treatment". This suggests a confusion in how to evaluate proportionate interventions as it would have been ethical to randomise at baseline to a control arm of care as usual. Nordin et al. [87] did not include a comparator for step one intervention, however, after step one those who continued to report stress symptoms were randomised to a group or individual format of intervention deliver (step two a or b) thus a comparison of delivery method was possible.

A variety of statistical analysis methods were used dependent upon the outcome measures and main aims. Longitudinal data were incorporated into many of the analyses. Mixed effects models, containing both fixed and random effects, were used as the statistical analysis method in 38% of studies (14 of 37), see Table 3.3. They were used to account for both longitudinal data and the clustering effects of NHS trusts, therapists, and other health professionals. Repeated measures ANOVA was used in three studies [75, 92, 99], however, this method does not successfully deal with missing values. In contrast the mixed effects model assume data are missing at random and allows for imbalance or missing observations within patient [60].

Six stepped-care studies included or planned some form of analysis of the different intervention stages. These included: summaries of outcome measures presented per intervention step [77]; analysis of outcomes after step one and step two [96]; analysis at the end of each step and the

end of the whole intervention as well as a comparison of differences in outcomes and patient characteristics (weight loss and self-monitoring characteristics) between those who were stepped down and those who remained in treatment in the stepped-care arm [100]; analysis comparing patient demographic characteristics of those who agree to participate in step two/three compared to those who decline (for eligible patients) [105]; percentages of children free of any anxiety disorder after each treatment phase and by intervention [82]; analysis of outcomes after step one and analysis of outcomes after step two adjusting for intervention received in step one and any interactions between step one and two interventions [91].

The objectives of the study by Lamb et al. [91] were to evaluate the effectiveness of step one, step two, and the combined effects of the interventions together. This was made possible by designing two linked, pragmatic, RCTs. In step one emergency departments were cluster randomised to The Whiplash Book or usual care, and individual consent was not sought at this stage. In step two participants who received either of the step one interventions and were eligible after step one (persistent symptoms at three weeks) were individually randomised at step two to either one physiotherapist advice session or up to six physiotherapist advice sessions.

Table 3.2: Abbreviations for Table 3.3

Abbreviation	**Description**
BAI	Beck Anxiety Inventory
CBT	Cognitive behavioural therapy
CES-D	Epidemiologic Studies Depression scale
CGI	Clinical Global Impression
CGI-S	Clinical Global Impression – Severity scale
CIDI	Composite International Diagnostic Interview
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders
EORTC-QLQ-C30	European Organization for Research and Treatment of Cancer QLQ-C30 quality of life questionnaire
GAD-7	Generalised Anxiety Disorder-7
GHQ-12	12-item General Health Questionnaire
HADS	Hospital Anxiety and Depression Scale
HADS-A	Hospital Anxiety and Depression Scale – Anxiety
HADS-D	Hospital Anxiety and Depression Scale – Depression
ICD-10	International Statistical Classification of Diseases and Related Health Problems-10th revision
IDS	Inventory of Depressive Symptomatology
IES	Impact of Events Scale
MASC	Multidimensional Anxiety Scale for Children
MINI	Mini International Neuropsychiatric Interview
PHQ-9	Patient Health Questionnaire
PST	Problem solving treatment
SCL-20	20-item Symptom Checklist Depression Scale
STAI	State-Trait Anxiety Inventory
WSAS	Work and Social Adjustment Scale
Y-BOCS	Yale-Brown Obsessive-Compulsive Scale

Table 3.3: Characteristics of included stepped care studies

First Author	Intervention	Tailoring variable and decision rules (response unless otherwise stated)	Primary outcome	Statistical analysis	Analysis of stages
Ell [71]	Stepped-care, 3 steps: 1) based on patient preference, patients start PST or antidepressant medication 8 weeks, 2) a different antidepressant medication or the addition of antidepressant medication or PST 4 weeks, 3) considered for additional PST, augmentation of low-dose Trazodone for insomnia, and referral to speciality mental health care.	50% SCL-20 reduction	Depression remission was assessed by SCL-20<0.5 or PHQ-9<5	Logistic regression model to compare the odds of achieving clinically meaningful improvement between treatment groups.	No
Van't Veer-Tazelaar [72]	Stepped-care, 4 steps: 1) watchful waiting, 2) bibliotherapy, 3) PST and 4) antidepressant medication. Stages were in 3 month cycles.	CES-D<16	MINI/DSM-IV diagnostic status of depressive and anxiety disorders	Incremental effectiveness computed as the difference in the probability of a disorder-free period between groups.	No
Braamse [73]	Stepped-care, 2 steps: 1) internet based self-help program, 2) contracting, individual face-to-face counselling, medication, or referral to other services.	PHQ-9≤10 and/or HADS<8 and/or STAI < 40.	Psychological distress using HADS and physical role function using EORTC-QLQ-C30.	ANOVA	No
Patel [74]	Stepped-care, 4 steps: 1) psychoeducation, 2) antidepressants, 3) interpersonal psychotherapy in addition to antidepressants or an alternative to antidepressants for those who did not respond to them, 4) referral to psychiatrist.	Varying	ICD-10 diagnosis	Chi-square and t-test. Mixed effect models for longitudinal data.	No
Gilliam [75]	Stepped-care, 2 steps: 1) short therapist sessions and bibliotherapy, 2) longer therapist directed sessions.	Y-BOCS reduction≥5 points plus a post-treatment score of ≤13	Y-BOCS total score and the clinician's CGI severity rating.	Repeated measures ANOVA	No
Kay-Lambkin [76]	Stepped-care, 4 steps: 1) brief integrated CBT/motivational interview (MI) intervention one session, 2) 4 CBT/MI sessions: 3) 4 CBT/MI sessions, 4) 4 CBT/MI sessions.	Varying	Depression and methamphetamine use.	Small sample size so no statistical analyses.	No
Richter [77]	Stepped-care, 6 steps: incremental therapy included the following add-on therapies at 4-week intervals: aliskiren 150–300 mg once daily, hydrochlorothiazide 12.5–25 mg once daily, and finally amlodipine 5–10 mg once daily, as needed.	Meet the target blood pressure at 4-week intervals.	Estimated cumulative probability of patients achieving BP target	Probability of reaching the BP target, assessed by estimating control rates of patients who reached target per visit using life-table survivor estimates at each visit. Summaries presented of change in blood pressure per treatment step.	Yes
Mitchell [79]	Stepped-care, 3 steps, 1) therapist assisted self-help for 18 weeks, 2) fluoxetine until 1 year follow-up, 3) full CBT for 6 months.	70% or more reduction in frequency of purging by the end of session six.	Recovery (no binge eating or purging behaviours in the past 28 days). Remission (no longer meeting DSM–IV criteria).	ANOVA with the site × treatment interaction.	No
Seekles [80]	Stepped-care, 4 steps: 1) watchful waiting 4 weeks, 2) guided self-help, 3) short face-to-face PST 5 sessions, 4) pharmacotherapy and/or specialised mental health care.	IDS<14 and HADS<8 and WSAS<6	IDS and HADS	t-tests	No
Tolin [81]	Stepped-care, 2 steps: 1) bibliotherapy 6 weeks, 2) therapist directed ERP sessions.	Y-BOCS decreased by ≥5 and ≤13.	Y-BOCS and cost	Mixed effects model	No

Table 3.3 – *Continued from previous page*

van der Leeden [82]	Stepped-care, 4 steps: 1) randomised to group or individual CBT sessions for children and parents, 2) five manual-based Parent-Child Treatment for Anxiety (PCTA) sessions, 3) additional five PCTA sessions.	Children diagnosed with an anxiety disorder and/or who scored below the cut-off of the MASC	Change in proportion of children with any DSM-IV anxiety disorder	Percentages of children free of any anxiety disorder after each treatment phase and by intervention (step 1 only, 1-2, and 1-2-3, and all combined). Mixed effects models for changes on continuous variables.	Yes
Apil [83]	Stepped-care, 4 steps: 1) watchful waiting 6 weeks, 2) bibliotherapy self-help booklet 6 weeks, 3) individual CBT 12 weekly sessions, 4) referral to physician or psychotherapist for any indicated treatment.	CES-D $\leq$16	Incidence of new depressive episode	Feasibility evaluated descriptively. Chi-square used to test if selective drop-out biased results on incidence of a new depressive episode.	No
Karp [84]	Stepped-care, 2 steps: 1) 6 weeks open treatment with venlafaxine xr 150 mg/day and supportive management (SM), 2) 14 weeks in which non-responders are randomised to high-dose venlafaxine xr (up to 300 mg/day) with PST for Depression and Pain or high-dose venlafaxine xr and continued SM.	PHQ-9 of $\leq$5 for 2 weeks and at least 30% improvement in the average numeric rating scale for pain.	Univariate pain and depression response and both observed and self-report disability.	Number needed to treat between 2 interventions. Repeated measures mixed-effect models for self-reported and observed physical disability between the 2 interventions across time.	No
Dozeman [86]	Stepped-care, 4 steps: 1) watchful waiting 3 months, 2) activity-scheduling 3 months, 3) life review and consultation with GP 3 months, 4) consultation with GP discuss further treatment 3 months.	Improvement of $\geq$5 points on the CES-D.	Incidence of major depressive disorder or anxiety disorder using MINI	Incidence rate ratio using an unadjusted and adjusted Poisson regression analysis of the MINI/DSM-IV depressive and anxiety cumulative incidence (1=developed a disorder, and 0=remained disorder-free) on the treatment indicator.	No
Nordin [87]	Stepped-care, 2 steps: 1) low intensity stress-management intervention given to all patients, 2a) more intensive group stress management treatment, 2b) more intensive individual stress management treatment .	A decrease in stress related symptoms measured by IES and/or HADS from clinical levels to normal results.	Subjective distress (intrusion and avoidance) assessed by IES.	Repeated measures ANOVA (continuous variables) and Chi-square test (categorical variables)	No
Jakicic [88]	Stepped-care, 6 steps: 1) monthly group intervention session + weekly mailed lessons and submission of self-monitoring diaries, 2) continue step 1 + 10-minute monthly telephone contact, 3) step 2 + a second 10-minute telephone contact each month, 4) step 3 + 1 individual in-person intervention contact per month, 5) step 4 + provided meal replacement shakes and bars to replace 1 meal and 1 snack per day, 6) step 5 + replace 1 of the telephone contacts with a second individual session per month. Modified based on weight loss achievement at 3 month intervals.	Weight loss goals 5% at 3 months, 7% at 6 months, 10% at 9 months, and remained at 10% at 12, 15, and 18 months.	Change in weight over 18 months	t-test to compare mean weight loss between groups. Mixed effects models for longitudinal data.	No
Pommer [90]	Stepped-care, 3 steps: 1) 4 sessions of extensive psychoeducation, 2) a course on coping with depression and/or anxiety, 10 consultations, 3) coaching (6 booster sessions on top of step 2) complemented with optional anti-depressant and/or anxiolytic medication.	PHQ-9<7 and/or GAD-7<8.	PHQ-9 & GAD-7 & MINI	Chi-square and t-test. Mixed effect models for longitudinal data.	No

Table 3.3 – *Continued from previous page*

Lamb [91]	Stepped-care, 2 steps: 1) The Whiplash Book advice/active management advice, 2a) single session of physiotherapist advice or 2b) up to six sessions of physiotherapy.	Non-response if persistent symptoms 3 weeks after emergency department attendance (WAD grades I–III).	Neck Disability Index (NDI)	Mixed effects model (accounts for clustering effects from NHS trusts and therapists in step 2).	Yes
Krebber [92]	Stepped-care, 4 steps: 1) watchful waiting 2 weeks, 2) guided self-help via internet or booklet 5 weeks + 6 phone/email coaching sessions, 3) PST administered by a specialised nurse, 4) specialised psychological intervention or antidepressant medication chosen in cooperation between patient and care co-ordinator.	HADS-A or HADS-D $\leq$7	HADS	Repeated measures ANOVA (continuous outcomes). GEEs used to evaluate longitudinal changes.	No
Borsari [93]	Stepped-care, 2 steps: 1) brief advice session, 2a) brief motivational intervention, 2b) assessment only.	Non-response if student has heavy episodic drinking (HED) $\geq$4 and/or alcohol-related consequences $\geq$5 in the past month they were randomised to receive step 2 or control (assessment only).	HED and peak blood alcohol content	Comparison of outcomes at 3, 6 and 9 months between those assigned to 2a or 2b using generalised estimating equations for longitudinal data.	Yes
Watson [95]	Stepped-care, 3 steps: 1) behavioural change counselling 1 session, 2) motivational enhancement therapy, 3 sessions, 3) local specialist alcohol services .	AUDIT–Consumption (3-item) (AUDIT–C) $<$ 5	Average drinks per day	Mixed effects model (accounts for variation in GP practice and allocated therapist).	No
Oosterbaan [96]	Stepped-care, 2 steps: 1) self-help course, 2) CBT in combination with antidepressant medication.	CGI-S$<$3	% of patients responding to and remitting after treatment measured using CGI-S	Logistic mixed effects models. Analysis after step 1 and step 2.	Yes
van Dijk [97]	Stepped-care, 4 steps: 1) watchful waiting, 2) guided self-help, 3) PST, 4) referal to GP.	PHQ-9$\geq$6	Cumulative incidence of DSM-IV major depressive disorder using MINI	Logistic mixed effects models.	No
Arving [98]	Stepped-care, 2 steps: 1) low-intensity stress-management consisting of 2 counselling sessions over 6 weeks, 2) more intensive stress-management treatment consisting of 4-7 sessions.	IES and/or HADS score at 6 week assessment not clinically significant.	Avoidance and intrusions	Repeated measures ANOVA (continuous variables) and Chi-square test (categorical variables).	No
Mattsson [99]	Stepped care, 2 steps: 1) Self-help material, chat forum and FAQ section, 2) CBT.	HADS subscale$<$7 at 1, 4, or 7 months after inclusion.	HADS, 20% change as clinically relevant.	Repeated measures ANOVA (regarding anxiety, depression, post-traumatic stress, and health-related QoL).	No
Carels [100]	Stepped-care, 3 steps: 1) group-based behavioural weight loss programme 6 weeks, 2a) behavioural weight loss programme 6 weeks or 2b) self-help, 3a) behavioural weight loss programme 6 weeks or 3b) self-help.	Meet the 3% weight loss target.	% weight loss	Repeated measures ANOVA (continuous variables) and Chi-square test (categorical variables) to compare differences between treatment groups at the end of each stage and the end of the whole intervention	Yes
van der Aa [101]	Stepped-care, 4-steps: 1) watchful waiting, 2) guided self-help, 3) PST, 4) referral to GP.	CES-D$<$16 or HADS-A$<$7	MINI	Survival analysis and mixed effects model.	No

Table 3.3 – *Continued from previous page*

Muntingh [103]	Stepped-care, 4 steps: 1) guided self-help, 2) CBT 6 sessions, 3) antidepressant medication prescribed by GP, 4) optimization of medication in primary care or referal to secondary care.	50% reduction in BAI score and BAI≤11	BAI score	Difference in gain BAI gain scores from baseline. Inverse probability weighting used, accounts for variation in receiving treatment.	No
Hamall [105]	Stepped-care, 3 steps: 1) family resilience and well-being fact-sheet, 2) family resilience and well-being activity booklet, 3) family resilience information support group or waitlist control.	Step 2: parents eligible if have a child attending one of 4 selected outpatient clinics at the paediatric hospital. Step 3: eligible if K10≥15.	Parental well-being (K10). Family functioning (McMasters Family Assessment Device, FAD). Social connectedness (Medical Outcomes Study Social Support Survey, MOSSSS). Family beliefs.	Descriptive statistics used for Step 1. ANOVA for effect of booklet intervention for all participants in Step 2 and sustained change tested using a repeated measures mixed effects model for the participants who did not move into Step 3. ANOVA to examine additional effect of the information support group relative to wait-list control group.	Yes
Gureje [106]	Stepped-care, 3 steps: 1a) 8 weekly psychoeducation & PST sessions, 1b) 8 weekly weekly psychoeducation & PST sessions + doctors advice on treatment, 2a) 4 monthly weekly psychoeducation & PST sessions, 2b) 8 weekly weekly psychoeducation & PST sessions, 2c) consult doctor + 8 weekly weekly psychoeducation & PST sessions, 3a) 4 monthly weekly psychoeducation & PST sessions, 3b) consult doctor + 8 weekly weekly psychoeducation & PST sessions.	Step 1: 1a) if PHQ-9 11-14 , 1b) if PHQ-9≥18. Step 2: 2a) PHQ-9<11, 2b) PHQ-9 11-17, 2c) PHQ-9≥18. Step 3: 3a) PHQ-9<11, 3b) PHQ-9≥11.	Recovery of depression at 12 months as shown by a PHQ-9≤6	Mixed effects model.	No
Stoop [107]	Stepped-care, 3 steps: 1) 4 weekly psychoeducation individual meetings, 2) 10 weekly individual meetings covering the coping with depression/anxiety course, 3) advice to meet GP to discuss optional medication and 6 booster sessions during 6 months. Followed by monitoring of symptoms of depression or anxiety in case of remission.	PHQ-9<7 and/or GAD-7<8.	Symptoms of anxiety and depression after 12-months intervention and 6 months post intervention.	ANCOVA and clinical significance in terms of effect size.	No
Stam [108]	Risk factor guided intervention including: 1) medication adjustment in case of three or more prescribed fall-risk-increasing drugs, 2) stepped care in case of anxiety disorder and/or depression, and 3) exercise therapy in case of impaired functional mobility. Those eligible for more than one intervention start them at the same time. Stepped-care, 4 steps: 1) watchful waiting 6 weeks, 2) guided self-help treatment 6 weeks, 3) problem-solving treatment max 6 sessions, 4) referral to GP.	GAD-7<10, a PHQ-9<10, or a positive PHQ-PD score.	Dizziness-related impairment, assessed using the Dizziness Handicap Inventory (DHI).	Mixed effects models for longitudinal data to compare intervention and control group, regardless of number of interventions. Subgroup analysis for 3 groups separately that received 1 of 3 interventions.	No
Lock [109]	Adaptive intervention, Intensive Family Coaching, consisting of family-based treatment (FBT)/Intensive Parental Coaching (IPC): 4 sessions of FBT + 3 session of IPC.	Weight gain≥2.3kg after FBT, proceed to IPC.	Retentions and treatment use, suitability and expectancy, clinical outcomes, changes in parental self-efficacy.	Feasibility and acceptability compared across the randomised groups (FBT vs. FBT/IPC) using chi-square test and t-test.	No

Table 3.3 – *Continued from previous page*

Schuurhuizen [110]	Targeted selection by a nurse (HADS≥13 and/or "Lastmeter"≥5), enhanced care (treatment process managed by a trained nurse) and stepped-care. Stepped-care, 4 steps: 1) watchful waiting 3 weeks, 2) a guided self-help program 5-7 weeks max 6 sessions in 10 weeks, 3) face to face PST, 4) psychotherapy, medication or a referral to other services (e.g. social work).	HADS<13	Psychological distress measured by HADS.	ANCOVA for difference between groups. Time patients entered stepped-care and the response to treatment (progression or not) are accounted for via a covariate.	No
Haug [111]	Stepped care, 3 steps: 1) short psychoeducation, 2) 10 weeks Internet-based self-help program, 3) 12 weeks individual CBT.	Two out of three of the following criteria: 1) loss of primary diagnosis (SCID-I), 2) CSR≤ 3 and reduced by at least two points, and 3) for PD, BSQ≤2.5, and for SAD SPS≤25.	Clinicians' Severity Rating (CSR) a 0-8 severity rating of the primary anxiety diagnosis	Multiple regression analyses.	No
Salloum [112]	Stepped-care, 2 steps, 1) 3 therapist-led sessions, 11 parent–child meetings at home over 6 weeks using a workbook, weekly brief phone support, online psychoeducation information and video demonstrations, 2) 9 trauma focussed CBT sessions.	PTS≤3, or a Trauma Symptom Checklist for Young Children PTS score ≤39, and an IE Clinical Global Impression-Improvement rating of 3, 2, or 1.	Trauma Symptom Checklist for Young Children - post-traumatic stress subscale (TSCYC-PTS).	Linear mixed effects models (continuous outcomes). Generalised linear mixed effects models (non-continuous outcome) for longitudinal data.	No
Painter [114]	Stepped-care, 5 steps: 1) watchful waiting, 2) depression care team treatment suggestions (counselling or pharmacotherapy, considering participant preference), 3) pharmacotherapy suggestions after review of treatment history ,4) combination pharmacotherapy and speciality mental health counselling, 5) referral to speciality mental health.	Non-response defined on 5 different measures, including: antidepressant adherence, counselling non-adherence, report of severe adverse effect, increase in PHQ-9 from baseline by ≥5, or <50% decrease from enrolment PHQ-9.	Quality-adjusted life years and percentage of participants with depression treatment response	Generalised linear models to calculate predicted expenditure for each participant to determine incremental cost. Logistic regression models to compare the odds of achieving clinically meaningful improvement (SCL-20 improved by ≥50%) between groups.	No

3.4.4 Optimal intervention strategy

A summary of the studies categorised as optimal intervention strategy is presented in Table 3.5. A total of 16% (7 of 44) of the review studies were aimed at finding an optimal intervention strategy when interventions consist of more than one stage. Unlike the majority of stepped-care studies, randomisation occured more than once and there was often no true control arm as the different proportionate intervention strategies were compared to one another. Six of the studies were explicitly defined as sequential multiple assignment randomised trials (SMARTs) with the other study based on a two phase trial design evaluating an adaptive smoking cessation intervention stratgey [94].

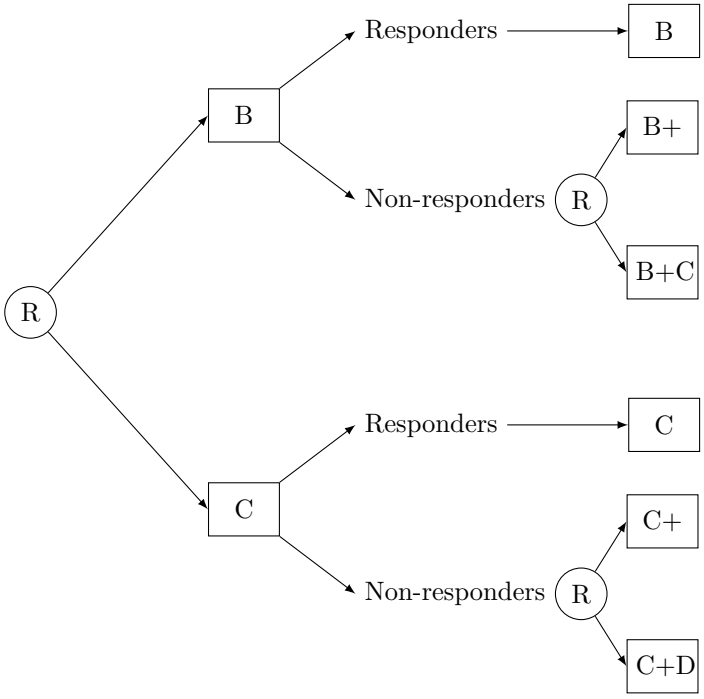
The optimal intervention strategy studies included three trial results studies [78, 94, 102], three secondary analyses of trials [85, 89, 113], and one trial protocol [104]. All seven studies were based in the United States, where the SMART design was developed [119, 120]. Therapeutic areas covered, include: oncology [89], schizophrenia [85], depression and anxiety [104], bipolar disorder [113], patients dependent on prescription opioids [78], smoking cessation [94], and communication for minimally verbal children [102].

The SMART study design has been developed to inform the development of an optimal intervention strategy. SMARTs compare groups of experimental conditions and provide the evidence to choose the intervention options at different stages. They enable the investigation of the intervention effects of different stages as part of a sequence rather than standalone interventions. The aim of a SMART is typically to develop an optimal adaptive intervention, it is recommended that the developed optimal adaptive intervention is then evaluated using a further randomised confirmatory trial comparing it to an alternative [120]. However, some SMART trials have also been run as confirmatory trials with control arms for comparison.

Five of the optimal intervention strategy studies were based on two stages of intervention and two studies used a three stage design [89, 104]. A measurement at the end of each stage was used to assess response and thus progression to the next stage. Participants were generally randomised to stage one interventions and if they were classified as responders to stage one they continued this treatment and non-responders were randomised to the following stage treatments, an example design is represented in Figure 3.4. The number of treatments at each randomisation stage varied greatly between studies, at stage one there were between two and six treatments randomised (two treatments [78, 102, 104], three treatments [113], four treatments [89], or five

treatments [85]). No control group was used in four of the studies [85, 89, 102, 104], one study used placebo in stage one [113], and usual care was used in another [78].

Figure 3.4: Example of SMART design with second randomisation dependent upon an intermediate outcome response status (R, Randomise; B, C and D, treatments; B+, enhanced treatment B, and C+, enhanced treatment C)



In general, more complex analysis methods were used for the optimal proportionate intervention strategy category compared to the stepped-care trials. Inverse probability weighting methods were used to estimate the outcome means associated with each of the two-stage dynamic treatment regimes [89]. A comparison of two treatment conditions was done using the stage two endpoint and GEEs (to account for correlation among measurements of patients from the same site) [78]. Other methods to estimate optimal intervention strategy included mixed effects models [102, 104] and Q-learning [113] .

The studies generally aimed to estimate the optimal proportionate intervention strategy as a whole rather than considering the effects of each treatment stage. Different stages of the interventions were considered in some studies. Weiss et al. [78] measured participants who responded after stage one and randomised those who did not respond to stage two. Weighted regression was used by Kasari et al. [102] to compare outcomes between the three embedded proportionate interventions including an indicator for stage one and stage two treatment and accounting for the probability of a participant following their assigned sequence of treatments based on randomisation sequence.

Table 3.4: Additional abbreviations for Table 3.5.

Abbreviation	**Description**
EF	External facilitator
EMT	Enhanced milieu teaching
IF	Internal facilitator
JASP	Joint attention symbolic play enagagement and regulation
PANSS	Positive and Negative Syndrome Scale
PSA	Prostate-specific antigen
QOL	Quality of life
REP	Replicating Effective Programs
SF-12	12-Item Short Form Health Survey
SGD	Speech-generating device

Table 3.5: Characteristics of included optimal intervention strategy studies

First Author	Intervention	Tailoring variable and decision rules (response unless otherwise stated)	Primary outcome	Statistical analysis	Analysis of stages
Weiss [78]	2 stage intervention. Stage 1: buprenorphine-naloxone induction, 2 weeks of stabilization, a 2-week taper, and 8 weeks of follow-up. Stage 2: 12 weeks of buprenorphine-naloxone stabilization, a 4-week taper, and 8 weeks of follow-up. In each phase, patients randomised to (1) standard medical management (SMM) or (2) SMM plus individual drug counseling	Stage 1: self-reported opioid use on $\leq$4 days in a month, absence of 2 consecutive opioid-positive urine test results, no additional substance use disorder treatment, and $\leq$1 missing urine sample. Stage 2: abstaining from opioids during week 12 and during $\geq$2 of the previous 3 weeks.	Composite measures indicating minimal or no opioid use based on urine test–confirmed self- reports	Compare 2 treatment conditions using the stage 2 endpoint. GEEs to account clustering of patients by site.	Yes
Shortreed [85]	2 stage intervention. Initially randomised to newer atypical antipsychotics or to perphenazine. Patients randomised at stage 1 to perphenazine who discontinue were randomised to a newer atypical antipsychotic. Patients randomised at stage 1 to newer atypical antipsychotic who discontinue were given the choice of 2 randomisation arms, including, ziprasidone, olanzapine, risperidone, or quetiapine, excluding their previous treatment or either clozapine, olanzapine, risperidone, or quetiapine, again excluding their previous treatment. Dissatisfied patients could opt to switch treatment again, at this stage treatment was neither randomised nor blinded.	Non-response if patient discontinues treatment and then eligible for randomisation to next stage.	12-month PANSS score and 12-month QOL score.	Marginal structural modelling using a weighted analysis to compare treatment regimes, the always atypical antipsychotic regime or the perphhenazine and atypical regime.	No
Wang [89]	3 stage intervention. Stage 1: randomised to 1 of 4 combination chemotherapies. Stage 2: 2a) responders receive second course of same chemo, 2b) non-responders, randomised to second-line treatment. Stage 3: After 2a, 3a) responders receive second course of same treatment, 3b) if not treatment finished. After 2b, 3c) if overall success finish treatment, 3b) if not randomised to second treatment and process repeated once more. After 3a, finish treatment.	Response defined as: Prostate-specific antigen (PSA) decline of at least 40% from baseline, objective regression (of any magnitude) of any measurable disease, improvement in any cancer-related symptoms, and no new lesions or new cancer-related symptoms. Success defined as: a PSA decline of at least 80% from baseline, resolution of all cancer-related symptoms, an objective tumour regression of at least 50% from baseline for all measurable lesions, and no new lesions or cancer-related symptoms.	Long term survival using log survival time. Efficiency in diminishing disease burden over 32 weeks using three specific scoring functions defined as functions of toxicity and efficacy taking values in the interval [0,1].	Inverse probability weighting methods to estimate the mean of counter-factual outcome for dynamic treatment regimens and sequentially randomised trials.	No

Table 3.5 – *Continued from previous page*

Rose [94]	2 stage intervention. Stage 1: all received nicotine patch treatment 2 weeks before quit date. Responders continue nicotine patch treatment. Non-responders randomised to 1) control (nicotine patch), 2) nicotine patch and bupropion, or 3) verenicline alone. Stage 2: for precessation nicotine patch responders, nonlapsers continue nicotine patch and for those who lapsed in the first week after quit date randomised to 1) control (nicotine patch), 2) nicotine patch and buprpion, or 3) verenicline alone.	Ad lib smoking (expired carbon monoxide levels) decreased by >50% after 1 week	Continuous smoking abstinence at end of treatment	Logistic regression compared each rescue treatment against the control.	Yes
Kasari [102]	2 staged intervention. Stage 1: sessions of a) JASP+EMT or b) JASP+EMT+SGD. Stage 2: early responders continue stage 1 treatment. Slow responders from 1a) randomised to receive intensified stage 1 treatment or augmented stage 1. Slow responders from 1b) receive intensified stage 1 treatment.	After stage 1 if child demonstrated 25% or greater change on at least half of the variables (7 out of 14), then the participant was considered an early responder.	Total spontaneous, communicative utterances coded from a standardised Natural Language Sample	Mixed effects models compared outcome between stage 1 treatments. Secondary aim analysis used a weighted regression to compare mean outcomes between the 3 embedded adaptive intervention, including an indicator for stage 1 and stage 2 treatments.	Yes
Kilbourne [104]	SMART design for adaptive implementation strategy. Run-in phase: sites offered REP to implement life goals (LG) for patients with mood disorders. Sites not initially responding to REP are randomised to receive additional support from an EF or both EF/IF. Additionally, sites randomised to EF and still not responsive will be randomised to continue with EF alone or to receive EF/IF.	<50% patients receiving ≥3 evidence based practice sessions	SF-12 mental health-related quality of life and PHQ-9 scores	Mixed effects models. Compare interventions in non-responding sites beginning with REP + EF/IF versus interventions beginning with REP + EF on longitudinal patient-level change in number of LG sessions received. Compare whether continuing REP + EF vs. augmenting with REP + EF/IF leads to changes in outcomes, among sites who are non-responsive to REP + EF at month 12.	Yes
Wu [113]	2 staged intervention. Stage 1: patients randomised to Bupropion, Paroxetine, or placebo. Stage 2: non-responders assigned 2nd intervention. If receive Buporopion or Paroxetine at stage 1, current doses increased. If placebo at stage 1, Bupropion or Paroxetine.	≥50% improvement over initial Scale to Assess Unawareness of Mental Disorders and not meeting DSM-IV criteria for hympomania or mania	Scale to Assess Unawareness of Mental Disorders	Q-learning to estimate optimal regime	Yes

3.4.5 Intervention induced clustering

Any potential intervention induced clustering was identified in studies included in this review. Table 3.6 presents a summary of intervention induced clustering of the trials, including the clustering type (therapist or group), the clustered intervention/s and which intervention stages had the potential for intervention induced clustering. Of the 44 trials in this review, 37 (84%) were identified as having some form of potential intervention induced clustering.

Of the stepped-care trials 95% (35 of 37) described interventions which were identified as having possible intervention induced clustering. Of the optimal treatment strategy trials 43% (3 of 7) described interventions which were identified as having possible intervention induced clustering. Though there were only a small number of these types of trials included in this review. Of the 37 studies 42% (16), 30% (11) and 27% (10), had clustering at one, two and three or more stages of the intervention, respectively. Furthermore, of these 37 studies, 54% (20), 68% (25) and 62% (23), had potential clustering at stage one, stage two, and stage three (and possibly subsequent stages) of the intervention, respectively.

Clustering was accounted for in the analysis of some trials. When cluster randomisation by site was undertaken this type of clustering was generally taken account of in both the design and analysis. Intervention induced clustering was accounted for in the analysis of two trials included in this review [91, 95]. Lamb et al. [91] ran a two stage trial with a cRCT for step one (randomisation unit was the NHS trust), evaluating The Whiplash Book versus usual advice, and an iRCT for step two, comparing physiotherapy versus reinforcement of advice given in emergency departments. The analysis used random effects to adjust for clustering of NHS trusts at step one and clustering of NHS trusts and of therapists within NHS trusts at step two. ICCs were reported for different outcomes and different follow-up times for both NHS trusts and NHS therapists. The AESOPS trial [95] involved clustering at step one of the intervention included (20-minute counselling session by a practice/research nurse) and at step two (three 40-minute motivational enhancement therapy sessions by a therapist such as an alcohol health worker, clinical nurse manager or drug and alcohol counsellor). The analysis reported that where the data allowed, the therapist/nurse identification in the AESOPS trial was added as a random effect, including three levels of data hierarchy: participant within therapist within practice [95]. In the majority of sites, step one and two were delivered by a different care provider. The same care provider delivered the interventions and steps one and two in four sites. However, the final model used was a two-level mixed model with participants nested within GP practice as

the three-level model (including nurse/therapist) resulted in a model that failed to converge. Both trials that reported to adjust for intervention induced clustering were reported in the HTA Journal (AESOPS [95] is also included in the HTA review in chapter 7).

Table 3.6: Summary of intervention induced clustering in trials included in systematic review. ✓indicates potential intevention induced clustering in that stage of intervention.

First author	Clustering type	Clustered intervention/s	Clustering stage		
			Stage 1	Stage 2	Stage 3 +
Ell [71]	Therapist	PST (stage 1, 2 & 3)	✓	✓	✓
Van't Veer-Tazelaar [72]	Therapist	PST (stage 3)			✓
Braamse [73]	Therapist	counselling (stage 2)		✓	-
Patel [74]	Therapist	Psycotherapy (stage 3), psychiatrist (stage 4)			✓
Gilliam [75]	Therapist	therapist sessions (stage 1 & 2)	✓	✓	
Kay-Lambkin [76]	Therapist	CBT/MI sessions (stage 1, 2 3, & 4)	✓	✓	✓
Weiss [78][d]	Therapist	Counselling (stage 1 & 2)	✓	✓	-
Mitchell [118]	Therapist	Psychoeducation (stage 1), depression/anxiety course (stage 2), coaching (stage 3)	✓	✓	✓
Seekles [80]	Therapist	PST (stage 3)			✓
Tolin [81]	Therapist	Therapist sessions (stage 1 & 2)	✓	✓	
Van der Leeden [82]	Therapist	Group or individual CBT (stage 1), parent-child sessions (stage 2 & 3)	✓	✓	✓
Apil [83]	Therapist	CBT (stage 3), psycotherpy (stage 4)			✓
Dozeman [86]	Therapist	Life review (stage 3) & consultation with GP (stage 3 & 4)			✓
Nordin [87]	Therapist & Group	Individual or group stress management (stage 2)		✓	-
Jakicic [88]	Therapist & Group	Group session (stage 1 - 6), individual session (stage 6)	✓	✓	✓
Pommer [90]	Therapist	Psychoeducation (stage 1), depression/anxiety course (stage 2), coaching (stage 3)	✓	✓	✓
Lamb [91]	Therapist	Pysiotherapy advice (stage 2a & 2b)		✓	
Krebber [92]	Therapist	PST (stage 3), psychological intervention (stage 4)			✓
Borsari [93]	Therapist	Advice session (stage 1), motivational intervention (stage 2)	✓	✓	
Watson [95]	Therapist	Counselling (stage 1), therapy (stage 2)	✓	✓	
Oosterbaan [96]	Therapist	CBT (stage 2)		✓	
Van Dijk [97]	Therapist	PST (stage 3)			✓
Arving [98]	Therapist	Counselling (stage 1 & 2)	✓	✓	
Carels [100]	Therapist	Group programme (stage 1-3b)	✓	✓	✓
Van der Aa [101]	Therapist	PST (stage 3)			✓
Kasari [102][d]	Therapist & Group	Speech clinician, special educator or child psychologist sessions (phase 1, children only & phase 2, children & parents)	✓	✓	
Muntingh [103]	Therapist	CBT (stage 2)		✓	
Kilbourne [104][d]	Therapist	Life goals program (LG) provider	✓	✓	✓
Hamall [105]	Group	Support group (stage 3)			✓
Gureje [106]	Therapist	Psychoeducation & PST sessions (stage 1-3)	✓	✓	✓
Stoop [107]	Therapist	Psychoeducation (stage 1), depression/anxiety course (stage 2), coaching (stage 3)	✓	✓	✓
Stam [108]	Therapist	PST (stage 3)			✓
Lock [109]	Therapist	Family coaching (stage 1), parent coaching (stage 2)	✓	✓	
Schuurhuizen [110]	Therapist	PST (stage 3), psychotherapy (stage 4)			✓
Haug [111]	Therapist	Psychoeducation (stage 1), CBT (stage 3)	✓		✓
Salloum [112]	Therapist	Therapist sessions (stage 1), CBT (stage 2)	✓	✓	
Painter [114]	Therapist	Counselling (stage 2 & 4)		✓	✓

[d] Optimal-intervention strategy subcategory. Stage is used interchangeably with step, but as a more generic term applicable to both stepped-care and optimal proportionate intervention strategies.

3.5 Discussion

3.5.1 Main findings

The results suggest that trials are being designed in various therapeutic areas that fit the proportionate universal framework. Most studies were conducted in developed countries. The term proportionate universalism was not used within the identified studies, other terminology used included: stepped-care, proportionate intervention strategy, dynamic treatment regimen, and SMARTs. In the review eligible studies fell into two main subcategories of designs: trials using the stepped-care design (to provide treatment dependent on need) or trials aimed at identifying an optimal intervention strategy (when more than one intervention was available at various stages and administered dependent upon need).

The stepped-care model begins with a lower level of intervention at the first step and treatment is only administered at further steps to those in need, randomisation generally only occurs at baseline. The optimal intervention strategy trials inform decisions on how and when to alter treatment, they generally involved randomisation at each stage dependent upon response at the end of the previous stage.

Mental health disorders were the most common therapeutic area of research in this review. This is most likely because a large majority of the results were stepped-care trials, which is a NICE recommended pathway for mental health care [115]. Reasoning for using a proportionate intervention was mainly based around costs and providing the level of care required by an individual. This is particularly relevant in mental health and complex interventions due to the sometimes resource intensive interventions (both in terms of time and costs).

Statistical methods used varied greatly based on the outcome measures, though longitudinal data is generally a feature of trials of proportionate interventions. The trials need to update and measure the changing needs of patients during delivery of the intervention, resulting in the collection of longitudinal data. ANOVA and repeated measures ANOVA were used in a number of analyses. However, these are not recommended as a general approach for longitudinal data due to the limitations in not being able to deal with missing data, failing to model the covariance among repeated measures and the use of a repeated measures ANOVA assumes an exchangeable auto-correlation structure between any two observations on the same individual [60]. More complex analysis methods were employed in the SMART studies which aimed to find the optimal intervention strategy.

Findings highlight that the nature of proportionate interventions commonly results in a complex hierarchical structure of data, with hierarchical clustering introduced by both intervention and/or centre, in addition to longitudinal data. If outcomes are correlated, not accounting for this in the analysis methods will result in an inflation in standard errors. The majority of trials in this systematic review were identified as potentially resulting in some form of intervention induced clustering.

A minority of studies considered the different stages of the interventions. Some stepped-care studies used an intention to treat analysis to compare the intervention group to the control group after each step individually and after the whole intervention period. Only one study explicitly evaluated the effectiveness of the different components as a key objective [91]. Without consideration of the separate component parts of a proportionate intervention the assumption is that each component will in itself be effective. Though this may be true, the effectiveness of the components might alter as they are incorporated with one-another. By design, the population size of a stepped-care trial on an intervention decreases as the trial progresses through the steps. This makes any comparisons between stages to be either impossible or very difficult without consideration at the design phase to account for this because of insufficent power. It is possible to evaluate the effectiveness of each stage of a proportionate intervention, as done by Lamb [91], by randomising patients who are eligible to the active or control treatment, regardless of the treatment they received at the previous stage. However, this requires large sample sizes and strong assumptions about response rates (hopefully relatively accurate) based on decision stages and rules to ensure there are enough patients to randomise at later stages of the trial. In certain scenarios it would be unethical or impossible to withhold the next stage treatment of a proportionate intervention if a patient were eligible. For example, if an unstaged version of the active treatment being tested was used as the control treatment or if each stage builds upon the previous stage in the following stage.

3.5.2 Limitations

Due to the resource limitations of this review it was not possible to supplement the database searching with reference list checking or trial registries. Only articles published in English were included. The studies included were mainly stepped-care, this may suggest limitations in the search criteria or eligibility criteria to identify other types of studies that were also trials of proportionate interventions. The review was limited to articles published after 1 January 2010,

however, this was with the aim to reflect current practice.

3.5.3 Wider context

Proportionate interventions have a role to play in the overarching goal of proportionate universalism, both in reducing health inequalities and providing care to those in need. If the use of early stage low intensity interventions provides similar outcomes to more intense interventions then costs are reduced and health interventions less onerous on both patients and health professionals. Increased intensity of treatments does not necessarily lead to increased effectiveness. Some patients are expected to respond to lower intensity interventions.

In broad terms proportionate interventions fit within the overarching goals of personalised medicine: to make decisions appropriate to an individual patient, decisions that lead to the best outcomes for the patient, and to formalise clinical decision-making and make it evidence based. Personalised medicine aims to assign individuals to interventions based on their individual characteristics and target interventions to patients likely to benefit. This requires evidence on what types of patients will benefit from different interventions, which is not always available [121]. In contrast, proportionate interventions can be self correcting with individuals failing to benefit from lower intensity interventions stepping up to more intense interventions.

The recommendation from the Marmot review for interventions to follow a proportionate universalism framework has not been supported by an evidence base on how to evaluate or implement such interventions [5]. This review provides examples of the types of interventions that fit under the proportionate universalism framework and the trial designs used to evaluate these at present.

3.5.4 Implications and recommendations

There have been recent developments in proportionate intervention strategies, trial designs now exist to develop optimal intervention strategies (SMARTs). Designs also exist to evaluate the effectiveness of stepped-care interventions as a whole. Further research considering how to design and analyse trials of proportionate interventions would benefit from considering when quantifying the effectiveness or the incremental effectiveness of each stage is necessary and how this may be implemented. This depends upon whether the separate stages have been evaluated in a trial before as well as the interactions between them, is the interaction of the different component parts expected and of interest? Without this aspect it may be unclear how all

components work and how they interact with one another.

Trialists need to consider the impact that multiple hierarchical levels (often present in proportionate interventions) have on the design and analysis. More complex mixed effects models accounting for the various correlations may be necessary. These may consider and build on methods for iRCTs with intervention induced clustering [48, 122]. Many proportionate interventions identified in this review were an extension of a partially nested trial, with intervention induced clustering often at different stages of the intervention.

Of the 52 studies excluded based on the full texts, 27 were excluded due to lack of or undefined nature of the decision stages or rules in the intervention. This lack of clarity was occasionally due to the decision rule being based on a health professionals opinion. However, lack of clarity was also repeatedly due to limited information in the articles' explanation of what the intervention actually entailed. For any trials to provide fully usable information and interventions replicable they require a clear explanation of the decision stages and rules, readers can then understand the reasoning and the process can either be implemented in a different setting or in a further trial.

When reporting trials it is important to follow both the relevant CONSORT guidelines checklist [16] and the template for intervention description and replication (TIDieR) checklist [123]. The CONSORT-NPT [17] statement also explicitly instructs authors to, where applicable, report details of whether and how clustering by care providers or centres was addressed in the sample size and statistical methods, and results (participant flow). Both CONSORT and TIDieR state that interventions must be reported with sufficient detail to allow replication, including how and when they were administered. This is particularly pertinent in proportionate interventions such as stepped-care as the how and when are often multifaceted.

3.6 Summary

The increasing demand on healthcare services has driven the move for proportionate universalism as well as the move towards fairer and more effective personalised medicine; appropriate treatment and service provision according to individual need is key. The proportionate universalism framework enables individuals to receive the care they require and reduces the burden of treatment on an individual whilst reserving resources for those most in need. This review has identified various contexts and therapeutic areas in which trials of proportionate intervention

are being designed and implemented in, mainly in the treatment of mental health disorders. Potential intervention induced clustering was identified in the majority of the studies included in this review. The term proportionate universalism was not used in any of the studies identified, though analogous terms will be used including the stepped-care model, adaptive treatment strategy, and dynamic treatment regimen. The two key types of study designs found in this review included stepped-care studies and SMART studies. The effectiveness of different stages was considered in a minority of studies and often only as a simple analysis using summary statistics.

The findings from this review are used to inform investigations of methods to analyse trials of proportionate interventions which result in within-arm nesting in chapter 5. The next chapter investigates analysis methods for pnRCTs, the more simple form of a within-arm pnRCT.

Chapter 4

Partially nested trials

4.1 Introduction

Clustering of any form needs to be considered in trial design and analysis, it can result in a reduction in statistical efficiency and if ignored standard errors and p-values for intervention effects are typically underestimated [31]. Chapter 2 introduced pnRCTs (clustering of individuals in only one arm of the trial), highlighting that these often arise in trials of complex interventions and are likely to occur in trials of proportionate interventions. Bauer et al. [49] undertook a literature review of RCTs published in four public health and clinical research journals; out of 94 RCTs, they identified 32% as pnRCTs, 40% as iRCTs and 27% as cRCTs. A lack of awareness for the need to account for the dependence of observations in the clustered arm of the pnRCTs was identified. Specific statistical methods need to be used for analysing pnRCTs, this chapter uses a simulation study to evaluate analysis methods for pnRCTs with continuous outcomes.

In theory, the mixed-effects models can be formulated for analysis of pnRCTs so that they do not model clustering in the control arm. However, when running these models in statistical software packages it is necessary to impose some form of clustering in the control arm. Although there is literature on the topic of analysing pnRCTs [42, 45, 48, 49, 122, 124], the analysis models have not been evaluated in a systematic manner. Research to date is lacking in addressing the best way to treat the non-clustered control arm when running the models in statistical software, using scenarios of relevance in the field of public health with small clusters and small ICCs [45, 125], and estimating the effect of the variance ratio of the residuals on the model fit.

In this chapter, a series of simulations are used to evaluate the statistical analysis models for

two-arm parallel pnRCTs with continuous outcomes. This follows on from work of Flight et al. [45], a range of scenarios are evaluated including the effect of ICC, heteroscedasticity of individual level variance across trial arms, cluster size and the number of clusters. In pnRCTs there may be small numbers of clusters, for instance only seven therapists treated patients in the intervention arm of the Accupuncture for lower back pain trial [126], thus the simulations evaluate the impact of the number of clusters on statistical inference and if statistical inference remains valid using mixed effects models with a small number of clusters.

In addition to my supervisors, this chapter acknowledges the collaborative support of the following researchers: Munya Dimairo; Laura Flight; Laura Mandefield; and Stephen J Walters, ScHARR Statisticians. They contributed during the simulation conception and design for quality control and to ensure simulation scenarios where chosen of relevance to real world trials. The work was led and implemented by myself (including set-up and running of all simulations and write-up) and has been published in BMC Medical Research Methodology journal [68].

4.2 Chapter aims

This chapter addresses the research aim to evaluate commonly used analysis methods for pnRCTs to establish which methods are most appropriate and why. The specific objectives are to explore:

1. where mixed effects models are necessary,
2. methods of specifying the clusters in the non-clustered arm when fitting a model and the impact these have on the analysis,
3. the impact of cluster sizes and the number of clusters on statistical inference and,
4. the impact of heteroscedastic individual variance between trial arms on statistical inference.

4.3 Literature on pnRCTs

Table 4.1 presents a summary of the key literature on the analysis of pnRCTs. Literature was identified using an updated literature search of the one undertaken by Flight et al. [45] for more recent work and findings. Sample size calculations for pnRCTs have been addressed elsewhere [127–131] and are discussed in chapter 6.

Analysis methods for pnRCTs have mainly focussed on using mixed effects models [45, 48, 49, 122, 132–135]. These allow us to control for baseline covariates and represent the different levels in the data, including cluster, individual, and repeated measures (where applicable) and estimate the ICC as part of the primary analysis.

The cluster inducing intervention may result in a decrease or increase of the individual level variability. Consequently, in addition to accounting for the clustering, it might be expected that the variance of the individual errors to differ between trial arms in pnRCTs, termed heteroscedastic variance [48]. When an intervention arm with clustering is compared to a non-clustered control arm the between-cluster variation in the intervention arm is not present in the control arm.

Table 4.1: Summary of relevant literature on analysis of pnRCTs

Author	Relevant themes	Range of values*	Findings
Schweig [135]	Describe and compare models for pnRCTs with non-compliance using a simulation study.	Simulation for two levels of clustering, exact cluster sizes (m) unclear in paper, $c_{school} = 37$, $c_{class} = 177$, $\lambda_B = 2, 8$, $\rho_{school} = 0.005, 0.05, 0.15$, $\rho_{class} = 0.0004, 0.10, 0.25$, and $\theta = 0.087$.	Clustering and non-compliance may have a substantial impact on statistical inference about intention-to-treat effects. Provide methods that may accommodate pnRCT with non-compliance, recommend using complier average causal effect estimate (CACE) and scaling by the proportion of compliers. No mention of degrees of freedom, assumed they used default degrees of freedom method available in R lme packages.
Flight [45]	Compare models applied to four examples of pnRCTs. Compare three different methods for classifying the non-clustered control arm in pnRCTs, including: singleton clusters, one large cluster and pseudo clusters.	Examples with $\{m, c\} = \{36, 8; 24, 7; 14, 8; 5, 6\}$, and estimated $\hat{\rho} =< 0.0001, 0.02, 0.007$.	Recommend use of the heteroscedastic model, recommendations based only on re-analysis of case studies. Methods for classifying the non-clustered control arm in pnRCTs had a large impact in fully clustered mixed effects models and no measurable impact in partially nested mixed effects models. ICCs in four examples were found to be small.
Sterba [124]	Review of modelling developments for pnRCTs, focused on those particularly relevant to psychotherapy trials.		Recommend the inclusion of cluster variability in analysis model as it provides insight into treatment process (rather than treating it as a nuisance). Annotated Mplus commands for models.
Lohr [133]	Report presenting a guide to design and analysis issues for pnRCTs in education research, using example trials. Discussion of degrees of freedom issue in Appendix.		Guidance document, defines pnRCT in context of education research and show methods to analyse these using SAS. Provides SAS commands for model fitting in examples.
Korendijk [132]	Compare models for pnRCTs using a simulation study, investigate misspecification for the estimation of the parameters and their standard errors.	Simulation study with $m = 5$, $c = 10, 30, 50, 100$, $\rho = 0.05, 0.1, 0.2$, $\lambda_A = 1$, $d = 0.3$.	All models perform comparably with respect to fixed effect estimates. Recommend use of partially nested mixed effects model. Simulations were under null hypothesis and ICC was always greater than zero. No mention of degrees of freedom, assumed default degrees of freedom used from MLwiN software, and homoscedasticity was assumed for individual variances between the two arms.
Sanders [134]	Compare models for pnRCTs using simulation study in terms of Type I error and power.	Simulation study with $\{m, c\} = \{2, 10; 4, 4; 5, 4; 10, 2\}$, $\rho = 0, 0.1, 0.2, 0.3, 0.4, 0.5$, $\lambda_A = 1$, and $\omega^2 = 0, 0.01, 0.059, 0.138$.	Type I error rate increased as ICC increased, Satterthwaite degrees of freedom had better control than Kenward-Roger degrees of freedom. Found using mixed effects model for pnRCT when ICC is zero likely leads to never detecting intervention effects, observed Type I error rates nearly non-existent under all scenarios with ICC equal to zero. Recommend to evaluate if ICC is significantly different from zero prior to selecting analysis method. Homoscedasticity was assumed for individual variances between the two arms.

Table 4.1 – *Continued from previous page*

Baldwin [122]	Compare analysis models for pnRCT simulation study, comparing three degrees of freedom calculations, and a pnRCT example.	Simulation for $m = 5, 15, 30$, $c = 2, 4, 8, 16$, $\rho = 0, 0.05, 0.1, 0.15, 0.3$, $\lambda_B = 0.25, 1, 4$, and $d = 0, 0.5$.	Recommend pnRCTs take account of heteroscedasticity. Satterthwaite and Kenward-Roger degrees of freedom control Type I error rate. The heteroscedastic model provides an unbiased estimate and little reduction in power compared to the homoscedastic model. Argue that using a partially nested mixed effects model only a problem for statistical inference when the number of clusters is small. The number of clusters has greater impact on power in pnRCTs. At least eight, preferably 16 clusters, to maintain Type I error rate.
Bauer [49]	Review of RCTs to ascertain the prevalence of pnRCTS in four public health and clinical research journals. Analysis models for pnRCTs extended to include pre-test measures as covariates, individual and group level covariates, and example of pnRCT	Example with clustering in one arm $c = 41$, $m = 9$, and estimated $\hat{\rho} = 0.02$.	Out of 94 RCTs, 32% were pnRCTs, 40% iRCTs and 27% cRCT. None used methods specific to pnRCTs. Example pnRCT data could be analysed using mixed effects models. Argue pnRCTs "often increase external validity at the expense of internal validity" (p.20).
Roberts [48]	Examine the case of pnRCTs, heterogeneity, comparison of analysis methods for simulation study and present an example.	Simulation for $m = 6$, $c = 8, \rho = 0, 0.1, 0.2, 0.3$, $\lambda_A = 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2$ and $d = 0$.	Recommend pnRCTs take account of heteroscedasticity. Satterthwaite unequal variances t-test gave robust to heteroscedasticity. The heteroscedastic model gives slightly inflated test size for large ρ: suggest Satterthwaite degrees of freedom as a solution.
Lee [41, 42]	Describe analysis models for iRCTs with clustering and apply to two examples (using Bayesian approach)		Show that ignoring clustering may underestimate uncertainty, leading to incorrect conclusions.
Hoover [136]	Statistical tests for RCTs with clustering that differ across trial arms.	Example with clustering in both arms with $m = 7 - 12,\ c = 3$.	Provide an adjustment for the independent samples t-test for pnRCTs. Statistical impact of heterogeneity effect increases as the cluster size increases, and as heterogeneity increases. The test does not allow for the inclusion of covariates, multiple treatments, baseline measures, or non-normally distributed outcomes.

*m = cluster size, c = number of clusters, ρ= ICC, d = standardised effect size, ω^2= Omega Squared effect size percent of variability accounted for by treatment condition, λ_A = ratio of total variance in control arm compared to clustered, λ_B= ratio of individual variance in control arm compared to clustered arm. Ordered by year of publication.

During research of the literature I found an error in the description of the simulation study in Baldwin et al. [122] paper, they had simulated a ratio of standard deviations as opposed to a ratio of variances. Following contact with the corresponding author a correction was made to this paper.

4.4 Analysis methods for partially nested trials

As discussed in chapter 2, it is not recommended to analyse clustered trial data using a fixed effect for each individual cluster in the model. This section provides more information on the main modelling approaches currently available and used for pnRCTs, including ignoring clustering altogether, imposing clustering in the non-clustered control arm, and explicitly modelling the partially nested design by modelling clustering only in the intervention arm.

Define y as a continuous outcome, i is the individual participant indicator, j is the cluster indicator, t_{ij} is the intervention indicator (0 = control, 1 = intervention), θ is the intervention effect, β_0 is an intercept term. Error terms are defined depending on the model procedure, represented using ϵ, u, and r.

4.4.1 Linear regression model

One approach, and the one often taken in pnRCTs, is to ignore the clustering altogether. Linear regression analysis ignores the grouping and uses analysis for non-clustered trials, assuming independence between individuals regardless of whether they are in the same cluster. The outcome for the ith individual ignoring any group level variation is given by the model

$$y_i = \beta_0 + \theta t_i + \epsilon_i, \tag{4.1}$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon).$$

This infers that the conditional variance of y in both the treatment and control arms is equal. Ordinary least squares (OLS) is a method to estimate parameters in linear regression model which minimises the residual (differences between observed and predicted responses) sum of squares. An important assumption of the linear regression model is that the outcomes are independent, if the outcomes of individuals in the same cluster are correlated, the independence assumption is violated and standard errors of the intervention effect are underestimated when

using the linear regression model in equation 4.1 [40].

4.4.2 Fully nested mixed effects model

We can use a fully clustered mixed effects model which includes the cluster as a random effect as in equation 4.2; this includes variability at both the individual and cluster level. The mixed effects model with imposed clustering of the control arm requires the estimation of a random cluster effect for both intervention and control arms. Some options for the imposed clustering in the control arm are given in Table 4.2. Here, the random intercept term u_j represents between cluster variation, this can model the cluster level variation and ϵ_{ij} is the individual level variation for the i^{th} individual in the j^{th} cluster. The outcome for the ith individual and the jth cluster is given by the model

$$\begin{aligned} y_{ij} =& \beta_0 + \theta t_{ij} + u_j + \epsilon_{ij}, \\ u_j \sim& N(0, \sigma_u^2), \\ \epsilon_{ij} \sim& N(0, \sigma_\epsilon^2). \end{aligned} \tag{4.2}$$

The variance of the control arm and intervention arm are assumed to be the same (homoscedastic). When the variance is believed to differ between control and intervention arm equation 4.2 is not appropriate as it does not account for heteroscedasticity [48]. Adding random coefficients at individual and group level to equation 4.2 can account for between treatment heterogeneity. Again the clustering in the control arm must be specified.

4.4.3 Partially nested mixed effects models

Alternatively the cluster effect can be applied to the clustered arm only, defined as partially nested models which accurately reflects the design of the study [48, 49, 122, 131, 135, 137]. Individuals in the non-clustered arm are assumed independent.

4.4.3.1 Homoscedastic model

In the partially nested homoscedastic model, the random effect u_j is applied to the clustered treatment arm only, between-cluster variability is only present for the intervention arm. The outcome for the ith individual and the jth cluster is given by the partially nested homoscedastic

model

$$y_{ij} = \beta_0 + \theta t_{ij} + t_{ij} u_j + \epsilon_{ij}, \tag{4.3}$$
$$u_j \sim \mathcal{N}(0, \sigma_u^2),$$
$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2).$$

Equation 4.3 equates to a random intercept model for each cluster in the intervention arm,

$$y_{ij}|(t_{ij} = 0) = \beta_0 + \epsilon_{ij},$$
$$y_{ij}|(t_{ij} = 1) = \beta_0 + \theta + u_j + \epsilon_{ij}.$$

Mutual independence of the random components $u_j \perp \epsilon_{ij}$ is assumed. The ICC in the treatment arm is $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\epsilon^2)$. There is no ICC for the control arm as there is no estimate of between cluster variance; total variance differs between the control (σ_ϵ^2) and treatment arm ($\sigma_u^2 + \sigma_\epsilon^2$).

Equation 4.3 is homoscedastic as the variance of the individual errors, ϵ_{ij}, between arms is the same, in practice this may differ between trial arms. It makes the assumption that the variance in the intervention arm is greater than the control arm which will not necessarily always be the case. The clustered intervention may result in a decrease or increase of the individual level variability.

4.4.3.2 Heteroscedastic model

The variance of the individual errors may differ between trial arms [48]. Therefore, equation 4.3 is extended to a partially nested heteroscedastic model in equation 4.4. This allows for differing individual errors between intervention and control arms but does not constrain the form of heteroscedasticity. For instance, in a pnRCT with a therapist led treatment in the intervention arm we might expect the individual level variance to differ between participants of the intervention and control arms. The partially nested heteroscedastic model is given by

$$y_{ij} = \beta_0 + \theta t_{ij} + t_{ij}u_j + (1 - t_{ij})r_{ij} + t_{ij}\epsilon_{ij}, \tag{4.4}$$
$$u_j \sim \mathcal{N}(0, \sigma_u^2),$$
$$r_{ij} \sim \mathcal{N}(0, \sigma_r^2),$$
$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

where ϵ_{ij} are the individual level residuals in the clustered arm and r_{ij} are the individual level residuals in the unclustered arm. Equation 4.4 gives,

$$y_{ij}|(t_{ij} = 0) = \beta_0 + r_{ij},$$
$$y_{ij}|(t_{ij} = 1) = \beta_0 + \theta + u_j + \epsilon_{ij}.$$

The variance of y in equation 4.4 differs between the control (σ_r^2) and intervention arm ($\sigma_u^2 + \sigma_\epsilon^2$) thus accounting for heteroscedasticity and not assuming the variance in treatment arm is always greater than the variance in the control arm. The ICC in the intervention arm is $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\epsilon^2)$.

4.4.4 ICC estimation

It is possible to use a mixed effects model to analyse pnRCT data with a fixed treatment slope and a random intercept related to cluster level residuals as in equations 4.3 and 4.4. Theoretically in such models, the residual cluster variance is only estimated for the clustered intervention arm.

We may consider that there are two ICC values in such a pnRCT trial: the ICC in the whole dataset which we can use to calculate how correlated the outcome is in relation to all participants in the trial across both arms (reported in Flight et al. [45]) given by

$$\rho = \sigma_u^2/(\sigma_u^2 + \sigma_r^2 + \sigma_\epsilon^2).$$

As no individual has both ϵ_{ij} and r_{ij} the above ICC is likely to be an underestimate of the true ICC. The ICC in the clustered intervention arm only (how correlated the outcome is in relation to participants in just the intervention arm) is given by

$$\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\epsilon^2).$$

In a sense, inference is based on overall ICC although the source of clustering is driven by one arm. On the other hand, if we assume the ICC in the unclustered arm is zero then one approach when designing a pnRCT is to inflate the sample size only in the clustered treatment arm using arm specific ICC. The assumption of zero ICC in the unclustered arm may be questionable and trial dependent considering what the control arm receives. The second approach for estimating ICC is used in this study.

4.4.5 Impose clustering in the control arm

Regardless of whether or not the model assumes clustering in one or both arms, there is still a need to create artificial clusters in the non-clustered arm to fit the model and within the statistical software package a decision must be made about how to code the cluster indicators in the control arm. One method is to impose clusters for all individuals, including those in the control arm, and use analysis for cRCTs with clustering in both arms.

Table 4.2 represents the different options for imposing clustering, j, in the control arm, l is the number of individuals in the control arm and k is the number of arbitrary clusters in the control group. Option one treats the control group as one single cluster; option two treats each individual in the control arm as their own cluster of size one (singleton clusters) giving $j = l$ clusters in the control arm. ICC estimation can be problematic with options one and two, in theory, it is not possible to estimate between-cluster variability in option one, or estimate within cluster variability in the control group using option two [138]. Option three imposes artificial pseudo-random clusters in the control group to overcome the problem of estimating within or between-cluster variability. The number of arbitrary clusters, k, needs to be considered. In this work it was chosen to be equal across treatment arms. In addition, option three will likely result in a lower ICC estimation due to the assumed independence of control participants.

Table 4.2: Options for imposing clustering of controls

Option	**Cluster**	**Intervention**
1	j= 0	j= 1, . . . ,c
2	j= 1, . . . ,l	j=l + 1, . . . ,c
3	j= 1, . . . ,k	j=k + 1, . . . ,c

Recent work by Flight et al. [45] investigated the effect of the different methods of imposing clustering in the control arm presented in Table 4.2 in four pnRCT case-studies. The four case-studies covered trials evaluating the effect of: specialist leg ulcer clinics (clustered by clinic), acupuncture for low back pain (clustered by acupuncturist), postnatal support in the community

(clustered by community support worker), and telephone befriending for maintaining quality of life in older people (clustered by volunteer facilitator) Little difference was found between the different methods for the fully clustered mixed effects models and there was no difference found between the different methods for the partially nested mixed effects models.

4.4.6 Degrees of freedom for fixed effect estimates

In addition to the correct choice of model, the test statistics and degrees of freedom in mixed effects models also need to be considered. In the mixed effects models described in sections 4.4.2 and 4.4.3 we typically wish to carry out significance tests for the intervention effect, θ. The intervention effect, θ, is estimated using MLE or REML and the significance is estimated using the likelihood-ratio or Wald test statistic. The significance of the intervention effect is typically determined using a Wald test statistic in statistical packages. The Wald test statistic is used to test the hypothesis that the estimated regression coeficient $\hat{\theta}$ is 0, it is obtained by comparing the estimated regression coeficent with its standard error [31]

$$W = \frac{\hat{\theta}}{se(\hat{\theta})}$$

For large sample sizes in mixed effects models, the Wald test statistic for fixed effects can be assumed Normally distributed. However, for small samples the large sample approximations may not be appropriate and the t and F distributions can be used as an approximation of the distribution of the Wald test statistic. The denominator degrees of freedom for the t and F statistics can be approximated using the Satterthwaite [139] or the Kenward-Roger [140] method to account for small samples and unbalanced data or balanced data with complicated covariance structures.

Comparison of degrees of freedom correction methods has been undertaken for cRCTs and pnRCTs with small numbers of clusters [122, 141]. The Satterthwaite small-sample degrees of freedom correction takes into account the variance structure of the data, for pnRCTs, it has been shown to be superior to the between-within method for maintaining Type I error rates (and comparable to the Kenward-Roger method) [122]. When an unadjusted analysis is suitable, Baldwin et al. [122] argue the Satterthwaite unequal variance t-test is appropriate. Following these results, the Satterthwaite approximation was used to calculate degrees of freedom (using `dfmethod()` option for `mixed`, available in Stata 14 onwards [142, 143]).

There are two commonly used packages for fitting mixed effects models in R: lme4 and nlme. The Satterthwaite approximation can be implemented for lme4 using the lmerTest package. However, lme4 does not currently implement nlme's features for modelling heteroscedasticity and correlation of residuals so it is not possible to fit the heteroscedastic model in equation 4.4 using lme4, nlme is required. At the time of running this simulation study, I was not aware of a method to implement the Satterthwaite approximation in nlme. Hence, the statistical software Stata was used for the simulation study. Stata allows the fitting of the heteroscedastic partially nested model (equation 4.4) with the Satterthwaite degrees of freedom correction.

4.4.7 Summary of analysis methods

Table 4.3 presents a summary of the models for the analysis of pnRCTs from the previous section and which were used in the simulation study.

In the simulation study, the fully clustered model 2 was parametrised using the imposed clustering from Table 4.3. The models were defined as:

1. Model 2.1 fully clustered mixed effects model with singleton clusters in the control arm;
2. Model 2.2 fully clustered mixed effects model with one large cluster in the control arm;
3. Model 2.3 fully clustered mixed effects model with pseudo clusters in the control arm.

It was anticipated that the method of imposing the clustering in the control arm would not affect the results of the methods which model clustering in only one arm, however, this was evaluated in the simulation study.

Table 4.3: Models for the analysis of pnRCTs

Model description		**Statistical model**	**Heteroscedastic residuals**
Model 1	Linear regression (ignore clustering)	$y_i=\beta_0+\theta t_i+\epsilon_i$ $\epsilon_i \sim N(0,\sigma^2_\epsilon)$ the individual level variation	No
Model 2	Fully clustered (impose clustering)	$y_{ij}=\beta_0+\theta t_{ij}+ u_j+\epsilon_{ij}$ $u_j \sim N(0,\sigma^2_u)$, a random effects term representing between cluster variation $\epsilon_{ij} \sim N(0,\sigma^2_\epsilon)$ the individual level variation	No
Model 3	Partially nested homoscedastic	$y_{ij}=\beta_0+\theta t_{ij}+ u_j t_{ij}+\epsilon_{ij}$ $u_j \sim N(0,\sigma^2_u)$, a random effects term representing between-cluster variation in clustered arm $\epsilon_{ij} \sim N(0,\sigma^2_\epsilon)$ the individual level variation	No
Model 4	Partially nested heteroscedastic	$y_{ij}=\beta_0+\theta t_{ij}+ u_j t_{ij}+r_{ij}(1-t_{ij}) +\epsilon_{ij}t_{ij}$ $u_j \sim N(0,\sigma^2_u)$, a random effects term representing between cluster-variation in clustered arm $r_{ij} \sim N(0,\sigma^2_r)$ the individual level variation in the non-clustered control arm. $\epsilon_{ij} \sim N(0,\sigma^2_\epsilon)$ the individual level variation in the clustered arm	Yes

4.5 Simulation study methods

4.5.1 Overview

A simulation study was conducted to evaluate the aims presented in 4.2, evaluating the statistical analysis models for pnRCTs presented in Table 4.3 and the imposed clustering of the control arm presented in Table 4.2. The simulation study utilised guidance on design, conduct and reporting of simulation studies [144, 145]. All models were fitted using a REML. REML has been shown to be more robust than MLE [40].

The number of iterations used in simulations was 1000, therefore, an estimated type I error of 5% would have a Monte Carlo standard error of approximately 0.7%. The Satterthwaite degrees of freedom correction computed using `dfmethod(sat)` is computationally intensive, hence, the number of simulations were partly limited due to computational time required. See Appendix A.2 for simulation code.

4.5.2 Software

All simulations were done in Stata [142] and graphs produced using ggplot2 [146] in R [147]. Stata is a statistical software package, which allows programming of simulations and analysis and presentation of results. Stata does have a `simulation` command, however, it was not used for this study as it did not provide enough flexibility. Simulations were instead written directly using a combination of loops and the Stata postfile commands.

Random number generation within Stata is technically pseudo-random, numbers are not truly random but generated from a specific algorithm. To allow the simulations to be replicable, a random number seed should be provided at the start of a simulation study and recorded. The seed for the Stata pseudo-random number generator was set at the start of each set of simulations and recorded so that simulations could be reproduced if required.

4.5.3 Data-generating mechanism

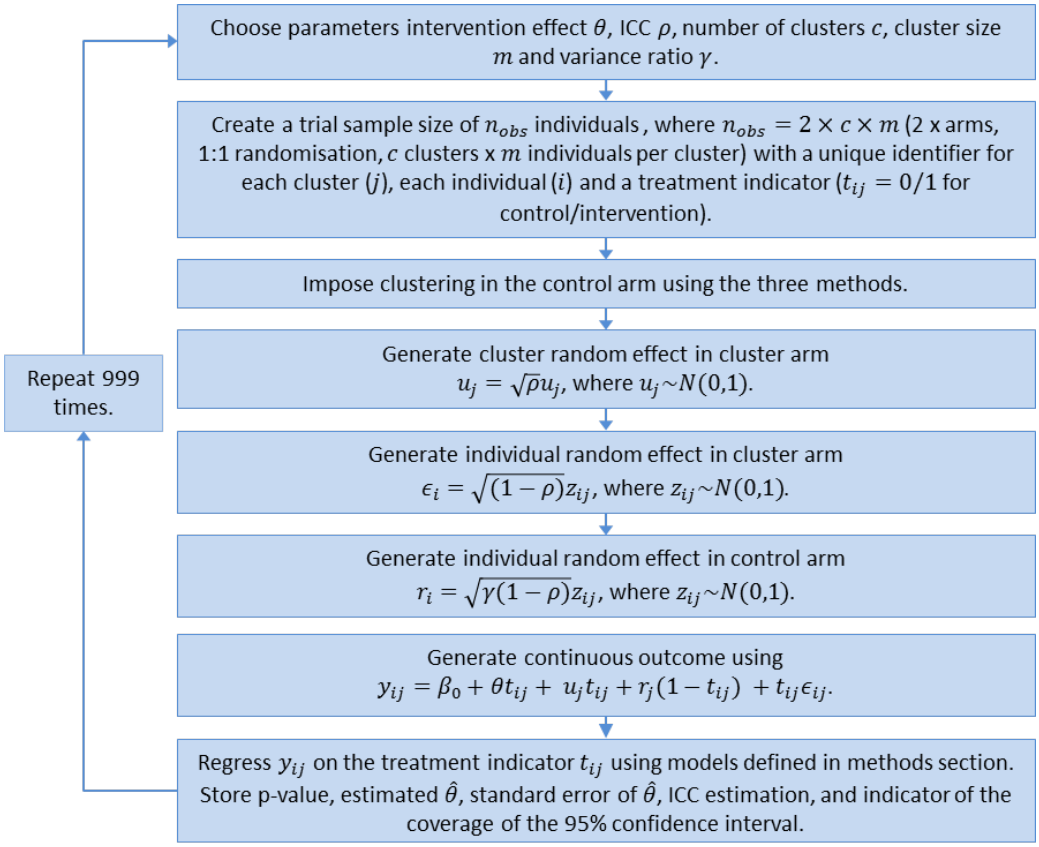
Data were simulated to replicate a two-arm parallel pnRCT trial with a non-clustered control arm and a clustered intervention arm (randomised 1:1) and a continuous outcome. Data were simulated under various design scenarios and under both the null ($\theta = 0$) and alternative hypothesis ($\theta = A$, where $A \neq 0$).

Data were simulated from the following model with the intercept set to zero and group allocation denoted by t ($t = 0$ for control arm, $t = 1$ for intervention arm):

1. For the control arm ($t = 0$) $y_{ij} = z_{ij}\sqrt{\gamma(1-\rho)}$
2. For the intervention arm ($t = 1$) $y_{ij} = \theta + \ u_j\sqrt{\rho} + z_{ij}\sqrt{(1-\rho)}$

Where $u_j \sim N(0,1)$ and $z_{ij} \sim N(0,1)$. This simulates an ICC of ρ and a ratio of individual level variance between the non-clustered control arm and the clustered intervention arm of γ. If $\gamma = 1$, there is homoscedasticity between the individual level variance in the control and intervention arms. Full simulation study steps, including the data generation process and modelling, are presented in Figure 4.1.

Figure 4.1: Flowchart representing the simulation study steps



4.5.4 Scenarios to be investigated

Simulation scenarios are presented in Table 4.4. The intervention effect, ICC, cluster size, number of clusters, and ratio of individual variance between the trial arms were varied across scenarios. Values were chosen based on literature on pnRCTs [42, 45, 48, 49, 122, 124, 132, 134], as well as extending these to more extreme cases of γ and ρ that may occur in rare instances. Reporting of ICCs in iRCTs with clustering is limited at present and it is plausible that ICCs in pnRCTs differ from those of cRCTs. Current evidence suggests ICCs in iRCTs with clustering in either one or both arms are generally small and often less than 0.05 [45, 46, 48, 50], hence the choice to include a small ICC $\rho = 0.01$ in the simulations. If $\gamma = 0.25$ then individual variance in the control arm is one quarter that in the intervention arm and if $\gamma = 4$ then individual variance in the control arm is four times that in the intervention arm. The number of clusters in the intervention group was 3, 6, 12 or 24. These figures reflect the small numbers of clusters recruited in many pnRCTs and, coupled with the cluster sizes of 5, 10, 20 or 30, they allowed alternative combinations of cluster size and number of clusters to be investigated for a given total trial size. For each of the total of 1,440 scenarios 1,000 datasets were generated.

Table 4.4: Simulation input scenario values (total 1,440 scenarios)

Variable	Notation	Values
Number of clusters	c	3, 6, 12, 24
Cluster size	m	5, 10, 20, 30
Intervention effect	θ	0, 0.2, 0.5
ICC	ρ	0, 0.01, 0.05, 0.1, 0.2, 0.3
Ratio of individual variance between control and cluster trial arms	γ	0.25, 0.5, 1, 2, 4

4.5.5 Methods

Each simulated dataset was analysed using the models described in Table 4.3: a linear regression model, a fully clustered mixed effects model, partially nested mixed effects homoscedastic model, and a partially nested heteroscedastic mixed effects model.

4.5.6 Estimand

The estimands of interest were the intervention effect θ and the ICC ρ.

4.5.7 Performance measures

The following performance measures were used:

- Bias of the intervention effect estimate: calculated as the difference between the average estimate of the effect and the true effect using Bias $= E(\widehat{\theta}) - \theta$.
- Mean square error (MSE): provides a measure of accuracy which incorporates both bias and variability and calculated using $E[(\widehat{\theta} - \theta)^2]$.
- Coverage of the intervention effect estimate 95% confidence intervals: proportion of simulations that the obtained 95% confidence interval contains the true specific intervention effect θ.
- Type I error rate: proportion of simulations in which the p-value < 0.05 when the null hypothesis is true, true intervention effect $\theta = 0$.
- Power: proportion of simulations in which the p-value < 0.05 when the alternative hypothesis is true, true intervention effect $\theta \neq 0$.

- Where applicable, model estimated ICC was calculated. ICC calculated using $\hat{\rho} = \hat{\sigma}_u^2/(\hat{\sigma}_\epsilon^2 + \hat{\sigma}_u^2)$, where $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_u^2$ are within- and between-cluster variance estimates for the clustered intervention arm.

4.6 Results

Models converged between 95% and 100% of the time across the different scenarios. Results in this chapter are presented mainly in figures and tables of results are included in the Appendix A.

4.6.1 Imposed clustering in the control arm

Methods for imposing clustering in the control arm, presented in 4.2, had a negligible impact on the performance measures of the partially nested mixed effects models (models 3 and 4). Under the simulation scenarios, the differences in the p-value, 95% confidence intervals and estimated ICC between the methods were only present at four decimal places. Model fitting was considerably faster (around four to five times faster) using either one large cluster or the pseudo clusters compared to the singleton clusters, however, this will likely be immaterial when fitting only a small number of models.

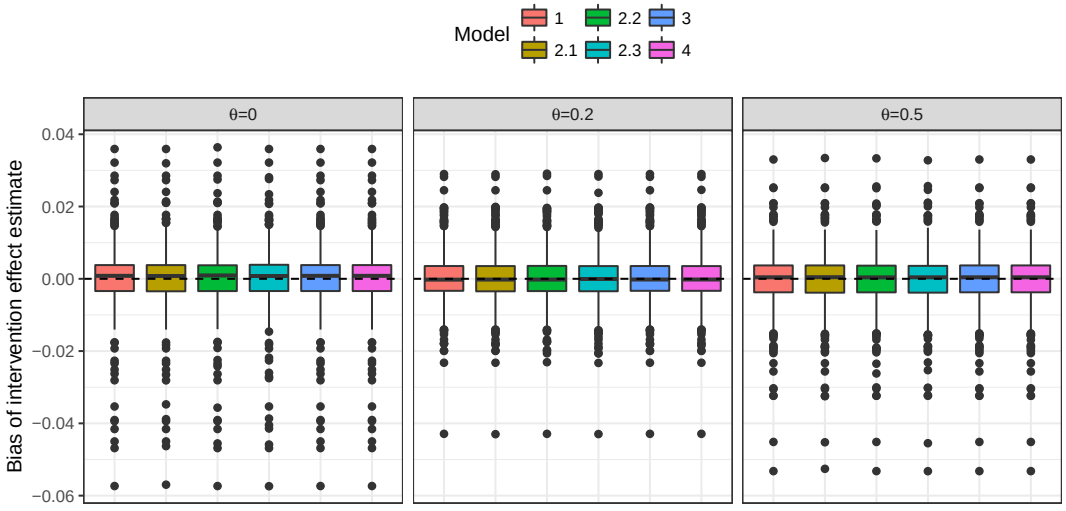
Methods for imposing clustering in the control arm had a large impact on the performance measures of the fully clustered mixed effects models (models 2.1, 2.2, and 2.3). Specific results for each performance measure are presented in the subsequent sections.

Results are reported only for the partially nested mixed effects models (models 3 and 4) with the non-clustered controls classified as one large cluster, as other methods were comparable. All three methods for classifying the non-clustered control arms for the fully clustered mixed effects model (models 2.1, 2.2, and 2.3) are reported.

4.6.2 Bias

The bias of the intervention effect estimate was not affected by the analysis model used, individual variances (γ) or the ICC (ρ). All models produced bias of the intervention effect less than $|0.057|$ under all scenarios considered. Figure 4.2 presents a box-plot of bias of intervention effect estimate $\hat{\theta}$ by θ and model.

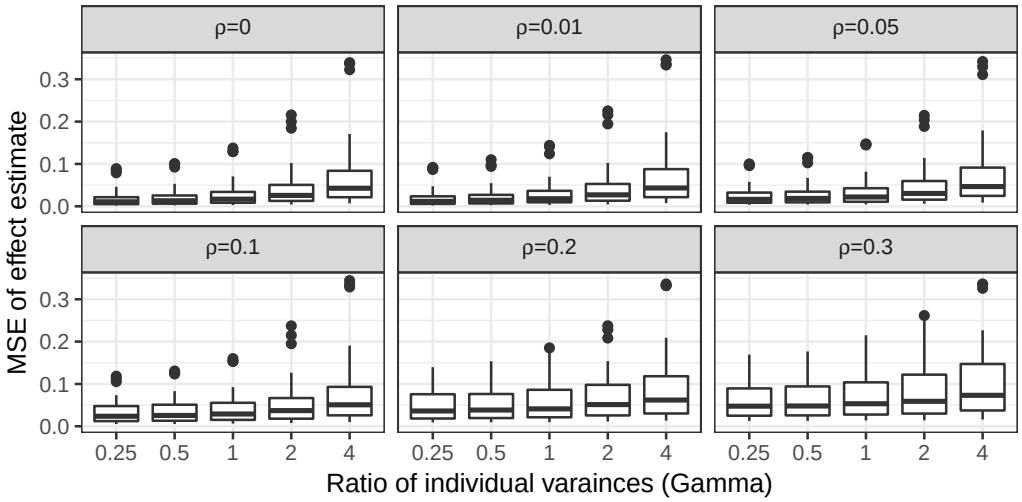
Figure 4.2: Bias of intervention effect estimate by θ and model



4.6.3 Mean square error

The models produced unbiased estimators with no difference in the observed MSE between the different models. The MSE of the intervention effect estimate had a mean of 0.051 (SD 0.056) and maximum of 0.346. Figure 4.3 shows the MSE of the intervention effect estimate by ρ and γ for model 4 as there was no difference across models.

Figure 4.3: MSE of intervention effect estimate by ρ and γ



4.6.4 Type I error

Plots of the mean Type I error rates split by model, the ratio of individual variances (γ) and the ICC (ρ) are presented in Figure 4.4. As would be expected the linear regression model which ignores clustering had inflated Type I error rates, with Type I error rate affected by ICC (ρ),

the ratio of individual variances (γ), number of clusters (c), and cluster size (m). Although the inflation was minimal when ICC $\rho = 0.01$, the mean Type I error was 0.061 (SD 0.010). When cluster size $m \leq 10$ and ICC $\rho = 0.01$ the mean Type I error rate was 0.056 (SD 0.007).

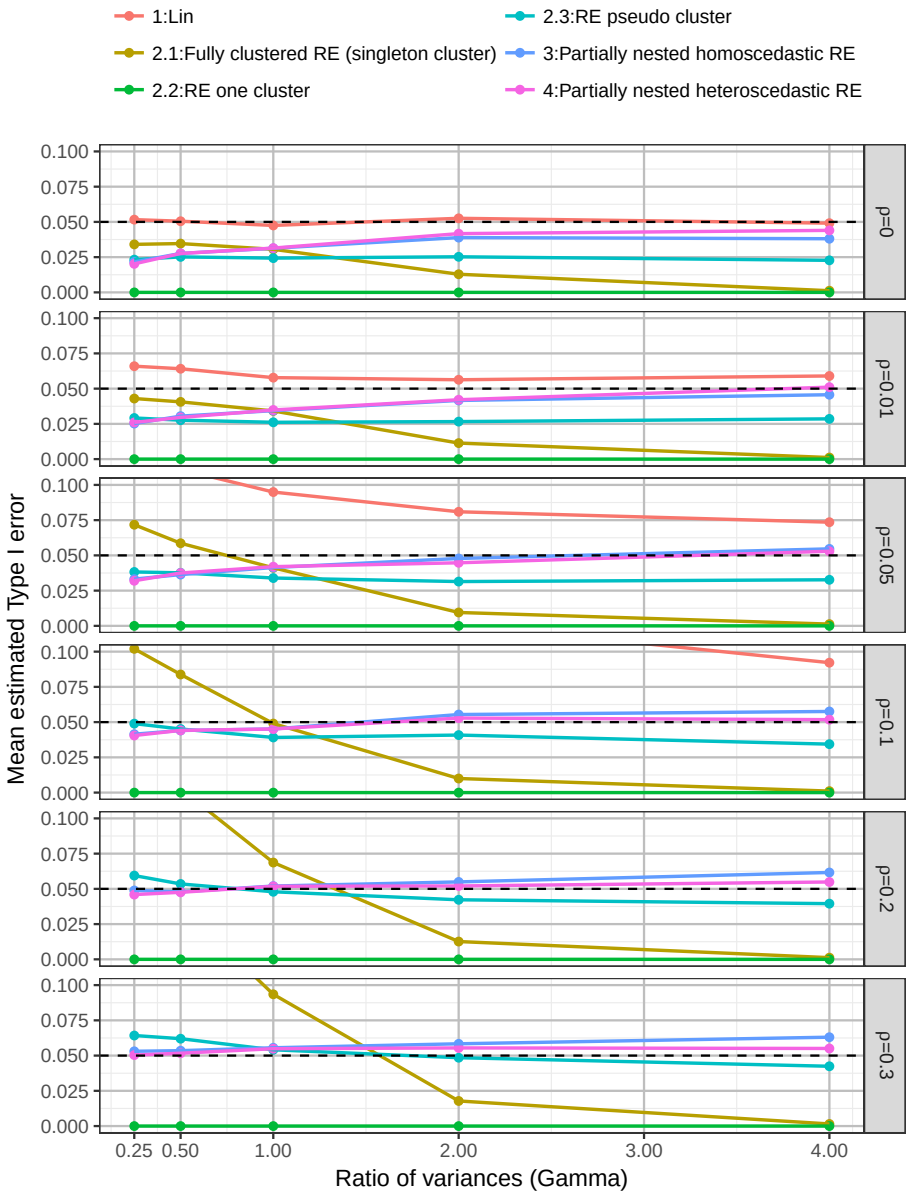
Model 2, the fully clustered models with imposed clustering in the control arm resulted in biased Type I error rates. Imposing clustering as singleton clusters (model 2.1) led to Type I error rates which were largely affected by the ratio of individual variances (γ) and ICC (ρ). Imposing one large cluster in the control arm (model 2.2) resulted in Type I error rates of zero (due to the Satterthwaite degrees of freedom correction resulting in large degrees of freedom when imposing one large cluster in the control arm). Imposing pseudo clusters in the control arm of the same size as the intervention arm (model 2.3) provided relatively good control of Type I error rates, mean Type I error of 0.039 (SD 0.018), but was affected slightly by both the ratio of individual variances (γ) and the ICC (ρ).

Both the homoscedastic and heteroscedastic partially nested models (models 3 and 4) provided good control of Type I error rates (model 3: mean Type I error 0.045 (SD 0.016) and model 4: mean Type I error 0.044 (SD 0.014)) with little difference present between the two models.

For more detailed comparison Figure 4.5 presents the Type I error rates for the linear regression model (model 1), the homoscedastic (model 3) and the heteroscedastic (model 4) partially nested models by ICC (ρ), the ratio of individual variances (γ), number of clusters (c), and cluster size (m). Higher ICC values resulted in higher Type I error rates in each model. The impact of ignoring clustering (model 1) depends on both ICC (ρ), cluster size (m), and number of clusters (c). Larger number of clusters (c) resulted in better control of Type I error rates for the partially nested models. When ICC $\rho = 0$, the Type I error rates of the partially nested models (models 3 and 4) were reduced from the nominal 5% level. This is due to the cluster variance components being estimated when they are not actually present in the data. When the ICC was small ($\rho \leq 0.05$) and the individual variance in the control arm was smaller than that in the intervention arm ($\gamma < 1$), the Type I error rates of partially nested models were reduced from the nominal 5% level. When ICC was large ($\rho = 0.3$) the partially nested models generally resulted in inflated Type I error rates. As ICC increased Type I error rates increased, with the partially nested models 3 and 4 only reaching above the nominal Type I error rate of 5% on average when ICC $\rho \geq 0.2$.

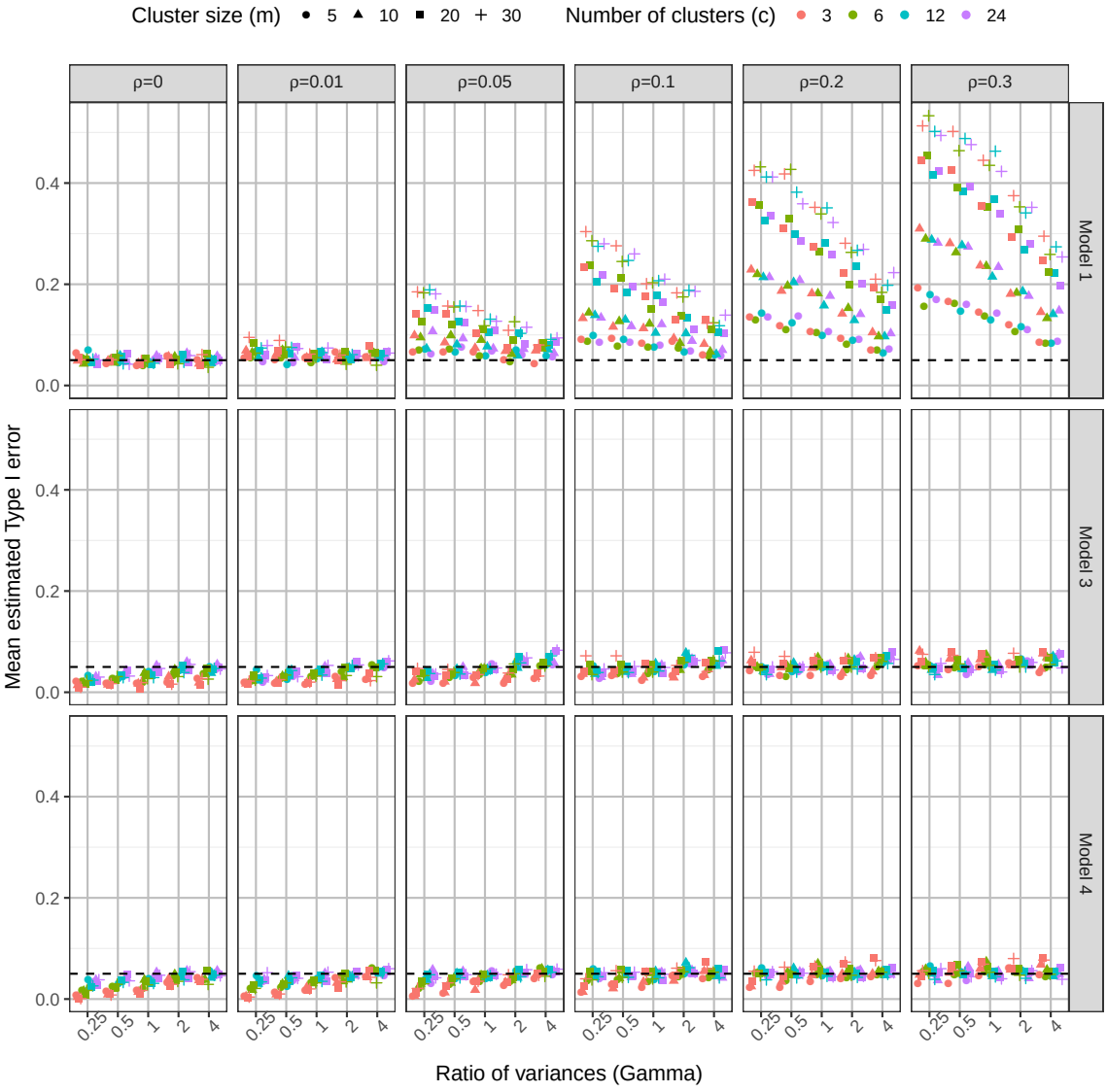
The Satterthwaite correction used in Stata `mixed` (`dfmethod(sat)`) did not fully correct the Type I error rates to the nominal level of 5%, even with the use of the heteroscedastic model 4.

Figure 4.4: Mean Type I error rate by γ and ρ over all scenarios, for each model



The heteroscedastic model 4 did have slightly improved control of Type I error rates compared to the homoscedastic model 3.

Figure 4.5: Type I error rate of models 1, 3 and 4, by ρ, γ, c, and m

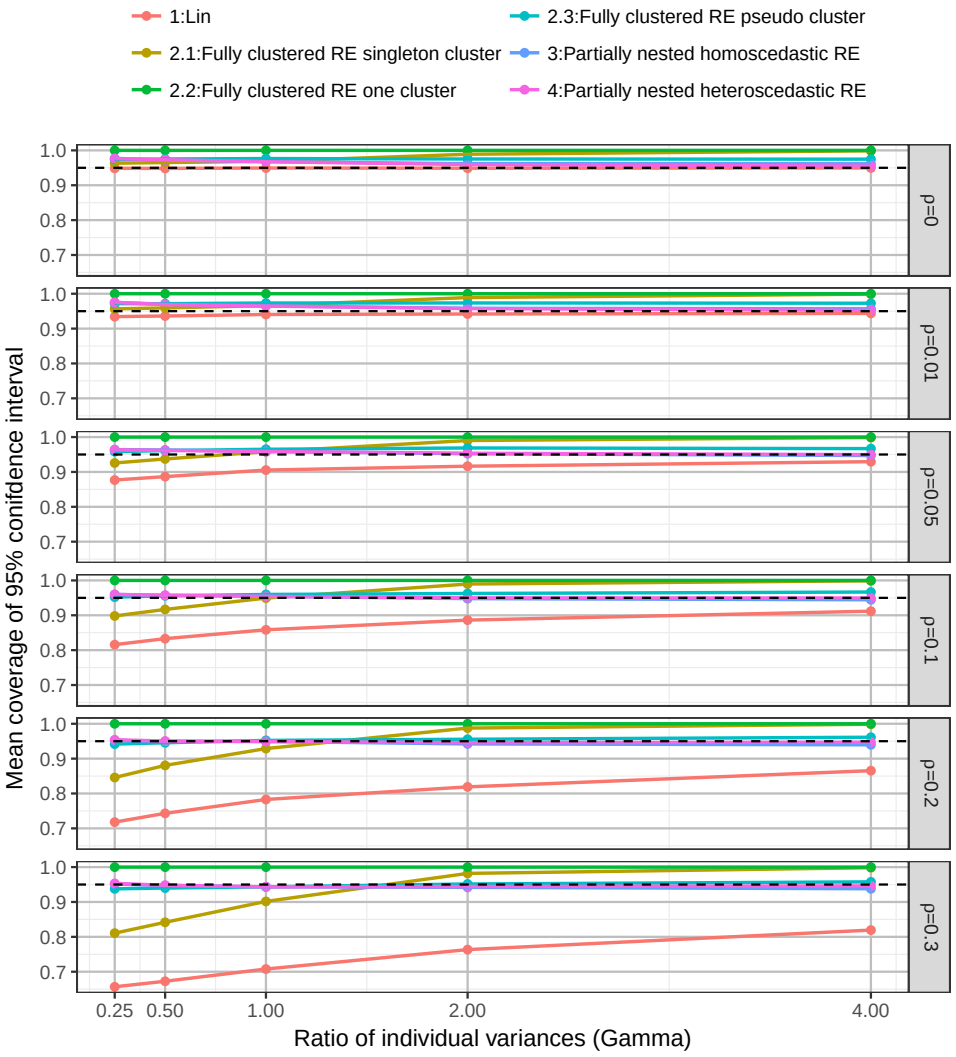


4.6.5 Coverage

Plots of the mean coverage of the 95% confidence intervals of the intervention effect estimate, split by model, ICC (ρ) and the ratio of individual variances (γ) are presented in Figure 4.6 under the alternative hypothesis. The linear regression model (model 1) resulted in under coverage when ICC was small ($\rho \leq 0.05$) and the coverage rates decrease as ICC (ρ) increases. The fully clustered models with imposed clustering in the control arm resulted in both over and under coverage dependent on the direction of the variance ratio and the method of imposed clustering. Imposing clustering as singleton clusters (model 2.1) resulted in coverage rates largely affected by the ratio of individual variances (γ). Imposing one large cluster in the control arm (model 2.2) resulted in over coverage, due to the Satterthwaite degrees of freedom correction. Imposing pseudo clusters in the control arm (model 2.3) provided mean coverage rates of 0.961 (SD 0.018).

Both the homoscedastic and heteroscedastic partially nested models (models 3 and 4) provided good control of coverage rates of 95% confidence intervals (model 3: mean coverage rate 0.956 (SD 0.014) and model 4: mean coverage rate 0.956 (SD 0.014)) with little difference between the two models. In the simulations over coverage of the 95% confidence intervals for the heteroscedastic model 4 occurred when ICC $\rho \leq 0.05$, except when the ratio of individual variances $\gamma = 4$. Hence, the results were generally conservative when ICC was small. Under coverage of the 95% confidence intervals for the heteroscedastic model 4 only occurred for large ICC (ρ) and ratio of individual variances (γ).

Figure 4.6: Mean coverage of 95% CI, by ρ and γ over all scenarios



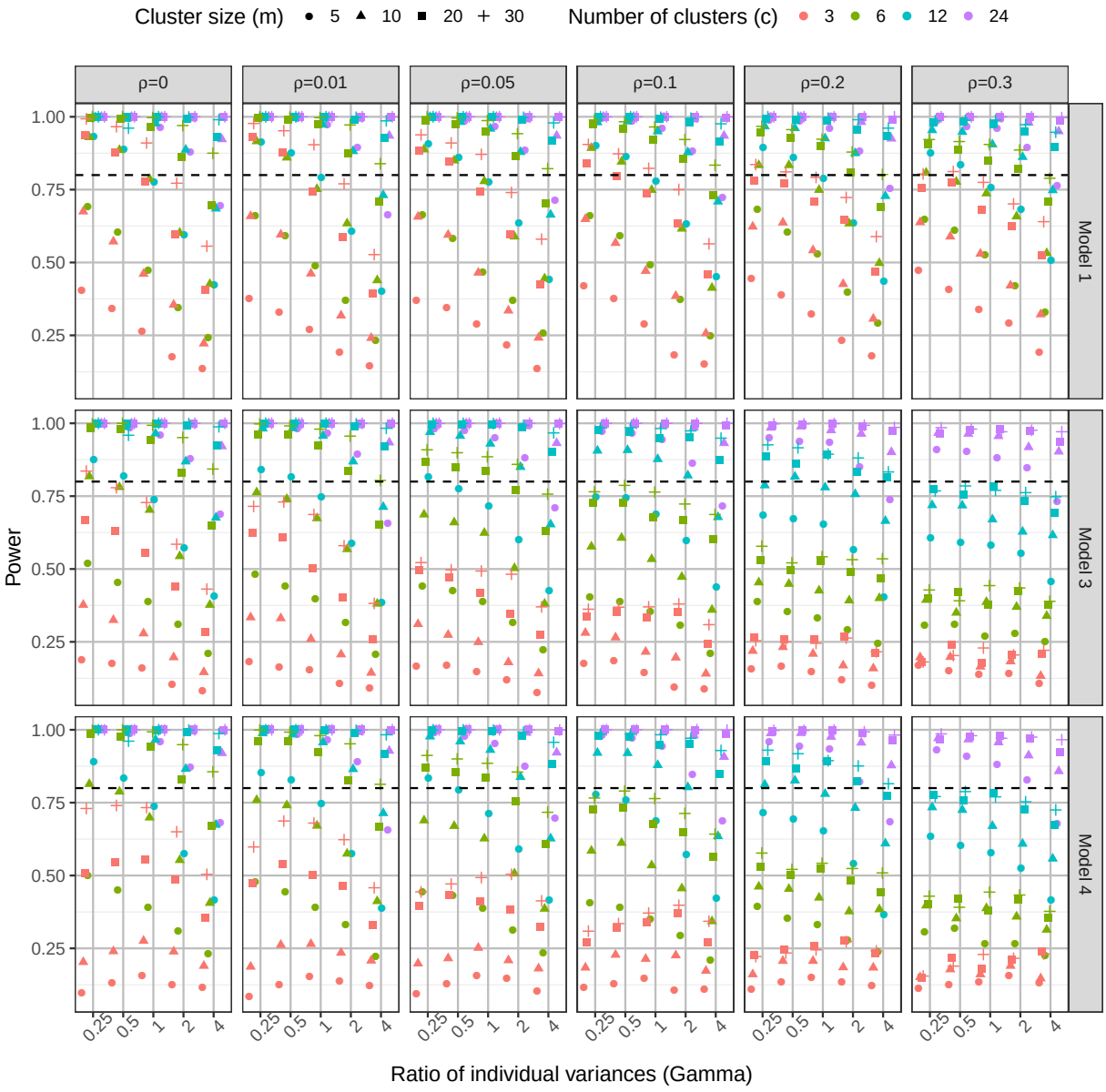
4.6.6 Power

Increasing the number of clusters as opposed to increasing the cluster size had a bigger impact on power with a fixed sample size. Figure 4.7 shows the power of the linear regression model (model

1), the homoscedastic (model 3) and the heteroscedastic (model 4) partially nested models when intervention effect $\theta = 0.5$ by ICC (ρ), the ratio of individual variances (γ), number of clusters (c), and cluster size (m) (see Appendix A for when $\theta = 0.2$). Under the simulation scenarios conducted, 12 or more clusters and cluster sizes of ten or more were generally required for a power greater than 80%. Using three or six clusters rarely gave power over 80%. Only for ICC $\rho \leq 0.05$ and relatively large cluster sizes $m \geq 20$ did the power go over 80%.

For ICC $\rho \leq 0.05$, which is commonly assumed when planning complex intervention trials in healthcare, power of 80% was generally achieved with: 24 clusters of any size, 12 clusters of size ten or more, and six clusters of size 20 or more (120 in each arm).

Figure 4.7: Power when $\theta = 0.5$, by ρ, γ, c and m



Under a ratio of individual variances $\gamma = 1$ the total residual variance in both trial arms is equal

to one, hence, the intervention effect (θ) simulated is the standardised intervention effect. Figure 4.8 shows the power of models 1, 3 and 4 under homoscedastic individual variances ($\gamma = 1$). The heteroscedastic model 4 is over-parametrised in the case of the ratio of individual variances $\gamma = 1$, however, it did not result in a substantially lower power than the homoscedastic model.

Figure 4.8: Power with standardised intervention effect of 0.5 ($\theta = 0.5$ and $\gamma = 1$)

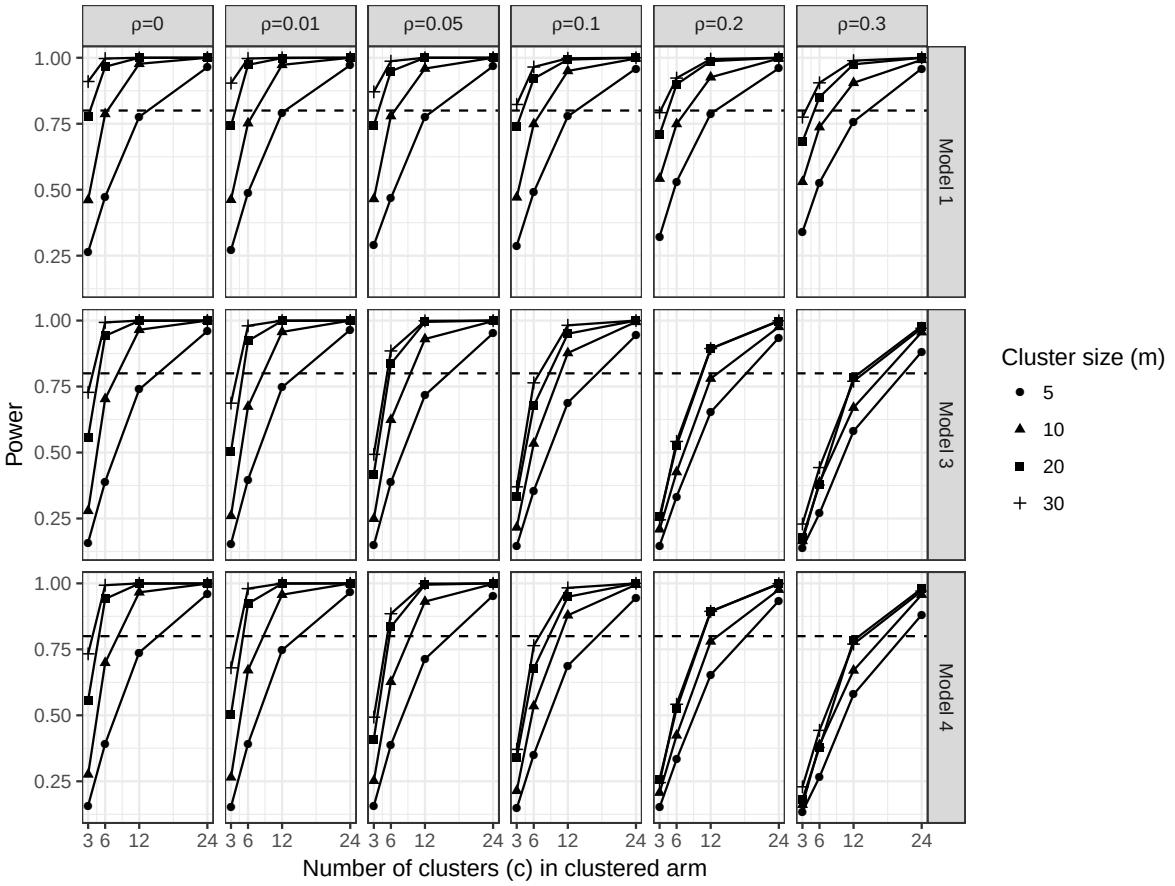


Table 4.5 presents the power of model 4 and model 1 under ICC $\rho = 0$, model 4 is over-parametrised here. There is a loss in mean statistical power which ranged between 2-6%.

Table 4.5: Mean and SD of power of model 4 versus model 1 under $\rho = 0$ over all scenarios

Intervention effect (θ)	**Model**	**Power Mean**	**SD**
0	1	0.050	0.007
	4	0.033	0.014
0.2	1	0.388	0.276
	4	0.327	0.286
0.5	1	0.803	0.254
	4	0.740	0.298

4.6.7 ICC

Figure 4.9 presents the mean estimated ICC across the fully clustered and partially nested mixed effect models, by the ratio of individual variances (γ) and ICC (ρ). ICC estimation was consistent under the heteroscedastic partially nested model (model 4). The homoscedastic partially nested model (model 3) resulted in biased ICC, with the direction of bias dependent upon the ratio of individual variances (γ).

Figure 4.10 presents the ICC for the homoscedastic (model 3) and heteroscedastic (model 4) partially nested models by the ratio of individual variances (γ), ICC (ρ), number of clusters (c), and cluster size (m). The ICC estimation from the homoscedastic model was highly affected by γ. The ICC from the heteroscedastic model was not affected by γ. Using the heteroscedastic model, there was a slight positive bias in the ICC estimation when ICC ≤ 0.05, and when ICC ≥ 0.2 there was slight negative bias in the ICC estimation. For example, when ICC = 0.0 the mean ICC estimate was 0.028 (SD 0.018), and when ICC = 0.05 the mean estimate was 0.060 (SD 0.014). As expected ICC estimation improved as sample size increased. The ICC estimation was only consistent across all values of ICC (ρ) regardless of cluster size when there were 24 clusters. For an accurate estimate of ICC when ICC = 0.05, under the simulation scenarios cluster sizes (m) of 20 or 30 were required or at least six clusters of size ten or 24 clusters of size five.

Figure 4.9: Mean estimated ICC by γ and ρ over all scenarios, for each model

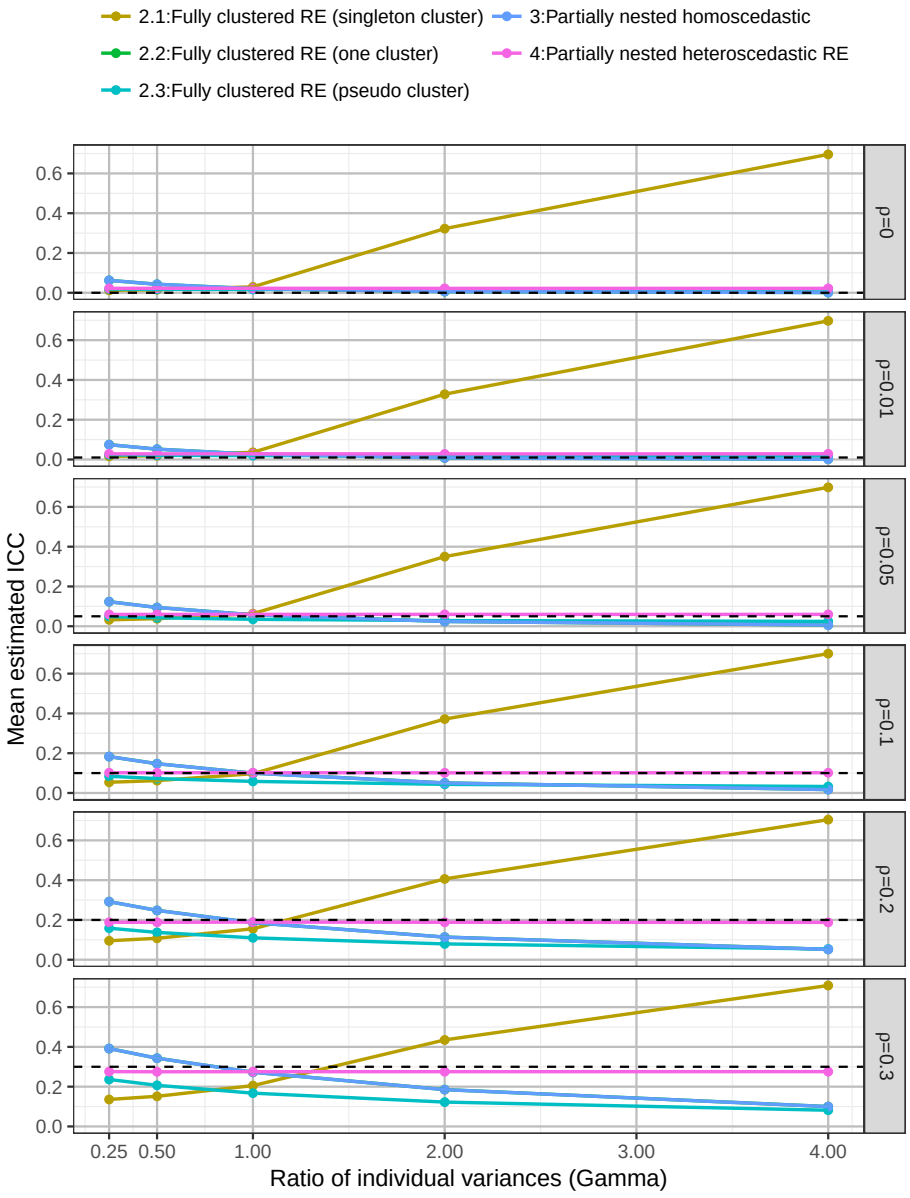
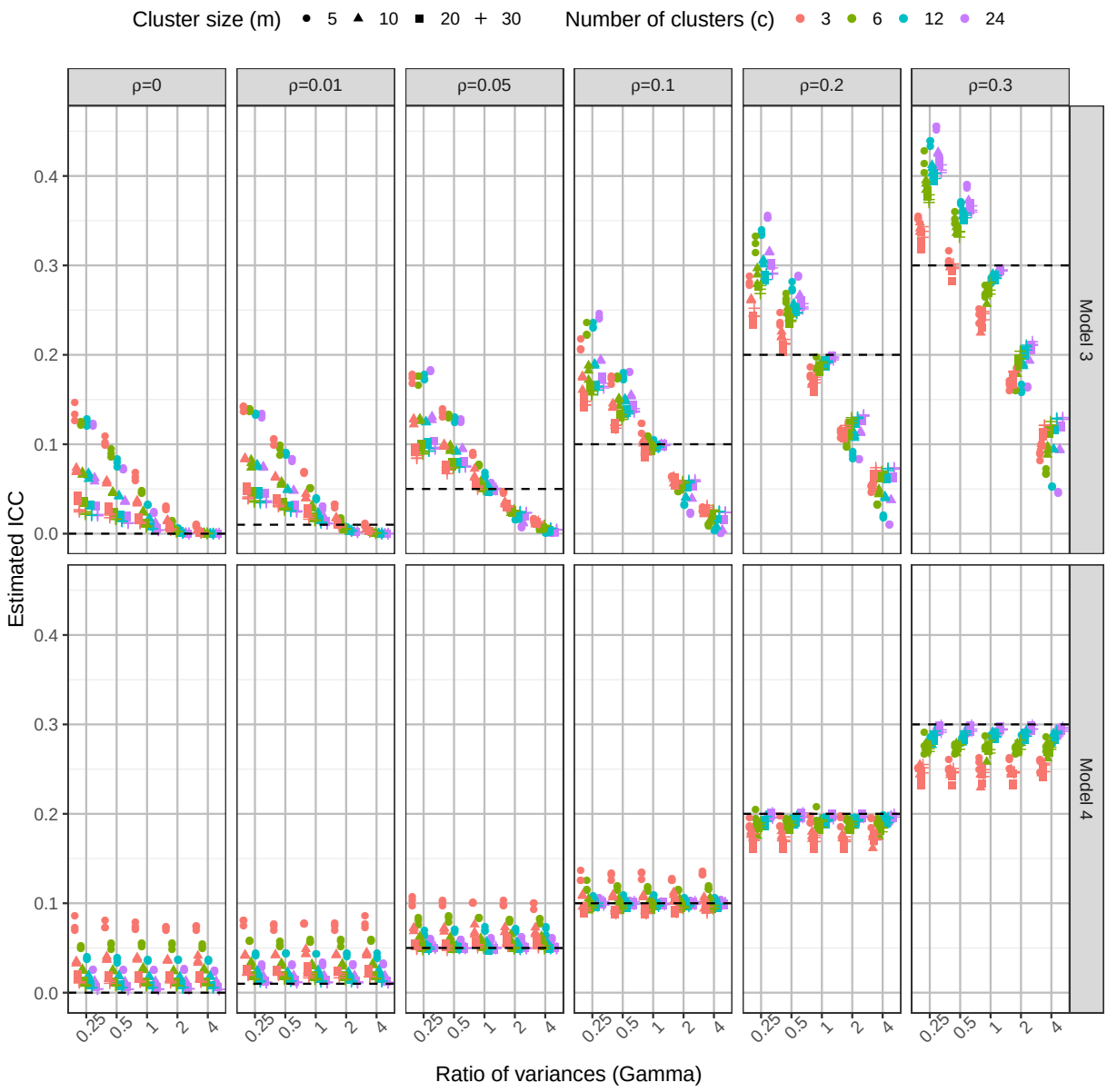


Figure 4.10: ICC estimation of heteroscedastic partially nested model, by γ, ρ, m and c



4.6.8 Summary of results

Simulation results are summarised in Table 4.6 presenting the performance of the simple linear regression model (model 1), homoscedastic partially nested mixed effects model (model 3) and heteroscedastic partially nested mixed effects model (model 4) under different design scenarios. Results from the fully clustered mixed effects models (model 2) are excluded from Table 4.6 as these are not recommended in any scenario regardless of the method used to impose clustering in the control arm. None of the fully clustered mixed effects models provided full control of the Type I error rates and the partially nested mixed effects models always outperformed them.

Table 4.6: Summary of simulation results in terms of Type I error ($\hat{\alpha}$) and ICC estimation ($\hat{\rho}$) for models 1, 3, and 4 split by ρ, m and c averaged over all γ.

ICC (ρ)	**Cluster size** (m)	**No. clusters** (c)	**Mean (SD)**				
			Model 1	**Model 3**		**Model 4**	
			$\hat{\alpha}$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\alpha}$	$\hat{\rho}$
0	5 - 10	3 - 6	0.049 (0.007)	0.025 (0.009)	0.047 (0.043)	0.026 (0.013)	0.047 (0.020)
		12 - 24	0.052 (0.007)	0.040 (0.010)	0.035 (0.040)	0.042 (0.009)	0.023 (0.010)
	20 - 30	3 - 6	0.050 (0.007)	0.023 (0.011)	0.014 (0.012)	0.024 (0.014)	0.013 (0.004)
		12 - 24	0.050 (0.007)	0.038 (0.010)	0.010 (0.011)	0.040 (0.009)	0.006 (0.002)
0.01	5 - 10	3 - 6	0.058 (0.007)	0.028 (0.010)	0.052 (0.043)	0.030 (0.016)	0.052 (0.017)
		12 - 24	0.055 (0.006)	0.041 (0.011)	0.041 (0.044)	0.043 (0.007)	0.029 (0.010)
	20 - 30	3 - 6	0.064 (0.015)	0.029 (0.010)	0.021 (0.016)	0.029 (0.016)	0.019 (0.003)
		12 - 24	0.066 (0.008)	0.044 (0.012)	0.017 (0.016)	0.046 (0.008)	0.013 (0.001)
0.05	5 - 10	3 - 6	0.072 (0.016)	0.031 (0.011)	0.077 (0.057)	0.031 (0.016)	0.079 (0.016)
		12 - 24	0.071 (0.012)	0.047 (0.012)	0.067 (0.061)	0.048 (0.008)	0.058 (0.007)
	20 - 30	3 - 6	0.120 (0.035)	0.041 (0.008)	0.051 (0.031)	0.039 (0.011)	0.052 (0.002)
		12 - 24	0.123 (0.032)	0.052 (0.017)	0.050 (0.036)	0.050 (0.006)	0.050 (0.001)
0.1	5 - 10	3 - 6	0.093 (0.024)	0.037 (0.007)	0.108 (0.068)	0.037 (0.012)	0.114 (0.011)
		12 - 24	0.092 (0.025)	0.050 (0.013)	0.103 (0.082)	0.050 (0.008)	0.100 (0.004)
	20 - 30	3 - 6	0.192 (0.058)	0.053 (0.011)	0.090 (0.046)	0.050 (0.012)	0.093 (0.004)
		12 - 24	0.185 (0.055)	0.055 (0.015)	0.097 (0.056)	0.050 (0.007)	0.099 (0.002)
0.2	5 - 10	3 - 6	0.136 (0.049)	0.047 (0.012)	0.174 (0.091)	0.044 (0.012)	0.187 (0.008)
		12 - 24	0.135 (0.047)	0.054 (0.011)	0.183 (0.113)	0.051 (0.005)	0.193 (0.004)
	20 - 30	3 - 6	0.301 (0.087)	0.060 (0.009)	0.169 (0.072)	0.057 (0.012)	0.177 (0.011)
		12 - 24	0.286 (0.077)	0.051 (0.011)	0.188 (0.084)	0.049 (0.006)	0.196 (0.003)
0.3	5 - 10	3 - 6	0.181 (0.068)	0.056 (0.011)	0.242 (0.108)	0.053 (0.010)	0.262 (0.012)
		12 - 24	0.177 (0.065)	0.054 (0.012)	0.268 (0.135)	0.050 (0.007)	0.288 (0.006)
	20 - 30	3 - 6	0.383 (0.092)	0.065 (0.010)	0.245 (0.090)	0.061 (0.011)	0.258 (0.017)
		12 - 24	0.368 (0.094)	0.051 (0.009)	0.278 (0.105)	0.050 (0.007)	0.292 (0.005)

*Model 1: simple linear regression; Model 3: homoscedastic partially nested mixed effects model; Model 4: heteroscedastic partially nested mixed effects model. Green highlighted $\leq$ than expected, red highlighted $>$ than expected.

4.7 Discussion

This chapter has investigated six modelling strategies for the analysis of pnRCTs with a continuous outcome. The simulation study showed that when analysing pnRCTs the use of the heteroscedastic partially nested mixed effects model for normally distributed outcome data (using Satterthwaite degrees of freedom) in general provides: unbiased effect estimates; maintains relatively good control of Type I error rates; and did not noticeably cause a reduction in power even with homoscedastic individual variances across arms. The heteroscedastic partially nested model takes account of the between-cluster variance (if present) and therefore provides valid inferences for the intervention effect. When using the partially nested mixed effects model, the method of classifying the non-clustered controls had a negligible impact on statistical inference under the simulation scenarios, agreeing with findings from analysis of four example pnRCTs by Flight et al. [45].

The findings were broadly similar to those of Baldwin et al. [122]. However, they did not assess the method of classifying the non-clustered controls or performance of models under small ICC ($\rho = 0.01$) which commonly occur in pnRCTs [45, 46, 48, 50, 122]. Unlike findings from Baldwin et al. [122], the Satterthwaite degrees of freedom correction did not fully control the Type I error rate in the simulations. The most discrepancy from the nominal level occurred when the ICC was small, ratio of individual variances $\gamma < 1$, and under small sample sizes.

It was illustrated that using a naïve linear regression model which ignores clustering in pnRCTs gives inflated Type I error rates and results in under coverage of confidence intervals when clustering of outcomes was present. When ICC $0.01 \leq \rho \leq 0.05$, which is typical in pnRCTs [45], ignoring clustering led to largely inflated Type I error rates using the linear regression model. A low ICC may still have a large impact, particularly when there are large cluster sizes.

When ICC was small and/or with very few clusters and small cluster sizes using the partially nested mixed effects models 3 and 4 resulted in deflated Type I error rates. These models correctly reflect the design of the trials; however, they can result in conservatism regarding the precision of estimates due to the bias in estimating the variance estimates when there are a small number of clusters. Consequently, using the partially nested mixed effects models with small ICC may make it difficult to detect differences between the trial arms when present.

Sanders [134] recommend evaluating whether ICC is significantly different from zero prior to selecting an analysis method. The use of significance testing for ICC and similarly testing for

heteroscedasticity are generally discouraged in agreement with Roberts and Roberts [48] and Donner and Klar [148]. These tests will generally lack power in a pnRCT and it is not the statistical significance of the ICC that matters but impact of the magnitude on inference. In general, the use of the partially nested models when analysing pnRCT trials is recommended, particularly if conservatism and an ICC estimate are desired. However, model choice decision and the requirement or not for conservatism needs to be considered in the context of the specific trial setting.

Similar to cRCTs [40], in a pnRCT increasing the number of clusters rather than increasing the cluster size had a greater increase in power for a fixed total sample size. The simulation results showed that this will also provide a more accurate estimation of the ICC. When the number of clusters is small, for example, three clusters in the intervention arm, the ICC estimation will likely be upwardly biased. With six clusters in the intervention arm, the ICC estimate was relatively unbiased once the true ICC ≥ 0.1. ICC estimation became consistent regardless of cluster size or true ICC only once there were 24 clusters in the simulation scenarios. This reflects findings from previous research that a large number of clusters are required to reliably estimate the size of clustering effects [149].

This chapter investigated the case of analysing pnRCTs under complete compliance. Non-compliance in the clustered arm of a pnRCT may occur when some participants randomised to a particular treatment group or therapist do not attend any sessions or receive treatment as part of different treatment group or therapist intended at randomisation. Consequently, non-complier outcomes may be assumed independent if they do not receive the clustered intervention. Schweig and Pane [135] describe and compare models for pnRCTs with non-compliance using a simulation study. They argue that an unbiased intention-to-treat (ITT) estimate under non-compliance on a pnRCT may be obtained using a Complier Average Causal Effects (CACE) model. This method involves estimating the treatment effect for compliers and scaling this CACE effect estimate by the proportion of compliers to provide an ITT effect estimate. The issues posed by non-compliance warrant further investigation, considering a broader range of scenarios and investigating the degrees of freedom corrections for valid statistical inferences.

A wide variety of terminology is used in iRCTs with clustering, including partially nested, partially clustered, multi-level, and individually randomized group intervention. A more consistent use of terminology would reduce confusion, improve reporting and make finding relevant ICCs from previous trials easier. The terminology partially nested randomised controlled trial is

suggested to describe an iRCT with clustering in one arm.

4.7.1 Limitations

All the mixed effects models assume that the cluster effects follow a Normal distribution. This may not be a valid assumption, for example, when there are a small number of clusters.

The simulations used fixed cluster sizes. In practice, cluster size may vary, causing a loss in efficiency when estimating the intervention effect. A Monte Carlo simulation study by Candel and Van Breukelen [128] found the efficiency loss in the intervention effect estimate was rarely more than 10%, requiring recruitment of 11% more clusters for the intervention arm and 11% more individuals for the control arm. The loss of efficiency in the intercept variance reached up to 15%, requiring 19% more clusters in the clustered arm, and no additional recruitment in the control arm. Additionally, it has been shown in cluster trials if the coefficient of variation in cluster size is small, less than 0.23, then the correction on sample size is negligible [150]. It should be noted that cluster sizes are likely more similar in the group administered treatment compared to trials which impose clustering by being treated by the same care provider [48].

Throughout the simulations it was assumed that there was no effect of clustering in the control arm, this may not strictly be true in practice. In healthcare intervention trials, a commonly used control intervention is 'care as usual'. This type of control may induce some form of low-level clustering, for instance, treatment by a healthcare practitioner. If the same practitioner treats numerous individuals, it can be assumed, in the same sense as was done for the intervention arm, that these individuals are clustered and include this in the modelling procedure. However, this information is often not available in trial data and is not unique to pnRCTs.

Partially nested trials pose a number of challenges, in particular, the issue of internal validity [49]. The grouping of individuals as part of the delivery of a treatment may affects the outcome. However, taking a pragmatic viewpoint, the grouping is considered as part of the treatment as a whole if this is reflective of treatment delivery in real-world practice. In addition, if the ungrouped controls are the true comparison in real life a pnRCT design will provide external validity.

4.7.2 Model fit code for Stata, R, and SAS

It is still common for clustering in pnRCTs to be ignored in the design and analysis stages. To encourage the use of analysis methods which take account of clustering, Table 4.7 presents commands to implement the homoscedastic and the heteroscedastic partially nested models in three commonly used statistical packages, Stata, R, and SAS, along with the degrees of freedom correction options (where available). The homoscedastic model is included for clarity and where there is a strong priori belief of homoscedasticity it may be more suitable to use. Both the homoscedastic and heteroscedastic models can be fitted in Stata and SAS with the option of the Satterthwaite degrees of freedom correction. Both models can be fitted in R, however, at the time of submission of this thesis I was not aware of a method that allows the fitting of the heteroscedastic partially nested model (model 4) using the Satterthwaite degrees of freedom correction. Instead, it is possible to use bootstrapping to obtain confidence intervals and the likelihood ratio test to obtain p-values for the effect estimate.

Table 4.7: Stata, R and SAS model fitting commands for the partially nested models

Software	**Homoscedastic partially nested model**	**Heteroscedastic partially nested model**
Stata	*mixed* command with *dfmethod(sat)* option	mixed command with dfmethod(sat) option and residuals(independent, by(intervention))
Example code	mixed y treat || cluster:treat, nocons reml dfmethod(sat)	mixed y treat || cluster: t, nocons reml residuals(independent, by(treat)) dfmethod(sat)
R*	*lmer* function from *lme4* package, and package *lmerTest* for Satterthwaite df	*lme* function from *nlme* package with *weights=varIdent(form=∼1|treat)* option. No option for Sattertwaite df.
Example code	lmer(y ∼ x + (0 + treat|cluster))	lme(fixed=y ∼ treat, random=∼(0+treat)|cluster, weights=varIdent(form=∼1|treat), method="REML")
SAS	*proc mixed* command with the option *ddfm=sat*	proc mixed command with the option ddfm=sat and repeated / group = treat;

Example code	proc mixed covtest;
	class cluster treat;
	model y = treat / solution ddfm=sat;
	random intercept treat / subject = cluster;
	run;
		proc mixed covtest;
		class cluster treat;
		model y = treat / solution ddfm=sat;
		random intercept treat / subject = cluster;
		repeated / group = treat;
		run;

y = outcome, treat = intervention arm indicator, cluster = intervention cluster indicator. *At present cannot fit the heteroscedastic partially nested model with degrees of freedom correction in R. The function *lmer* in R does not allow for different variances for each level of a grouping factor.

4.8 Summary

This chapter has described the scenarios in which pnRCTs may arise, the reason clustering needs to be considered, and available analysis methods. Partially nested RCTs are increasingly used in complex intervention research. Ignoring clustering can lead to large inflations of the Type I error rates, even for small ICCs. When analysing a pnRCT it is recommended to use a heteroscedastic partially nested mixed effects model with corrected degrees of freedoms such as the Satterthwaite method, for continuous outcomes similar to those generated under the scenarios of the simulations study. The model is easy to implement in standard statistical software and does not cause a notable reduction in power under homoscedastic variances. The method used for classifying the non-clustered controls made a negligible impact on the results using the partially nested mixed effects model. With few clusters, small cluster sizes and small ICC, the partially nested model underestimated Type I error rates and gave largely inflated ICC estimates, hence, for such designs there is no optimal model and we need to be cautious in model interpretation. Finally, in order to aid the design and prior selection of an appropriate analysis plan for pnRCTs, it is strongly recommended to report both estimated ICC and 95% confidence interval for primary and secondary endpoints when publishing trials results.

The next chapter expands upon findings from this chapter and the systematic review in chapter 3, the analysis of within-arm pnRCTs is investigated, pnRCTs with clustering of only some of the intervention arm based on intermediate response.

Chapter 5

Within-arm partially nested trials

5.1 Introduction

In chapter 4 it was demonstrated that it is possible to analyse pnRCTs appropriately parametrising the model with fixed and random effects. When there are a sufficient number of clusters, partially nested mixed effect models provide an unbiased effect estimate, maintaining nominal Type I error rates and account for the clustering in one arm providing ICC estimation. When clustering is only present for some in the intervention arm, termed within-arm pnRCTs herein, analysis which reflects the design of the trial becomes more challenging. This chapter expands upon chapter 3 and 4 and investigates the analysis of within-arm pnRCTs through a simulation study.

Proportionate interventions can induce a complicated hierarchical data structure. The systematic review in chapter 3 found that, similar to the E-SEE trial, trials of proportionate interventions often induce some form of clustered outcomes at different stages of the intervention. One or more of the intervention stages can result in the presence of possibly non-ignorable clustering [54]. This clustering may be due to therapist effects, for example therapists running the trauma focussed CBT sessions in step two of the stepped care trial by Salloum et al. [112], or group effects, for example the Incredible Years-Infant parenting groups in the E-SEE trial [3], or both the group and the group facilitator effects (if group facilitators run more than one parenting group).

When clustered outcomes are present for all individuals in the intervention arm, for instance when stage one induces clustering, these can be defined in the same manner as pnRCTs and

analysis models parametrised accordingly using partially nested mixed effects models. However, the clustering may not affect all participants in the intervention arm, and if so it is typically not at random (participants who receive the clustered intervention stage are those who did not respond to the previous intervention stage), and defined as post-randomisation clustering.

Within-arm partial nesting can occur from both the design of the intervention, such as proportionate interventions, or from non-compliance within the trial. This chapter considers the case related to design, non-compliance has been considered elsewhere [135]. A series of simulation studies are used to evaluate appropriate analysis methods for within-arm pnRCTs with continuous outcomes. A range of scenarios are evaluated including the effect of ICC, cluster size and the number of clusters.

5.2 Chapter aims

This chapter aims to evaluate the performance of commonly used analysis methods for within-arm pnRCTs to establish which methods are most appropriate and why. The specific objectives are to:

1. evaluate models for within-arm pnRCTs using linear regression and mixed effects models;
2. evaluate linear regression with robust/bootstrap standard errors to account for clustering;
3. quantify the impact that ignoring clustering can have on Type I error and precision;
4. provide recommendations for the most appropriate analysis methods for within-arm partially nested randomised controlled trial considering the ICC and cluster size.

5.3 Analysis methods for within-arm partially nested trials

This section presents and discusses the modelling approaches available to analyse within-arm pnRCTs, expanding on those presented in chapter 4 including: ignoring clustering altogether using linear regression; linear regression with cluster robust standard errors; linear regression with cluster bootstrap standard errors; and mixed effects model.

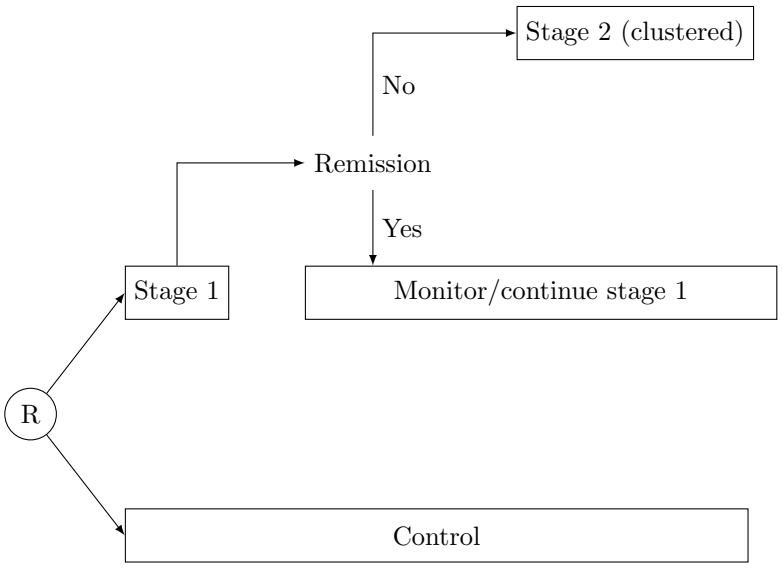
5.3.1 Within-arm partially nested trial design

In order to be able to evaluate the different methods of analysis, all models are tested on the same within-arm pnRCT trial scenarios. Attention is limited to a two arm, two stage intervention as follows:

1. first stage is delivered at the individual or universal level, for example, bibliotherapy for the treatment of depression. If individuals do not respond to stage one (based on pre-defined decision rules) they are offered,
2. second stage intervention which is delivered in a manner that results in intervention induced clustering of outcomes, for example, Cognitive Behavioural Therapy treatment by therapist for depression.

Figure 5.1 represents the trial design described above in a flow diagram form; this is a version of a stepped-care trial with two steps. The primary aim of a trial using this design (and primary analysis used to power such a trial) is typically to evaluate the effectiveness of the overall intervention being tested (comprised of stage one and two) compared to the control arm. The control is commonly care as usual or a comparator treatment. Therefore, the aim of the trial in Figure 5.1 would be to evaluate the overall effectiveness of the proportionate intervention compared to control, regardless of whether participants were only offered stage one, or were offered both stage one and two.

Figure 5.1: Diagram representing the simple proportionate intervention with clustering of outcomes at intervention stage 2.



There are specific issues that arise in the analysis of Figure 5.1. The outcome follow-up after

stage two is clustered due to intervention induced clustering of stage two. We ideally wish to adjust for this clustering in the analysis. However, not all individuals in the intervention arm get stage two and it is not randomly allocated, clustering is dependent on previous outcome after stage one intervention.

For the following modelling approaches, define y_{fij} as a continuous outcome at follow-up f for the ith individual and jth cluster, $i = 1, \ldots, N$ and $j = 1, \ldots, k$. The intervention indicators are t_{11j} is the trial arm indicator (0 = control, 1 = intervention, and all in the intervention arm receive stage one intervention) and t_{2ij} is the stage two indicator (0 = no stage two intervention received, 1 = stage two intervention received). Let the outcome follow-ups occur after stage one intervention, y_{1ij}, and after stage two intervention, y_{2ij}. Let θ denote the overall intervention effect and θ_1 and θ_2 denote the effect of stage one and stage two interventions, respectively. Error terms are again defined depending on the model, represented using ϵ, u and r, and β_0 is an intercept term.

5.3.2 Linear regression model and other naive analysis

One approach to analysis of within-arm pnRCTs (the one most commonly used to date and seen in the systematic review results of chapter 3) is to ignore the treatment induced clustering altogether. For instance, analysis could be done using a t-test, ANOVA, linear regression model, or a mixed effects model analysing the longitudinal follow-up data of patients by including a random effect for patient (accounts for within person correlation). The naïve linear regression model presented in equation 4.1 can be used to analyse the within-arm pnRCT.

Due to the design induced clustering we may wish to choose a more robust method than fitting OLS regression. Possible methods are described in the following sections.

5.3.3 Mixed effects regression

Chapter 4 recommended the use of a heteroscedastic partially nested mixed effects model. Following this, the initial modelling strategy was to try and extend and parametrise a mixed effects model for the within-arm pnRCT data, which would model the clustering for only those who receive the clustered intervention stage two. Schweig and Pane [135] formulated a similar mixed effects model to account for non-compliance in pnRCTs and estimate the intervention effect, this attributed random effects only to those individuals who received the cluster intervention

and suppressed random effects for those that did not receive the clustered intervention.

A mixed effects model for the analysis of within-arm pnRCT shown in Figure 5.1 would require a fixed intervention effect for intervention arm t_{1ij} (0 = control arm, 1 = intervention arm), individual random variation and an additional random effect for those who receive the clustered intervention. Hence, the random intercept term u_j represents the between cluster variation in the stage two intervention (clustered), where t_{2ij} is an indicator for stage two (0 = no stage two intervention, 1 = stage two intervention) and ϵ_{ij} are the individual level residuals in the control arm and r_{ij} are the individual level residuals in the intervention arm. Additional predictive covariates can also be added to the model in practice. The outcome at follow-up two ($f = 2$) for individual i in cluster j, $i = 1, \ldots, N$, $j = 1, \ldots, k$, is given by the model

$$
\begin{aligned}
y_{2ij} = \beta_0 + \theta t_{1ij} + u_j t_{2ij} + r_{ij}(1 - t_{1ij}) + \epsilon_{ij} t_{1ij}, \qquad (5.1)\\
u_j \sim & N(0, \sigma_u^2),\\
r_{ij} \sim & N(0, \sigma_r^2)\\
\epsilon_{ij} \sim & N(0, \sigma_\epsilon^2).
\end{aligned}
$$

The mixed effect model in equation 5.1 is expected to result in a biased estimate of the overall intervention effect. The rationale is based on the parametrisation of the model and the structure of the data. Firstly, the model includes a random intercept at the cluster level (u_j) for the individuals who receive stage two and no fixed effect for stage two intervention. Secondly, the clusters or treatment groups in within-arm pnRCTs are not organised randomly. Only those with interim measures below a threshold (non-responders) are assigned to groups. This effectively down-weights the contribution of these participants when allowing for non-dependence (with a random effect) because these participants are more similar even before they receive any stage two intervention and hence the estimate of the mean effect estimate is biased.

To account for the intervention effect and clustering effect of stage two both a fixed and random effect would need to be directly included in a mixed effects model. A partially nested mixed effects model accounting for clustering of stage two treatment and adding a fixed effect for stage one and one for patients who receive stage two using the following model

$$y_{2ij} = \beta_0 + \theta_1 t_{1ij} + \theta_2 t_{2ij} + u_j t_{2ij} + r_{ij}(1 - t_{1ij}) + \epsilon_{ij} t_{1ij}, \tag{5.2}$$

$$u_j \sim N(0, \sigma_u^2),$$

$$r_{ij} \sim N(0, \sigma_r^2),$$

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2).$$

Equation 5.2 will also provide an estimate of the ICC for stage two intervention, t_{2ij}, using $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_\epsilon^2)$. However, as stage two is not given at random it is important to note this will produce biased intervention effect estimates when randomisation occurs only at baseline. This is confirmed using a small simulation study described in section 5.4. Although the analysis method in equation 5.2 will typically not be powered for during the design phase, it will provide an estimate of the ICC. The ICC may be of descriptive interest to evaluate how the outcomes are correlated within clusters. Equation 5.2 can be considered a heteroscedastic partially nested mixed effects model accounting for clustering of stage two treatment and adding subgroup covariates.

Preliminary work to evaluate bias of the mixed effect models given in equations 5.1 and 5.2 was undertaken using simulations of a number of different scenarios, with a single large simulated dataset for each scenario. This work showed the models to be largely biased when estimating the intervention effect estimate. The preliminary work coupled with the statistical properties of the mixed effect models in equations 5.1 and 5.2 suggest that these models are inappropriate and largely bias the effect estimate. This is formally confirmed using a small simulation study described in section 5.4. Only a small simulation study is required to demonstrate the large-sample bias of the effect estimator identified in the preliminary work [151]. Guidance for the simulation study is taken from Morris et al. [151] to demonstrate a large-sample bias of an estimator. The sample size of the simulated dataset is large to show that the effect estimator is largely bias from its true value and two scenarios are chosen (one under the null and the other under the alternative hypothesis).

So far this chapter has explored how to construct a mixed effects model for within-arm pnRCTs. The following section will discuss alternative analysis which may be able to account for clustering and provide an estimate of the interventions effect (θ).

5.3.4 Linear regression with robust standard errors

Robust standard errors are a method of estimating standard errors in linear regression analysis. They are robust to minor concerns about the OLS regression assumptions, such as concerns about normality, heteroscedasticity, or having some observations with large residuals, leverage or influence. The robust variance estimator is also called the sandwich estimator, Huber-White sandwich estimators are used in Stata [142].

The robust sandwich variance estimator of the parameter estimate θ is given by

$$V_{Rob}\left(\hat{\theta}\right) = (\boldsymbol{X'X})^{-1}\left(\sum_{i=1}^{N} e_i^2 \boldsymbol{x}_i' \boldsymbol{x}_i\right)(\boldsymbol{X'X})^{-1}$$

where $e_i = y_i - x_i\theta$, x_i is the ith row of $\boldsymbol{X}$ which is an $n \times p$ matrix of covariates, n is the number of observations and p the number of parameters.

5.3.4.1 Cluster robust

When using the robust estimator of the variance it is possible to relax the assumption of independent observations by using cluster-robust standard errors. Cluster robust standard errors require that observations are independent across clusters but not necessarily within-clusters. They do not require specification of a model for within-cluster error correlation, but cluster-robust standard errors do assume "that the number of clusters, rather than just the number of observations, goes to infinity" [58, p.318].

The cluster robust variance estimator is given by

$$V_{ClusRob}\left(\hat{\theta}\right) = (\boldsymbol{X'X})^{-1}\left(\sum_{j=1}^{k} u_j' u_j\right)(\boldsymbol{X'X})^{-1} \tag{5.3}$$

where j is the cluster identifier, $u_j = \sum e_{ij}x_{ij}$ and n_c is the number of clusters. The clustered robust variance formula is that of the robust (unclustered) estimator with the individual $e_{ij}x_{ij}$'s replaced by their sums over each cluster.

Cluster-robust standard errors were incorporated into Stata by Rogers et al. [152]. They can be implemented in Stata using the `vce(cluster clustervar)` command, where `clustvar` specifies to which cluster each observation belongs, for example, `vce(cluster cluster1)` in data with observations in the clusters defined by `cluster1`. This affects the standard errors and vari-

ance–covariance matrix of the estimators but not the estimated regression coefficients. Using the cluster estimator is expected to result in more conservative errors when positive correlation exists [152].

5.3.5 Linear regression with bootstrap standard errors

Bootstrapping was first proposed by Efron and Tibshirani [153], it can used for estimating standard errors and other statistical measures whilst making few assumptions. It is a non-parametric approach using computational re-sampling techniques rather than formulae.

The basic concept of bootstrapping is presented in the following. Let $\hat{\theta}$ again denote the estimate of our parameter from the original sample. A nonparametric bootstrap procedure can be used to calculate standard errors using the following steps [154]:

1. Draw B independent random bootstrap samples, each consisting of n data values drawn with replacement from the sample dataset $\boldsymbol{x}$, generating B pseudo-samples (for estimating standard error B is usually in the range of 25-200 [153]);

2. Estimate the desired statistic corresponding to each of these bootstrap samples, which forms the sampling distribution of $\widehat{\theta}$;

3. Estimate standard error from the sample standard deviation of the sampling distribution using

$$\widehat{se_B} = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b - \overline{\theta} \right)^2 \right\}^{1/2}$$

where $\hat{\theta}_b,\ b = 1 \ldots, B$ denote the B estimates of β from the bootstrap samples, and $\overline{\theta^*} = (1/B) \sum_{b=1}^{B} \hat{\theta}_b$ is the mean of the estimates across the B bootstrap samples. Though the bootstrap estimates of the statistic are used to estimate the standard error, the actual estimated value used for the statistic is still the original observed value $\hat{\theta}$ computed using the original observations.

Bootstrap confidence intervals can be calculated using the Normal approximation or using the percentile method. The percentile method is constructed from the sampling distribution of the B bootstrap estimates of the parameter of interest, reading off the 2.5% and 97.5% percentiles of the distribution gives you the 95% confidence interval. Though an obvious choice for the confidence interval the percentile method can be biased, hence, it is generally not recommended

as the best method [153]. This bias can be estimated and corrected for using the bias corrected and accelerated (BC_a) method [153, 154]. Bootstrapped confidence intervals may, therefore, be asymmetric and be better able to deal with skewed data.

5.3.5.1 Cluster bootstrap

Bootstrap resamples from the original data and assumes the data are independently and identically distributed. If clustering is present and we ignore the dependence of the data by resampling at the level of the individual observation we cannot preserve the distribution of the estimate of the desired statistic, this can result in under estimated standard errors and give incorrect inferences [155].

It is possible to bootstrap at the cluster level. We can recognize that observations within a cluster are not independently distributed by, instead of drawing individual observation units with replacement, each sample drawn during each replication is a bootstrap sample of clusters drawn with replacement. Each bootstrap resample will have exactly k clusters. Some of the original clusters will not appear at all while other original clusters may be repeated in the bootstrap sample two or more times [58]. Cluster bootstrap standard errors also assume independence across clusters of observations.

Although the majority of the data in a within-arm pnRCT are independent, correlated outcomes exist for participants that receive the clustered stage two treatment. Pals et al. [43, p.1421] notes that "bootstrap standard errors are correct in the context of within-group correlation only when bootstrapping is done at the group level". Hence, it was assumed that the simple bootstrap ignoring the cluster level would be invalid. If the ICC is positive then the cluster bootstrap standard errors compared to the usual bootstrap standard errors will result in a larger standard error and consequently larger confidence intervals.

Bootstrap standard errors are included in many statistical packages with the option of bootstrapping at the cluster level. In Stata bootsrap standard errors of the parameter estimates can be caluclated using the `vce(bootstrap)` command. Again, issues arise when there are a small number of clusters. Bootstrap cluster confidence intervals may have poor coverage properties when there are a small number of clusters [155]. It is also possible to re-sample separately from each stratum (using the strata option in Stata).

5.4 Simulation study methods

5.4.1 Overview

A simulation study was undertaken to address the aims of this chapter, to evaluate the analysis models for within-arm pnRCTs presented in section 5.3 and summarised in Table 5.2. As in chapter 4 the study utilised guidance on design, conduct and reporting of simulation studies [144, 145].

5.4.2 Software

Software used for the simulations was the same as that used in chapter 4: simulations in Stata [142] and graphs produced using ggplot2 [146] in R [147]. See Appendix B.2 for example simulation code.

5.4.3 Data generating mechanisms

Data were simulated to replicate a parallel within-arm pnRCT with an unclustered control arm and a two stage treatment in the intervention arm (randomised on a 1:1 basis), with clustering in the second stage treatment and a continuous outcome. Data were simulated under various design scenarios and under both the null and alternative hypotheses. Data generating mechanisms are explained in more detail in this section.

Data were simulated from the following model with the intercept set to zero and trial arm allocation denoted by t_1 ($t_1 = 0$ for control arm, $t_1 = 1$ for intervention arm). All in the intervention arm receive stage one, $t_1 = 1$. Fifty percent of the intervention arm receive stage two, based on outcome after t_1 $t_2 = 1$ if $y_1 < \theta_1$, ($t_2 = 0$ no stage two intervention, $t_2 = 1$ stage two intervention). The outcome y was simulated after t_1 and after t_2, y_1 and y_2 respectively.

The model is:

1. For the control arm ($t = t_1 = t_2 = 0$):

$$y_{1ij} = y_{2ij} = r_{ij} \tag{5.4}$$

2. For the intervention arm:

(a) Stage one treatment ($t_1 = 1$)

$$y_{1ij} = \theta_1 + e_{ij} \tag{5.5}$$

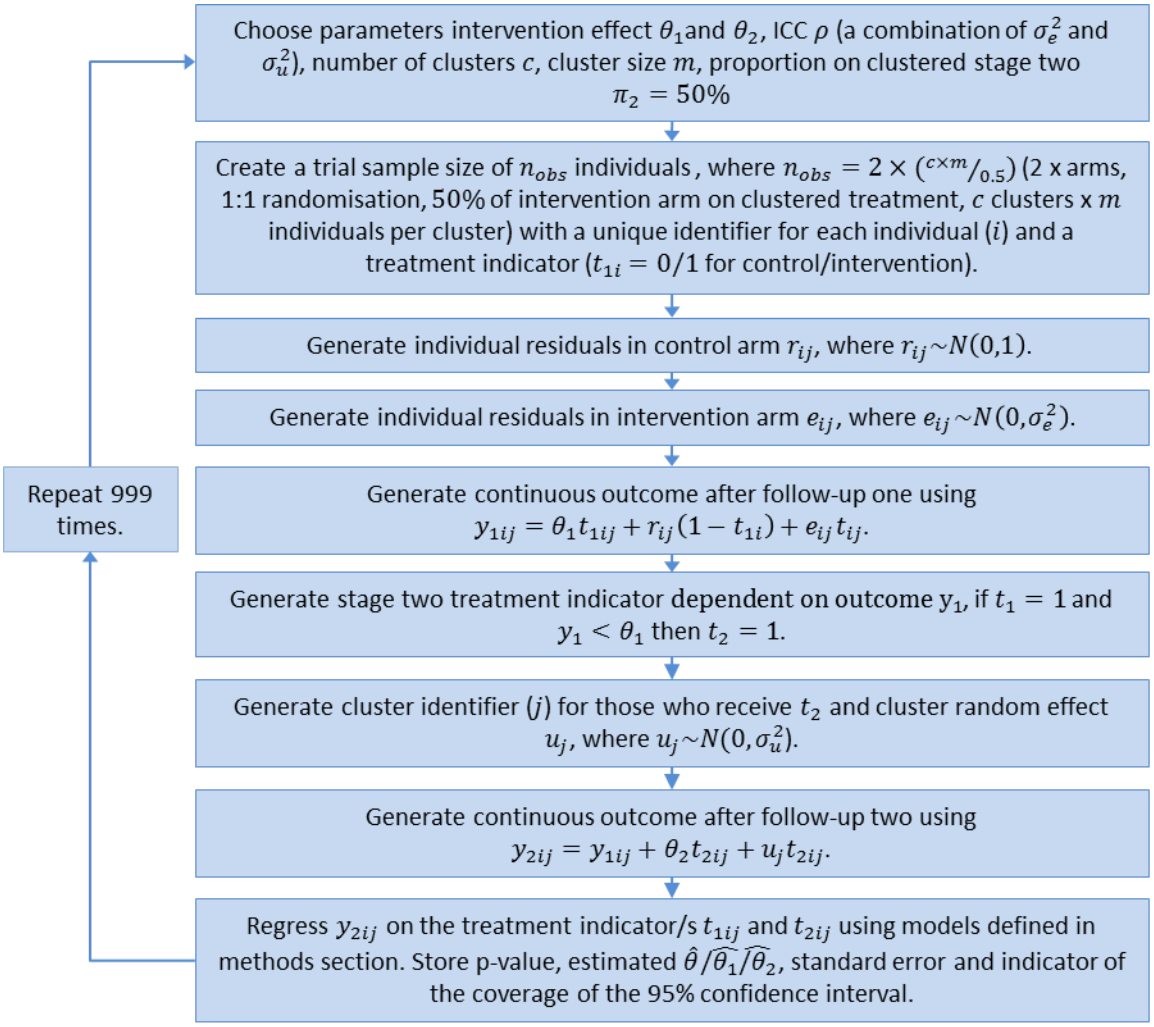
(b) Stage two treatment ($t_2 = 1$). If $y_1 < \theta_1$ (the non-responders) then $t_2 = 1$ and the outcome

$$y_{2ij} = \theta_1 + \theta_2 + u_j + e_{ij} \tag{5.6}$$

where $r_{ij} \sim N(0, \sigma_r^2)$ $u_j \sim N(0, \sigma_u^2)$, $e_{ij} \sim N(0, \sigma_e 2)$. Values of σ_e^2 and σ_u^2 were chosen to simulate the ICC, ρ, of clustered stage two intervention to be approximately 0.01, 0.05, 0.1, and 0.2. The variances $\sigma_e^2 + \sigma_u^2 = 1$ and the individual variance $\sigma_r^2 = 1$ for all scenarios.

Full simulation study steps, including the data generating process and modelling, are presented in Figure 5.2. An additional error term could be added for the time effect thus representing the repeated measures design, the error terms in the above equations would need to be changed so that the simulation still induces an ICC of ρ.

Figure 5.2: Flowchart representing the simulation study steps under null hypothesis.



5.4.4 Scenarios to be investigated

Simulation scenarios are presented in Table 5.1. For each of the total 98 scenarios 1,000 datasets were generated.

The bias of the mixed effect models were evaluated using simulated within-arm pnRCT data for two scenarios, one under the null hypothesis and one scenario under the alternative hypothesis. Only two scenarios were simulated as the parametrisation of the model and preliminary investigations showed the mixed effects model to be largely biased, the aim here was to demonstrate (large-sample) bias. It was decided that a large simulated dataset would provide this (number clusters was chosen to be 50 to ensure there were no issues with small number of clusters).

The other objectives of this chapter were to evaluate whether linear regression (with OLS standard errors, cluster robust standard errors or cluster bootstrap standard errors) in terms of bias and confidence interval coverage rates as well as quantifying the impact of ignoring clustering.

Simulation values were chosen based on findings from the systematic review in chapter 3 and some constraints of the data generating model. For all simulations equal allocation to two treatment arms was assumed and the proportion of the intervention arm who receive the clustered stage two treatment patients (π_2) was simulated as approximately 50%. The proportion of participants that receive the second clustered intervention stage will vary slightly based on the response to the first stage and the uptake of the intervention being offered. Due to the proportionate design of this type of trial the sample size of those who receive the second stage treatment is smaller than the sample size at randomisation. As $y_1|t_1 = 1 \sim N(\theta_1, \sigma_e^2)$ choosing $t_2 = 1$ for those with $y_1 < \theta_1$ we expect half the intervention arm to receive $t_2 = 1$. The overall intervention effect θ is made up of θ_1 and θ_2. Under H_A $\{\theta_1, \theta_2\} = \{0.25, 0.5\}$ and the expectation of the overall intervention effect is $\theta = 0.5$. The cluster size, number of clusters, ICC of the clustered stage two intervention and intervention effect were controlled and varied across scenarios. A description of the scenarios used are summarised in Table 5.1.

Table 5.1: Simulation scenarios for within-arm pnRCT

Parameter	**Notation**	**Value**
Investigation the bias of mixed effects model (two scenarios) under H_0 and H_A (2 scenarios)		
Proportion in stage two	$\pi_2\%$	50%
No. clusters	c	50
Cluster size	m	10
ICC*	ρ	0.05
Intervention effect	θ made up of θ_1, θ_2	Under H_0 $\{\theta_1, \theta_2\} = \{0, 0\}$ or Under H_A $\{\theta_1, \theta_2\} = \{0.25, 0.5\}$
Investigating bias and coverage of linear regression models (OLS, cluster robust and cluster bootstrap) under H_0 and H_A (96 scenarios)		
Proportion in stage two	$\pi_2\%$	50%
No. clusters	c	5, 10, 20
Cluster size	m	3, 6, 12, 24
ICC*	ρ	0.01, 0.05, 0.1, 0.2
Intervention effect	θ made up of θ_1, θ_2	Under H_0 $\{\theta_1, \theta_2\} = \{0, 0\}$ or Under H_A $\{\theta_1, \theta_2\} = \{0.25, 0.5\}$

5.4.5 Methods

Table 5.2 presents a summary of the analysis models fitted during the simulation study. Unclustered participants were treated as singleton clusters based on findings from the simulation study in chapter 4.

Table 5.2: Models used for the analysis of simulated within-arm partially nested trials

	Model description	Statistical model	Cluster definition
M1	Mixed effect model with single fixed effect	$y_{ij} = \beta_0 + \theta t_{1ij} + u_j t_{2ij}$ $+r_{ij}(1 - t_{1ij}) + \epsilon_{ij} t_{1ij}$, $u_j \sim N(0, \sigma_u^2)$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$	Cluster and singleton clusters for unclustered
M2	Mixed effect model with two fixed effects	$y_{ij} = \beta_0 + \theta_1 t_{1ij} + \theta_2 t_{2ij} + u_j t_{2ij}$ $+r_{ij}(1 - t_{1ij}) + \epsilon_{ij} t_{1ij}$, $u_j \sim N(0, \sigma_u^2)$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ $r_{ij} \sim N(0, \sigma_r^2)$	Cluster and singleton clusters for unclustered
L1	Linear regression (ignore clustering)	$y_i = \beta_0 + \theta t_{1i} + \epsilon_i$, $\epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2)$	Not applicable
L2	Linear regression with cluster robust SEs	$y_i = \beta_0 + \theta t_{1i} + \epsilon_i$, ϵ_i see section 5.3.4	Cluster and singleton clusters for unclustered
L3	Linear regression with cluster bootstrap SEs	$y_i = \beta_0 + \theta t_{1i} + \epsilon_i$, ϵ_i see section 5.3.5	Cluster and singleton clusters for unclustered*

*Resampling was stratified on t_2, whether a participant received stage two intervention or not.

5.4.6 Estimand

The estimand of interest was the overall intervention effect θ, and the intervention effect of stage one and stage two θ_1 and θ_2.

5.4.7 Performances measures

The following performance measures were used:

- For models M1, L1-L3, bias of the average intervention effect estimate for those in the intervention arm compared to control, regardless of whether they received stage two treatment or not using Bias $= E(\widehat{\theta}) - \theta$. For model M2 bias of intervention effect estimates of stage one and stage two using Bias $= E(\widehat{\theta_1}) - \theta_1$ and $= E(\widehat{\theta_2}) - \theta_2$.
- For models M1, L1-L3, mean square error of the intervention effect estimate using $E[(\widehat{\theta} - \theta)^2]$. For models M2 mean square error of the intervention effect estimates using $E[(\widehat{\theta_1} - \theta_1)^2]$ and $E[(\widehat{\theta_2} - \theta_2)^2]$.

- Type I error rate: proportion of simulations in which the p-value < 0.05 when the null hypothesis is true, true intervention effect $\theta = 0$.
- Coverage of the 95% confidence intervals of the intervention effect estimate: proportion of simulations that the obtained 95% confidence interval contains the true intervention effect θ when the alternative hypothesis is true (using Bias corrected confidence intervals for the cluster bootstrap standard errors).

5.5 Results: Mixed effects model

5.5.1 Bias, mean square error and coverage

Results of the bias and MSE for the overall intervention effect $\hat{\theta}$ from mixed effects model M1 with a single fixed effect are provided in Table 5.3. The model included a fixed effect for the intervention arm and random effect representing a random intercept for the clustering of stage two. The mixed effect model M1 (equation 5.1) resulted in a bias of the intervention effect estimate. The model largely overestimated the intervention effect, bias of the intervention effect estimate was 0.739 when $\theta = 0$ and 0.464 when $\theta = 0.5$, (equating to 73% and 93% percentage bias). The corresponding confidence intervals did not include the true intervention effect θ.

Table 5.3: Results of simulation investigating the bias of mixed effects model M1 with a single fixed effect for θ, under H_0 and H_A with ICC of t_2 $\rho = 0.05$, $m = 10$ and $c = 50$

Data generating model	**Bias**	**MSE**
$\theta = 0$	0.739	0.548
$\theta = 0.5$	0.464	0.217

Results of the bias and MSE for the intervention effect $\hat{\theta_1}$ and $\hat{\theta_2}$ from mixed effects model M2 with two fixed effects are provided in Table 5.4. The mixed effect model M2 resulted in bias of both intervention effect estimates $\hat{\theta_1}$ and $\hat{\theta_2}$.

Table 5.4: Results of simulation investigating the bias of mixed effects model M2 with two fixed effect for θ_1 and θ_2, under H_0 and H_A with ICC of t_2 $\rho = 0.05$, $m = 10$ and $c = 50$

Data generating model	$\hat{\boldsymbol{\theta_1}}$		$\hat{\boldsymbol{\theta_2}}$	
	Bias	**MSE**	**Bias**	**MSE**
$H_0 : \theta = 0$, $\{\theta_1, \theta_2\} = \{0, 0\}$	0.789	0.624	-1.580	2.497
$H_A : \theta = 0.5$, $\{\theta_1, \theta_2\} = \{0.25, 0.5\}$	1.039	0.624	-1.079	2.494

5.6 Results: Linear regression models (OLS, cluster robust and cluster bootstrap standard errors)

5.6.0.1 Bias

All three linear regression models (OLS, cluster robust and cluster bootstrap) compared in this simulation study led to unbiased estimates of the intervention effect. The maximum absolute bias of the intervention effect was $|0.015|$ for all models. Bias of the intervention effect was not affected by the analysis model used, the ICC or the proportion in stage two.

5.6.1 Mean square error

MSE did not differ by model. The MSE of the intervention effect had a mean of 0.002 (SD 0.073).

5.6.2 Type I error

Plots of the mean Type I error rates split by model and the ICC (ρ) are presented in Figure 5.3. As expected the linear regression model which ignores clustering had inflated Type I error rates, with Type I error rate affected by ICC (ρ), number of clusters (c), and cluster size (m). Although the inflation was minimal when ICC was small, mean Type I error when $\rho = 0.05$ was 0.057 (SD 0.008) and mean Type I error when $\rho = 0.01$ was 0.051 (SD 0.005).

The linear regression model with both cluster robust and cluster bootstrap standard errors (models 2 and 3) resulted in biased Type I error rates. The cluster robust standard errors over corrected the Type I error and the cluster bootstrap standard errors resulted in inflated Type I errors.

Figure 5.3: Type I error rate of linear regression model for analysis of within-arm partially nested trials: using OLS standard errors (model L1), cluster robust standard errors (model L2), cluster bootstrap standard errors (model L3), by ρ

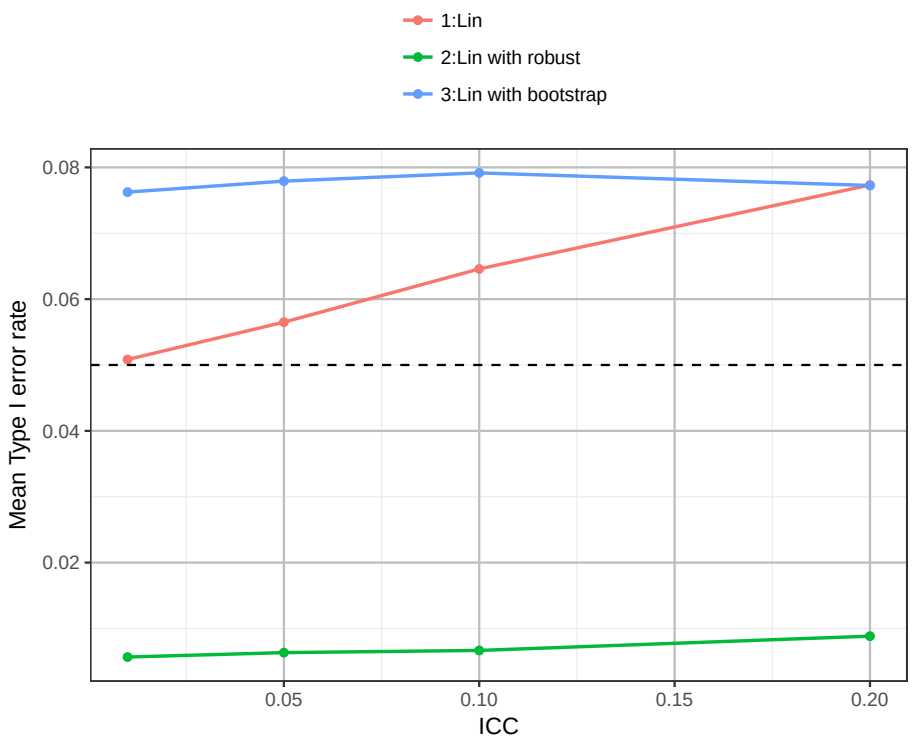
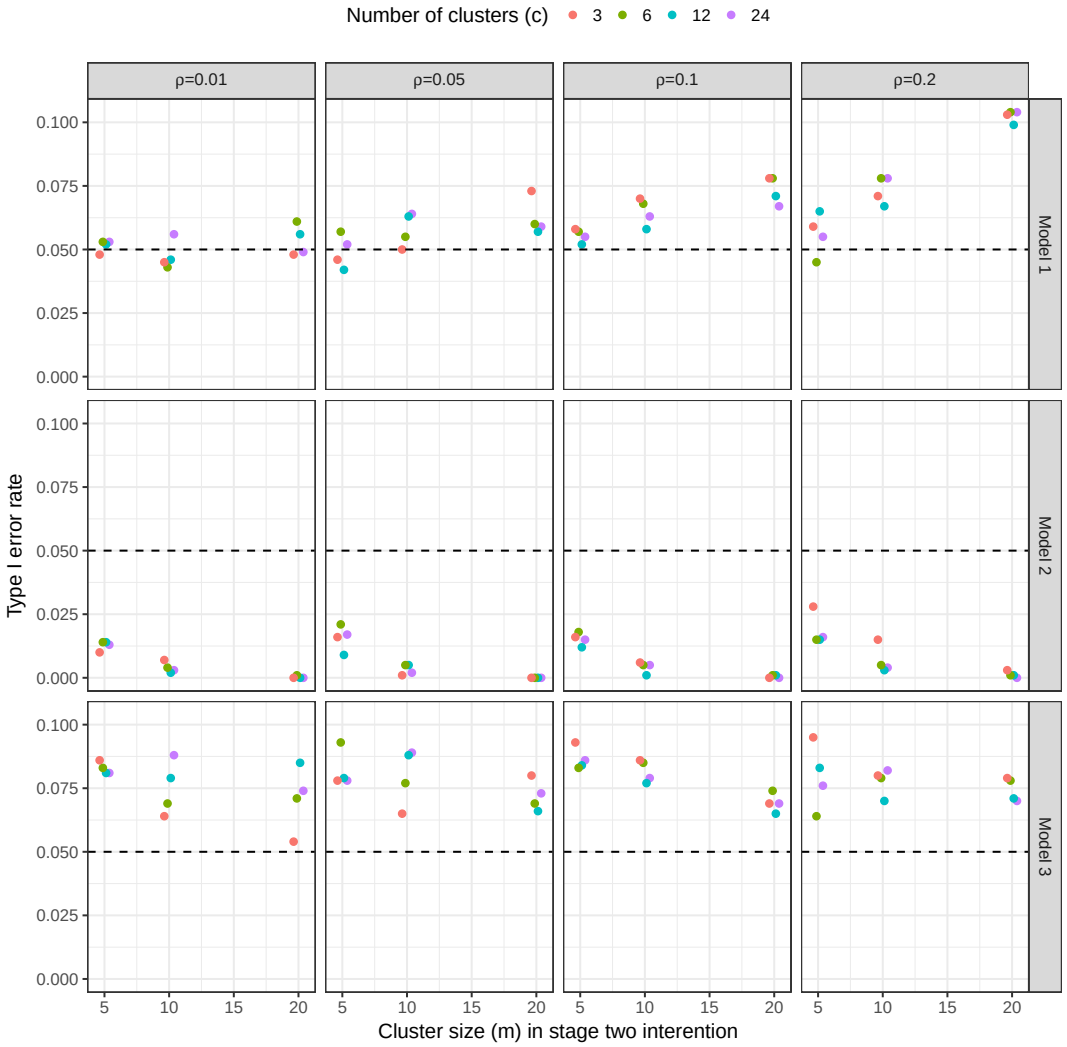


Figure 5.4 shows the Type I error rates for the linear regression model with OLS standard errors (model L1), cluster robust standard errors (model L2) and with cluster bootstrap standard errors (model L3) by ICC (ρ), number of clusters (c), and cluster size (m). When using the linear regression model with cluster robust standard errors Type I error rates were reduced from the nominal 5% level under all ICC values. In contrast, when using the linear regression model with cluster bootstrap standard errors Type I error rates were inflated from the nominal 5% level under all ICC values. Higher ICC values resulted in higher Type I error rates in each model. The impact of ignoring clustering (model L1) depends on both ICC (ρ) and cluster size (m).

Figure 5.4: Type I error rate of linear regression model for analysis of within-arm partially nested trials: using OLS standard errors (model L1), cluster robust standard errors (model L2), cluster bootstrap standard errors (model L3), by ρ, c, and m



5.6.3 Coverage

Plots of the mean coverage of the 95% confidence intervals of the intervention effect estimate split by model and the ICC (ρ) under the alternative hypothesis are presented in Figure 5.5. The linear regression (which ignores clustering) resulted in under coverage, with coverage affected by ICC (ρ) and cluster size (m). The coverage rate of the 95% confidence intervals decreased as ICC (ρ) increased.

Linear regression with either cluster robust or cluster bootstrap standard errors (models 2 and 3) resulted in biased coverage rates. Cluster robust standard errors resulted in coverage rates above the nominal 95% level. In contrast, the cluster bootstrap standard errors resulted in coverage rates below the nominal 95% level. Figure 5.6 presents the coverage rates in more detail for the linear regression model with OLS standard errors (model L1), cluster robust standard errors

(model L2) and with cluster bootstrap standard errors (model L3) by ICC (ρ), number of clusters (c), and cluster size (m).

Figure 5.5: Coverage rate of 95% confidence intervals of linear regression model for analysis of within-arm pnRCTs: using OLS standard errors (model L1), cluster robust standard errors (model L2), bootstrap standard errors (model L3), by ρ

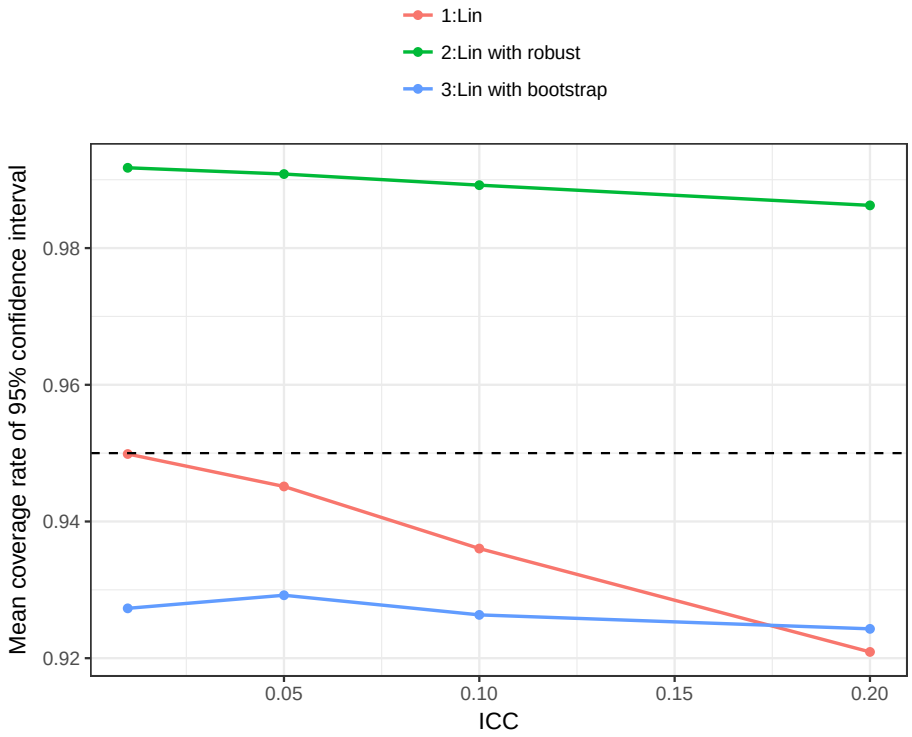
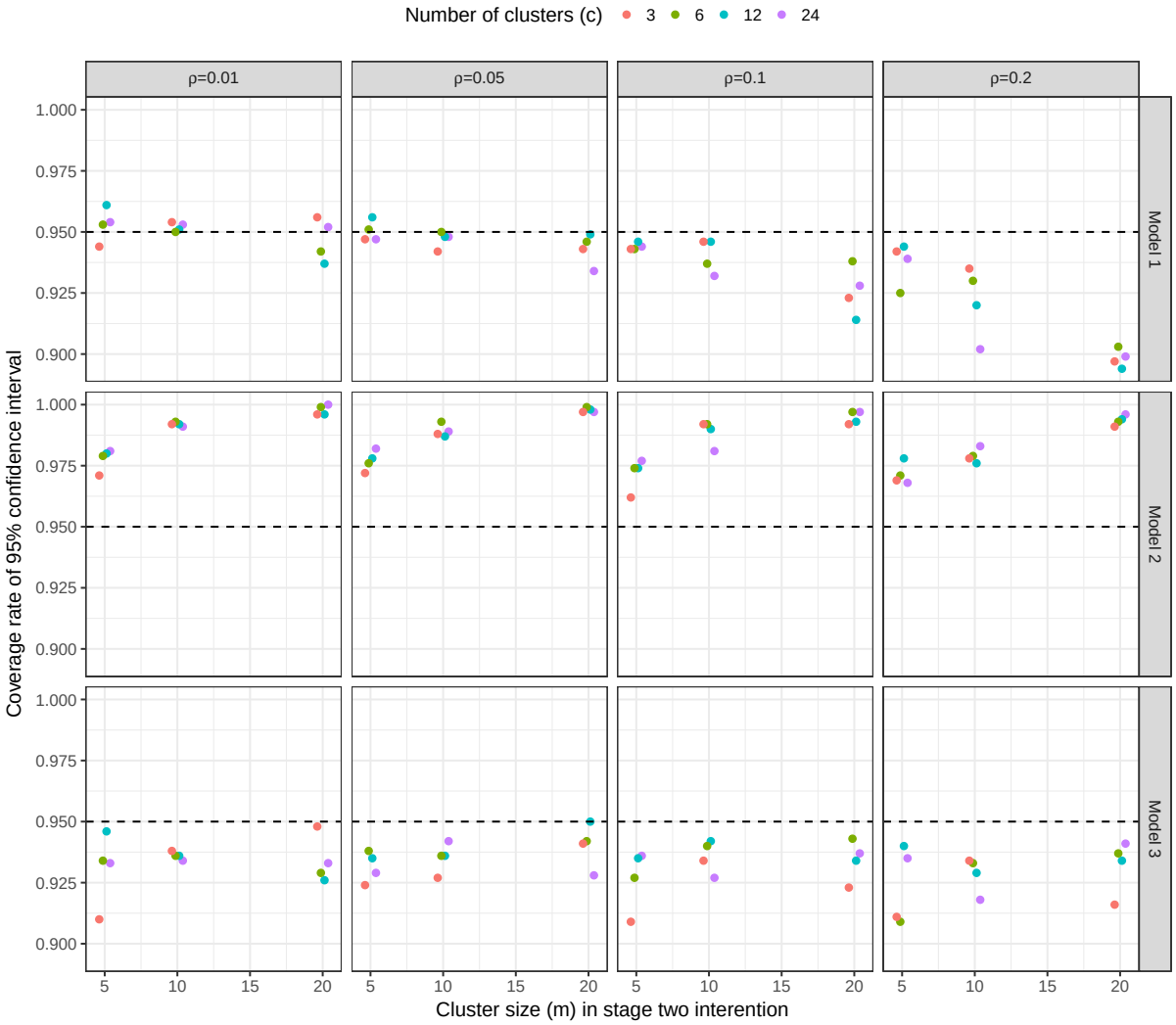


Figure 5.6: Coverage rate of 95% confidence intervals of linear regression model for analysis of within-arm pnRCTs: using OLS standard errors (model L1), cluster robust standard errors (model L2), bootstrap standard errors (model L3), by ρ, c, and m



5.7 Discussion

This chapter expands upon the simulation study in chapter 4 to the more complex within-arm pnRCTs. Different modelling strategies have been investigated for the analysis of a within-arm pnRCT with a continuous outcome and a two staged intervention with clustering at the second stage. Initial investigations into formulating an appropriate mixed effects model for within-arm pnRCTs showed that it was not possible to parametrise a mixed effects model that both accounts for the clustering of stage two and provides an unbiased intervention effect. The bias of the intervention effect estimate from a mixed effect model was demonstrated with a simulation study under ICC $\rho = 0.05$. The model was biased under both the null and alternative.

The linear regression models were found to provide unbiased effect estimates regardless of the

method used for standard error estimation. The findings presented in this chapter demonstrate that ignoring within-arm partial nesting using a linear regression model can result in under coverage rate of 95% confidence intervals. When ICC $\rho \geq 0.05$ the coverage rate of the linear regression model was almost always under the nominal 95% level regardless of the cluster size m or number of clusters c. This in turn may result in misleadingly over precise estimation of intervention effect estimates. However, when the ICC was small $\rho = 0.01$ the Type I error was only minimally inflated.

Neither of the methods used with the aim to account for clustering of stage two intervention, cluster robust or cluster bootstrap standard errors, provided coverage at the nominal level. In addition, the mixed effect model presented in equation 5.1 resulted in biased effect estimates. Caution in the analysis of this trial design can result in more spurious effect estimates and/or coverage rates.

One of the objectives of this chapter was to provide recommendations for the most appropriate analysis methods for within-arm pnRCTs considering the ICC estimates. The scenarios explored in this simulation study suggest that it is not possible to account for clustering in a within-arm pnRCT at the a latter stage of intervention whilst also obtaining an unbiased effect estimate using the methods evaluated in this chapter. In a proportionate intervention if either the ICC and/or the proportion of participants expected to receive the conditional clustered intervention stage(s) are expected to be large then the findings from this chapter suggest caution should be taken in considering what a trial of such an intervention will be able to provide us. Such a within-arm pnRCT will likely not be able to provide estimates of the intervention effect with usual expected precision. It is recommended to ignore the clustering when it is realistic to assume ICC is small. Where feasible, methods to reduce the affect of clustering such as small clusters and standardising the intervention are also suggested. It may be possible to group or cluster those who do not receive the clustered intervention into pseudo clusters. Pseudo clusters impose a clustered structure for participants who do not have one. Although the affects of this would need to explored, as those in the pseudo clusters do not have clustered outcomes.

An ideal solution for analysing within-arm pnRCTs was not identified in this chapter. However, as will be shown in chapter 7, treatment induced clustering is often small ($\rho \leq 0.01$) and thus its impact on results may be expected to be relatively small particularly if cluster sizes are small.

If non-random clustering occurs it is not recommended to account for it using the methods evaluated in this chapter (cluster robust or cluster bootstrap standard errors). If a study aim is

to evaluate the effect of different intervention stages with a treatment pathway and also account for the clustering of these stages randomisation after each stage of the intervention would be required. For instance, a SMART trial design re randomises individuals sometimes multiple times as seen in chapter 3. These designs are typically used to find an optimal intervention as opposed to a confirmatory trial of effectiveness. Multiple randomisations could result in requiring prohibitively large sample sizes for the overall trial design and a greater administrative burden. Landau and Chalder [52] argue that if the allocation of clusters is driven by patient characteristics (for example, the "best" therapist sees "worst" patients) then therapist effects cannot be separated from patient effects. Within-arm pnRCTs do not necessarily link allocation of care providers to patient characteristics. However, typically the non-responders or high risk patients proceed to more intense latter stages of the interventions, thus it is not possible to untangle the effect of clustering from the fact that these patients are higher risk (non-responders). Being unable to untangle the confounding of cluster effects and patient characteristics is not helpful if we wish to estimate clustering effects for the use in future sample size calculations. However, ICC estimates can still be obtained from such trials and are recommended to be able to better understand the validity of the precision of effect estimates. For individuals who receive the clustered interventions given proportionately, it may be possible to report the pre-treatment ICC (at the decision stage) and compare this with post-treatment ICC to see if ICC has increased by treatment.

Ideally, to ensure internal validity, it is recommended that the allocation of participants to clusters should be done at random. However, the nature of a using a RCT to evaluate the overall intervention effect of a proportionate intervention with randomisation at baseline will mean this is not possible.

5.7.1 Limitations

Fixed cluster sizes have been used in this simulation study and it has been assumed there is no effect of clustering in the control arm (as in chapter 4). Both of these things may not strictly be true in practice, the implications have been discussed in more detail in chapter 4.

This simulation study only considered a two stage intervention design. As seen in the systematic review in chapter 3 there are often more than two stages in a proportionate intervention stage. Although, as we move further along the treatment pathway it seems plausible that there will be fewer patients that receive the latter stepped-up intervention stages. The proportion of

participants clustered at latter stages will likely reduce to small percentages of those in the intervention arm. The fewer patients that have clustered outcomes the smaller affect this has on the precision of the standard errors.

5.8 Summary

This chapter has built on findings from the systematic review in chapter 3 and the simulation study in chapter 4. Different analysis methods were evaluated for within-arm pnRCTs with a continuous outcome and a two staged intervention with clustering at the second stage. Neither a mixed effects model nor a linear regression model using either cluster robust or cluster bootstrap standard errors provided both an unbiased effect estimate and coverage rates around the nominal 95% rate. Consequently, no ideal solution for analysing within-arm pnRCTs was identified in this chapter. In a proportionate intervention if either the ICC and/or the proportion of participants expected to receive conditional clustered intervention stage(s) are expected to be large then it is important to bear in mind what a trial of such an intervention will be able to provide us. Such a within-arm pnRCT will not be able to provide estimates of the intervention effect with usual expected precision.

In the next chapter that follows sample size formulae for pnRCTs are summarised. No optimal analysis strategy which accounts for clustering was found for within-arm pnRCTs, hence, sample size formulae in chapter 6 are not expanded to the case of within-arm pnRCTs at present.

Chapter 6

Sample size methods for partially nested trials

6.1 Introduction

Previous chapters highlighted the importance of the analysis method for iRCTs with clustering. Chapter 2 also introduced the concept of the use of the design effect to inflate sample sizes for trials with clustering. This chapter identifies and collates sample size formulae currently available for pnRCTs with continuous outcomes, presents a worked example of these methods, and identifies some possible areas for further work. There is a growing body of literature on sample size methods relevant for pnRCTs. There are accessible methods and corresponding software to allow for the design complexities of clustering in trials with continuous outcomes and these can be extended for further complexities such as variable cluster sizes and incorporation of baseline measures in the analysis.

A priori sample size calculations are used when designing trials. The calculation is used to estimate the minimum number of participants required to be able to detect an intervention effect to a specified probability. The size of an intervention effect, with the precision conveyed by a confidence interval, also need to be estimated. Sample size calculations are an important and often challenging part of a trial. Lack of independence in pnRCTs introduces complexities to the design phase of a trial, similar to that which occur in cRCTs. Ignoring the clustering in the sample size calculation could result in underpowered studies.

6.2 Chapter aims

This chapter aims to identify and collate a comprehensive summary and resource for sample size methods available for pnRCTs. The specific objectives are to:

1. provide a summary of existing methods;
2. provide practical guidance around the use of different methods;
3. link these methods to relevant available software.

This chapter also highlights the sensitivity of sample size calculations to both the ICC and cluster size. The focus is on parallel group trials with continuous outcomes.

6.3 Methods

6.3.1 Literature search

A comprehensive literature search was used to identify published methods for sample size calculations for pnRCTs. The database Ovid MEDLINE was searched on 30^{th} June 2017 for relevant articles. The search criteria presented in Figure 6.1 was implemented. Search terms were chosen based on literature already reviewed for this thesis, carefully looking for search terms in the documents, particularly at the title, abstract and the key words used.

To keep up to date, an Auto Alert was created within Ovid MEDLINE to provide a monthly notification of new citations that match the search specifications after the 30^{th} June 2017 and reviewed until 30^{th} June 2018.

Figure 6.1: Ovid MEDLINE search criteria for relevant articles on pnRCTs sample size calculations

```
Controlled Clinical Trials as Topic/mt, sn OR
Randomized Controlled Trials as Topic/mt, sn AND
Cluster Analysis OR
"partial$ nest$".mp.  OR
"partial$ cluster$".mp.  OR
"Individual$ Randomized Group Treatment
Trial$".mp OR.
"Multilevel data".mp.
```

The results were hand searched based on titles, abstracts and where necessary the full article, to identify relevant results. In addition to the database search, papers known by myself or

supervisors to be relevant were included. The results were reviewed and summarised, identifying the most relevant articles describing methodology for sample size calculations in pnRCTs.

6.3.2 Literature search results

The Ovid MEDLINE search identified 296 unique articles. After the searching 16 full text articles were selected (15 from MEDLINE and one already known to author). These texts include methodology, application, and software specific articles with methods for pnRCTs.

6.4 Results: sample size formulae

The sample size formulae for pnRCTs are presented. For clarity, this chapter builds up from the standard parallel two arm trial, to the fully clustered trial design which introduces the need to account for clustering, and finally moving onto the fully and partially nested design with results from the literature search. Sample size formulae have been re-expressed in consistent terminology and all assume a continuous outcome.

All sample size in this chapter assume a two-sided significance level α and power of $1-\beta$. The following considers a superiority RCT used to test the hypotheses,

$$H_0 : \delta = 0 \text{ versus } H_A : \delta = \delta^*,$$

where δ^* is the clinically meaningful difference.

Sample size formulae commonly assume large samples ($n \geq 30$); it has been suggested that the asymptotic Normality assumption of the test statistic is acceptable if there are 30 or more individuals in each trial arm [156]. When sample sizes are small the additional uncertainty can be accounted for by using a t-distribution. Although the total number of individuals in an RCT with clustering may be large the number of clusters is often small.

6.4.1 Trial design features that impact on sample size

In conducting a priori sample size calculation certain trial design features need to be estimated, key design features are presented in Table 6.1. These are typically decided upon using a combination of previous trial data, observational data and health records, financial and practical

constraints, and expert opinion. Some features can be manipulated by the researcher, however, others will be constrained by the limits of the data, the ICC, or financial resources. Chapter 7 adds to the evidence of ICC for the choice of ICC ρ in iRCTs with clustering.

Table 6.1: Trial design features required for a sample size calculation in a pnRCTs

Design feature	**Description**	**Notation**
Endpoint	Endpoint used for the primary outcome	y
Error rates	Type I error	α
	Type II error	β
Effect size	Minimum clinically important difference measured on primary endpoint	δ
Variance	Population variability for primary outcome.	
	Total variance	σ^2
	Individual variance control arm	σ_r^2
	Individual variance intervention arm	σ_ε^2
	Cluster variance	σ_u^2
Cluster size	Number of participants in each cluster, may be fixed or researcher may be able to choose the size.	m
Variability in cluster size	The variability in cluster sizes: Coefficient of variation	cv
Variability between clusters	How much of the outcome variability is due to clustering: ICC $\rho = \sigma_u^2/\sigma_u^2 + \sigma_\epsilon^2$	ρ
Other	Expected drop-outs/attrition rates	

6.4.2 Sample size formulae for individually randomised trial

Consider an iRCT using a parallel group design with Normally distributed outcomes y, with mean $\bar{y}$ and standard deviation σ and assume equal variance between arms. The intervention effect size is the expected mean of the outcome value in the intervention arm minus the expected mean value in the control arm

$$\delta = \bar{y}_1 - \bar{y}_0 \tag{6.1}$$

The standard parametric test for the hypotheses for a superiority RCT (expressed above) is an independent two-sample t-test. The test is based on the test statistic, T, which under the null hypothesis has a central t-distribution on $v = n_I(r+1) - 2$ degrees of freedom where n_I is the

sample size in one trial arm and r is the allocation ratio between the two arms. The T test statistic is given by

$$\begin{aligned} T &= \sqrt{r(1-r)n_I}\left(\frac{\bar{y}_1 - \bar{y}_0}{\sigma}\right) \\ &= \sqrt{r(1-r)n_I}\delta_s. \end{aligned} \tag{6.2}$$

where δ is the expected effect size, σ is the estimated population standard deviation and δ_s is the standardised effect size, $\delta_s = {}^{\delta}/_{\sigma}$.

6.4.2.1 Intervention effect estimator variance

We wish to minimise the variance of the intervention effect estimator, hence, it is used as the optimality criterion; minimal variance results in maximal power to detect an intervention effect. In general terms for a two-sided significance level α and power of $1-\beta$ we require,

$$\begin{aligned} \mathrm{Var}(\hat{\delta}) &= \left(\frac{\delta}{Z_{1-\beta} + Z_{1-{}^{\alpha}/_{2}}}\right)^2 \\ &\text{and} \\ \mathrm{Var}(\hat{\delta}) &= \frac{\sigma^2}{n_I} + \frac{\sigma^2}{n_B} = \frac{(r+1)}{r}\frac{\sigma^2}{n_I} \end{aligned} \tag{6.3}$$

where $n_B = rn_I$ and $Z_{1-\beta}$ and $Z_{1-{}^{\alpha}/_{2}}$ are the $100(1-\beta)\%$ and $100(1-{}^{\alpha}/_{2})\%$ points of a standard Normal distribution.

6.4.2.2 Approximate sample size formulae using asymptotic methods

For moderately large sample sizes (≥ 30) the t-distribution approximates to a standard Normal distribution. The approximate sample size per arm using the Normal approximation, from equations in 6.3, is

$$n_I = \frac{(r+1)(Z_{1-\beta} + Z_{1-{}^{\alpha}/_{2}})^2\sigma^2}{r\delta^2} \tag{6.4}$$

where r is the allocation ratio, δ is the expected effect size, σ^2 is the estimated population variance [33]. Assuming an allocation ratio of 1:1 this gives

$$n_I = \frac{2\sigma^2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{\delta^2}. \tag{6.5}$$

6.4.2.3 Exact sample size formulae

When a clinical trial is undertaken the population variance σ^2 is usually considered to be unknown and a sample variance is used in its place, s^2, estimated with $v = n_I(r+1) - 2$ degrees of freedom. Uncertainty in the variance estimate can be reflected by replacing the Z-statistic with a t-statistic. The following sample size formulae can be used to achieve a power of at least $1 - \beta$

$$n_I \geq \frac{(r+1)(Z_{1-\beta} + t_{1-\alpha/2,n_I(r+1)-2})^2\sigma^2}{r\delta^2}.$$

This equation does not have a direct solution as n_I appears on both sides of the equation. The equation can be written in terms of power and solved using an iterative technique

$$1 - \beta = \Phi\left(\sqrt{\frac{rn_I\delta^2}{(r+1)\sigma^2}} - t_{1\,-\,\alpha/2,n_I(r+1)-2}\right) \tag{6.6}$$

where $\Phi(.)$ is the cumulative density function of $N(0,1)$. When a sample variance is being used, s^2, the power should be estimated from the cumulative t-distribution instead of the Normal distribution. In addition, power is estimated under the alternative hypothesis and Senn [157] showed more accurate sample size calculations can be done using the non-central t-distribution. The power is actually being estimated under $\delta \neq 0$, hence, the corresponding t-distribution should be non-central which represents the distribution of the test statistic under the alternative hypothesis of unequal means. The non-centrality parameter is given by

$$\lambda = \sqrt{\frac{rn_I\delta^2}{(r+1)\sigma^2}}.$$

The power equation can be rewritten to give

$$1 - \beta = 1 - T^{-1}\left(t_{1-\alpha/2,n_I(r+1)-2}, n_I(r+1) - 2, \sqrt{\frac{rn_I\delta^2}{(r+1)\sigma^2}}\right) \tag{6.7}$$

where $T^{-}1$ is the cumulative density function of the non-central t-distribution. Again this equation cannot be solved for n_I explicitly. Practically, the Normal approximation can be used

as an initial sample size and then use an iterative solution until the required power is reached.

Alternatively, to allow for the Normal approximation to the t-distribution a correction factor of $\frac{Z_{1-\alpha/2}}{4}$ can be added to equation 6.4 to approximate and give

$$n_I = \frac{(r+1)(Z_{1-\beta} + Z_{1-\alpha/2})^2\sigma^2}{r\delta^2} + \frac{Z_{1-\alpha/2}^2}{4}. \tag{6.8}$$

When sample sizes are small the Normal approximation to the t-distribution is poor, this can result in overestimation of power. The differences between the Normal distribution and the non-central t-distribution are generally minimal [33]. In general, the central t-distribution is used, Machin et al. [158] propose that the use of t rather than Z adds 1 and 2 per group for the 5% and 1% significance level, respectively.

6.4.3 Sample size formulae for cluster randomised trial

A simple approach to take account of the clustering effect when designing a cluster trial was proposed by Donner et al. [159]: calculate a sample size for an iRCTs (as in section 6.4.2) and inflate this by the design effect (DE) to reach the required statistical power under cluster randomisation

$$DE = 1 + (m-1)\rho \tag{6.9}$$

where m is the number of individuals per cluster and ρ is the ICC. For example, when the ICC is 0.05 and the cluster size is eight, the design effect equals 1.35 meaning that 35% more participants would need to be recruited to achieve a sufficient sample size. This design effect (in equation 6.9) is actually for cluster level analyses which compares relevant summary statistics from the cluster level. Different design effects are summarised by Eldridge et al. [150] for the different analyses methods; when individual level analyses are planned using equation 6.9 is conservative.

6.4.3.1 Approximate sample size formulae using asymptotic methods

Assuming an allocation ratio of 1:1 the required number of individuals per arm for a cRCT is,

$$\begin{aligned} n_C =& \frac{2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{\delta_s^2}(1 + (m-1)\rho) \\ n_C =& n_I \times DE \end{aligned} \tag{6.10}$$

where δ_s is calculated as $\delta_s = \delta/\sqrt{\sigma_u^2 + \sigma_\epsilon^2}$, σ_u^2 is the between cluster variance and σ_ϵ^2 is the within cluster variance. The number of clusters required per arm, k, is then given by

$$k = \left(\frac{n_C(1 + (m-1)\rho)}{m} \right). \tag{6.11}$$

6.4.3.2 Exact sample size formulae and small number of clusters

Sample size formulae in equations 6.11 assume a relatively large number of clusters, thus the Normal approximation will be appropriate. When the number of clusters is small, using the Normal approximation is likely to overestimate power and underestimate sample size [148]. As in the standard parallel two-arm iRCT, it is possible to replace the Normal distribution with a non-central t-distribution. Donner and Klar [148] presented a power calculation for cluster-level analyses which uses the non-central t-distribution with $v = 2(k-1)$ degrees of freedom and a non-centrality parameter of

$$\lambda = \frac{\delta_s}{\sqrt{2\left(\frac{1+((m-1)\rho)}{mk}\right)}}. \tag{6.12}$$

The above gives us a formulae for power of

$$1 - \beta = 1 - T^{-1}\left(t_{1-\alpha/2,2(k-1)}, 2(k-1), \frac{\delta_s}{\sqrt{2\left(\frac{1+((m-1)\rho)}{mk}\right)}} \right) \tag{6.13}$$

The degrees of freedom for the t-distribution are determined by the number of clusters, which is what we are trying to estimate using sample size calculations. Hence, the above requires an iterative process of estimation. Harrison, Brady, et al. [160] also show that more accurate power calculations can be done using the non-central t-distribution, presenting formulae for sample size and power under the non-central t-distribution alongside a Stata command to perform these calculations `sampncti`.

6.4.3.3 Unequal cluster sizes

The standard design effect given in equation 6.9 assumes that the cluster sizes are equal. However, it is common to have variable cluster sizes, for example variable GP practice sizes or variable school sizes. Exact cluster sizes may not be known at the design stage due to lack of

information and variability over time. However, a priori estimates of the distribution of cluster sizes may be available (mean and standard deviation). Where the cluster sizes are unequal it is helpful to consider the impact this can have when designing a trial.

The design effect in equation 6.9 can be approximated by including a simple correction for the variable cluster size as shown by Eldridge et al. [150]

$$DE = 1 + ((cv^2 + 1)\bar{m} - 1)\rho \tag{6.14}$$

where cv is the coefficient of variation in cluster size. Coefficient of variation is the ratio of the standard deviation of cluster sizes, σ_m, to the mean cluster size, $\bar{m}$ ($\sigma_m = \sum \big(m_i - \bar{m})^2/(k-1)\big)$ where m_i is the size of cluster i and k is the number of clusters) calculated using

$$cv = \frac{\sigma_m}{\bar{m}}. \tag{6.15}$$

Eldridge et al. [150] demonstrate that using the coefficient of variation formula in cluster size provides either good or conservative estimate of sample size requirements.

6.4.4 Sample size formulae for partially nested randomised trials

Sample size formulae for iRCTs with treatment induced clustering have similarities to those for cRCTs. Where there is differential clustering between arms, as in a pnRCT, the variance may differ and thus the analysis needs to account for this variation. The analysis of a pnRCT may in its simplest form correspond to that of an unequal variances t-test [136], however, it is typically the case that analysis adjusted for baseline covariates is required and can provide more precise effect estimates (analysis methods were evaluated in chapter 4 of this thesis).

Assuming the individual variance in the unclustered control arm is given by σ_r^2, the individual variance in the cluster arm by σ_ε^2 and the between cluster variation in the clustered arm by σ_u^2. The ICC in a pnRCTs, ρ, is only present in the clustered intervention arm, and is again given by $\rho = {\sigma_u^2}/{\sigma_u^2 + \sigma_\varepsilon^2}$.

6.4.4.1 Partially nested randomised trials sample size literature

A simple solution to incorporate clustering effects in pnRCTs sample size calculations is to: calculate sample size for the corresponding iRCT; calculate the design effect as in a cluster trial; and inflate only the clustered arm by this design effect. Similarly, for nRCTs both trial arms may be inflated, either using the same design effect for both arms or if different ICCs and/or cluster sizes are anticipated for different trial arms then two separate design effects can be calculated one for each trial arm as suggested by Pals et al. [43]. Although this method does recognise the clustering effect and has been used, it does not fully account for the potential differential variance between arms in pnRCTs and may not provide an optimal allocation ratio. A review of recent methodological developments in group randomised trials design by Turner et al. [161] gave a brief summary of relevant literature dedicated to what they term Individually Randomised Group-Treatment trials, with references to various sample size articles. This chapter aims to extend this, with further information, a focus on pnRCTs, and a summary of software for sample size calculations in pnRCTs.

There are various articles reporting specific sample size formulae for trials with clustering of outcomes in only one trial arm. Roberts and Roberts [48] present an optimal allocation ratio for pnRCTs. Moerbeek and Wong [131] derive sample size formula for pnRCTs, which account for both the clustering in one arm and the possible heteroscedastic variance between arms. Similar formulae assuming homoscedastic variances have also been derived [57, 122, 132, 137]. Lohr et al. [133] derive a design effect for pnRCTs, however, this makes the assumption that the individual variance parameters are equal across arms $\sigma_r^2 = \sigma_\epsilon^2$, which may not always be the case. During the design phase of a trial specific information on the individual variances may not be available, however, with improved reporting of such trials this will have hopefully lead to more information on variance estimates and a better understanding of potential levels of heteroscedasticity of variances across arms. Hoover [136] suggested a sample size and power formula for small studies for the Satterthwaite unequal variance t-test. Roberts and Roberts [48] state that where variance is larger in the larger group (feasibly that this is true when comparing group therapy to individual therapy) the formula underestimates power.

Multi-centre trials introduce an additional level of data hierarchy in pnRCTs. Heo et al. [137] consider centre level clustering, they derive sample size formulae for pnRCTs under homoscedastic variances. They derive sample size formulae for testing main effects of a group-based intervention compared to individual-based control based on two- and three-level mixed-effects linear

models. The two-level mixed effects model represents when there is only clustering due to the group base intervention. The sample size formulae in Heo et al. [137] differs from Moerbeek and Wong [131] due to the assumption of homoscedastic variances and the assumption of equal allocation across arms. Appendix C provides a table comparing the two methods under different designs. The three-level mixed effects model represents when there is clustering due to the group based intervention and clustering due to the centre. In practice, the centres in multi-centre trials are commonly included as a stratification variable in the randomisation procedure and included as a fixed effect in the mixed effect model as opposed to a random effect. However, the use of random effects can have advantages compared to fixed effects in many scenarios [54].

Other complex designs, including clustered designs with clustering by therapist in both trial arms, or the crossed design where each therapist provides therapy in both trial arms, have been discussed with a focus on psychotherapy but are more widely applicable [38, 39].

6.4.4.2 Intervention effect estimator variance

The variance of the intervention effect estimator, $\hat{\delta} = \bar{y}_1 - \bar{y}_0$, in a pnRCTs was derived by Moerbeek and Wong [131] as

$$\text{Var}(\hat{\delta}) = \frac{\sigma_\epsilon^2 + m\sigma_u^2}{mk} + \frac{\sigma_r^2}{n_{pn}}$$

and

$$\text{Var}(\hat{\delta}) = \sigma_r^2 \left(\gamma \frac{(m-1)\rho + 1}{mk} + \frac{1}{n_{pn}} \right) \tag{6.16}$$

where $\gamma = {\sigma_u^2 + \sigma_\varepsilon^2}/{\sigma_r^2}$ is the ratio of variance in the clustered arm to the unclustered arm. The cluster size and number of clusters are denoted by m and k and n_{pn} is the number of individuals in the control arm.

Equation 6.16 can be simplified under the assumption of equal individual variances across the two treatment arms $\sigma_r^2 = \sigma_\epsilon^2$ to give

$$\text{Var}(\hat{\delta}) = \frac{\sigma_r^2 + m\sigma_u^2}{mk} + \frac{\sigma_r^2}{n_{pn}}$$

and

$$\text{Var}(\hat{\delta}) = \sigma_r^2 \left(\frac{(m-1)\rho + 1}{mk} + \frac{1}{n_{pn}} \right). \tag{6.17}$$

The variance of the intervention effect estimator, δ, expressed in equations 6.16 and 6.17 is dependent upon the cluster size m, number of clusters k, and the number of individuals in the control arm n_{pn}. The combination of these parameters is referred to as a design, $\xi = \{m, k, n_{pn}\}$. Different designs can result in the same power level, with an optimal design providing the best combination of $\xi = \{m, k, n_{pn}\}$ under the structure, budget and practicalities of a particular trial.

6.4.4.3 Approximate sample size formula using asymptotic methods

Equal allocation is commonly used for RCTs, this typically maximises power for a given total sample size. However, if costs and variances differ between trial arms, often the case in pnRCTs, the power may be maximised by changing the allocation ratio between the trial arms. The following presents both optimal allocation ratios and sample size formulae to achieve desired power and Type I error rate.

Sample size for fixed cluster size and optimal allocation ratio

If the cluster size is known and we wish to choose the optimal sample size calculations for a pnRCT a number of steps are required. An optimal allocation ratio for large partially clustered trials is given by Roberts and Roberts [48] as the ratio of the individuals in the unclustered control arm to the clustered intervention arm

$$\frac{mk}{n_{pn}} = \sqrt{1 + (m-1)\rho}. \tag{6.18}$$

This can be extended to account for heteroscedastic variance by adding the variance ratio γ, which gives

$$\frac{mk}{n_{pn}} = \sqrt{\gamma(1 + (m-1)\rho)}. \tag{6.19}$$

An increase in the design effect and the variance ratio both result in an increase in the sample size in the clustered arm. The number of individuals required in the control arm is given by Moerbeek and Teerenstra [130] as

$$n_{pn} = \left(\sqrt{\gamma(1 + (m-1)\rho)} + 1\right)\left(\frac{Z_{1-\beta} + Z_{1-\alpha/2}}{\delta_s}\right)^2 \tag{6.20}$$

where the standardised effect size expresses the intervention effect in relation to the standard

deviation in the control arm, $\delta_s = {}^{\delta}/_{\sigma_r}$. In practice, the number of control participants required is calculated using equation 6.20 in conjunction with the allocation ratio formula in 6.19 to calculate the number of intervention groups, k (and the number of individuals in the clustered intervention arm, mk). The steps to undertake a pnRCT sample size for a known cluster size calculation are illustrated below.

Calculate sample size for a partiality nested trial assuming two-sided test with $\alpha = 0.05$ and power $1 - \beta = 0.80$:

1. Information assumed provided by investigator: Intervention effect size of interest δ_s=0.5 and expected number of patients per cluster m=10

2. Obtain an estimate of ρ (from previous studies, observational data, and/or discussions with investigator): ICC ρ=0.05

3. Preliminary sample size: Estimated sample size of control arm $n_{pn} = 69.20$ and clustered intervention arm $mk = 83.33$, total sample size of 152.53, using equations 6.19 and 6.20 and assuming variance ratio $\gamma = 1$.

4. Calculate sample size required: Round clustered intervention arm to multiple of cluster size m.

    - Rounding down gives $mk = 80$ and recalculate control arm sample size $n_{pn} = 73$ to achieve desired power, total sample size of 153.
    - Rounding up gives $mk = 90$, nine clusters of size ten, and recalculate control arm sample size $n_{pn} = 64$ to achieve desired power, total sample size of 154.

5. Allowing for 10% drop-out in participants from baseline to follow-up, this gives $mk = 108$ nine clusters of size 12 and control arm sample size $n_{pn} = 72$ to achieve desired power, total sample size of 180.

Table 6.2 illustrates the sample sizes required in a pnRCT to achieve 1-β=0.80 in a two-sided test with α=0.05 to detect a standardised intervention effect $\delta_s = 0.5$, $\gamma = 1$ (assuming homoscedasticity between individual variance across trial arms) for various values of cluster size m and ICC ρ. For instance, for $\rho = 0.05$, $m = 10$ eight clusters are required in the intervention arm $mk = 80$ and $n_{pn} = 73$ in the control arm. For a parallel group iRCT, with no clustering, the required sample size would be 64 per arm. Calculations in Table 6.2 use equations 6.19 and

6.20. When calculating sample sizes which include clusters, it is generally necessary to round the clustered intervention arms to a multiple of the cluster size. The required sample size in the clustered intervention arm can be either rounded down or up to the nearest multiple and the control arm sample size calculated accordingly using equation 6.20. Similar total sample sizes are required in the examples given in Table 6.2, however, this is not always the case if cluster size or ICC are large rounding can have a larger influence on the required total sample size to achieve desired power.

It is evident from sample sizes presented in Table 6.2 that for larger cluster sizes m larger overall sample sizes are required to achieve the same power. This agrees with findings from chapter 4. For instance, with an ICC of $\rho = 0.05$ we require a total sample size of 138 with a cluster size of $m = 5$ and a total sample size of 181 with a cluster size of $m = 20$.

Table 6.2: Sample size for pnRCT to achieve 1-β=0.80 in a two-sided test with α=0.05 to detect a standardised intervention effect δ_s=0.5, γ=1 (assuming homoscedasticity between individual variance across trial arms) for various values of cluster size m and ICC ρ.

ρ	Rounding method*	Cluster size (m)											
		5				10				20			
		k	mk	n_{pn}	Total	k	mk	n_{pn}	Total	k	mk	n_{pn}	Total
0.01	Up	13	65	64	129	7	70	62	132	4	80	59	139
	Down	12	60	69	129	6	60	74	134	3	60	84	144
0.05	Up	15	75	64	139	9	90	64	154	6	120	65	185
	Down	14	70	68	138	8	80	73	153	5	100	81	181
0.1	Up	17	85	66	151	11	110	69	179	8	160	73	233
	Down	16	80	70	150	10	100	78	178	7	140	90	230

*Rounding method of sample size calculation refers to the method of rounding from the original calculation: either rounding number of clusters up or rounding number of clusters down.

Sample size for fixed cluster size and fixed number of clusters

The number of individuals in the control arm for fixed cluster size m and number of clusters in the intervention arm k [131] is

$$n_{pn} = \frac{\sigma_r^2}{\left(\frac{\delta^2}{\left(Z_{1-\beta}+Z_{1-\alpha/2}\right)^2} - \frac{\sigma_\epsilon^2+m\sigma_u^2}{mk}\right)} = \frac{1}{\left(\frac{\delta_s^2}{\left(Z_{1-\beta}+Z_{1-\alpha/2}\right)^2} - \frac{\gamma(1+(m-1)\rho)}{mk}\right)}. \tag{6.21}$$

Sample size including costs

The optimal allocation ratio and sample size formulae can also be extended to account for differential costs. It may be the case that costs do not depend only on the total sample size but

are based on other parameters. If a trials budget includes the intervention running costs then there are likely to be differing costs between trial arms. The cost ratio c_1/c_0 where c_1 and c_0 are the costs per individual in the intervention and control arms can be added to sample size formulae [130] giving

$$\frac{mk}{n_{pn}} = \sqrt{\gamma(1+(m-1)\rho)\frac{c_1}{c_0}}. \tag{6.22}$$

The number of individuals required in the control arm accounting for costs is

$$n_{pn} = \left(\sqrt{\gamma(1+(m-1)\rho)\frac{c_1}{c_0}}+1\right)\left(\frac{Z_{1-\beta}+Z_{1-\alpha/2}}{\delta}\right)^2. \tag{6.23}$$

The notion of including intervention costs to calculate the optimal allocation ratio to reduce overall costs for a pnRCT design was put forward by Moerbeek and Teerenstra [130] and Moerbeek and Wong [131]. No trials of where differential intervention costs had been used in the sample size could be found to provide an example in this chapter. It is likely that this is more suitable for certain healthcare settings dependent upon the funding models.

6.4.4.4 Exact sample size formulae: small sample sizes and unequal variances

The sample sizes calculations in equations 6.19-6.21 are based on the standard Normal approximation of the test statistic, and thus assume asymptotic Normality which is only suitable for large number of clusters. For small studies using these formulae will underestimate sample size and overestimate power. In addition, Roberts and Roberts [48] noted that in small studies the optimal allocation ratio they derived (equation 6.19) will give a smaller allocation ratio than is optimal.

Again for small studies the t-distribution can be used as an approximation to the Normal distribution and the test statistic under the alternative hypothesis can be approximated with a non-central t-distribution.

The variance of the effect estimator can be used to give a non-centrality parameter for a pnRCT [130]

$$\lambda = \frac{\delta}{s.e.(\delta)} = \frac{\delta}{\sqrt{\frac{\sigma_\epsilon^2+m\sigma_u^2}{mk}+\frac{\sigma_r^2}{n_{pn}}}} = \frac{\delta_s}{\sqrt{\gamma\frac{(m-1)\rho+1}{mk}+\frac{1}{n_{pn}}}} \tag{6.24}$$

Using the non-central t-distribution the power is derived as

$$1-\beta = 1 - T^{-1}\left(t_{1-\alpha/2,v}, v, \frac{\delta_s}{\sqrt{\gamma\frac{(m-1)\rho+1}{mk}+\frac{1}{n_{pn}}}}\right) \tag{6.25}$$

with degrees of freedom v. The degrees of freedom for the t-distribution are determined by the number of clusters, which is what we are trying to estimate using sample size calculations. Equation 6.25 requires an iterative process of estimation, the approximate sample size using asymptotic methods can be used as an intial sample size (from section 6.4.4.3) and then use an iterative solution until the required power is reached. As shown in chapter 4, the Satterthwaite degrees of freedom correction can be used to estimate v, which avoids the assumption of equal variances and makes a small sample correction for small number of clusters [139, 162]. The Satterthwaite approximation for effective degrees of freedom is made up of a weighted linear combination of the variances in this instance given by

$$v = \frac{\left(\frac{\sigma_\epsilon^2+m\sigma_u^2}{mk}+\frac{\sigma_r^2}{n_{pn}}\right)^2}{\frac{1}{k-1}\left(\frac{\sigma_\epsilon^2+m\sigma_u^2}{mk}\right)^2+\frac{1}{n_{pn}-1}\left(\frac{\sigma_r^2}{n_{pn}}\right)^2}. \tag{6.26}$$

For small samples Candel and Van Breukelen [162] used a numerical evaluation to formulate corrections when using the Normal approximation in iRCTs with clustering in both arms; they recommend that in a two-tailed test using REML, a correction for 80% and 90% power, is to add three clusters to each trial arm for a 5% type I error rate and four clusters to each trial arm for a 1% type I error rate.

6.4.4.5 Unequal cluster size

A pnRCT may have unequal cluster sizes. For example, the CASPER plus (CollAborative care for Screen-Positive ElderRs) trial evaluated collaborative care for older adults with major depressive disorder compared to usual care, a total of 20 collaborative care case managers treated a mean of 11.9 patients ranging between 1 to 46 patients [163].

This section extends the optimal allocation ratio and sample size formulae for the number of individuals in the control arm from Moerbeek and Wong [131], to now account for the variable cluster size by including the coefficient of variation, cv, and the mean cluster size (combining equation 6.14 with 6.22 and 6.23)

$$\frac{mk}{n_{pn}} = \sqrt{\gamma(1 + ((cv^2 + 1)\bar{m} - 1)\rho)\frac{c_1}{c_0}} \tag{6.27}$$

and

$$n_{pn} = \left(\sqrt{\gamma\left(1 + ((cv^2 + 1)\bar{m} - 1)\rho\right)\frac{c_1}{c_0}} + 1\right)\left(\frac{Z_{1-\beta} + Z_{1-\alpha/2}}{\delta_s}\right)^2. \tag{6.28}$$

If homoscedastic variances and/or equal costs are assumed, γ and/or c_1/c_0 can be removed from the above equations. The effect of varying cluster sizes in trials with differential clustering was investigated by Candel and Van Breukelen [128] using a Monte Carlo simulation study. Under the simulations scenarios it was found the efficiency loss in the intervention effect estimate was rarely more than 10%, requiring recruitment of 11% per cent more clusters for the intervention arm and 11% more individuals for the control arm. Additionally, Eldridge et al. [150] showed in cRCTs if the coefficient of variation in cluster size is small, less than 0.23, then the correction on sample size is negligible. However, the coefficient of variation may be high in iRCTs with treatment induced clustering. In the SHEAR trial there were 402 patients in the intervention arm treated by 79 clinicians. Clinicians treated between 1 to 35 patients and a coefficient of variation of cluster size of $cv = 1.14$ (this was calculated using the data provided by the trial statistician for work in the following chapter 7 $cv = \sigma_m/\bar{m} = 5.72/5.03$) [164]. Another approach for incorporating cluster size viability in sample size calculations would be to take a simulation approach to sample size. Table 6.3 presents the effect of cv on sample size using equation 6.27 for a number of different scenarios of ρ and cv.

Table 6.3: Effect of coefficient of variation of pnRCT sample size, with $1 - \beta = 0.80$, a two-sided test with $\alpha = 0.05$, $\delta_s = 0.5$, $\gamma = 1$, and $m = 10$.

ρ	cv	**Sample size ignoring cv**				**Sample size including cv**			
		k	mk	n_{pn}	**Total**	k	mk	n_{pn}	**Total**
0.05	0.2	9	90	70	160	9	90	70	160
	0.4	9	90	70	160	9	90	71	161
	0.6	9	90	70	160	9	90	72	162
	0.8	9	90	70	160	10	100	74	174
	1	9	90	70	160	10	100	76	176
	1.2	9	90	70	160	11	110	78	188
0.1	0.2	11	110	75	185	11	110	76	186
	0.4	11	110	75	185	11	110	77	187
	0.6	11	110	75	185	12	120	79	199
	0.8	11	110	75	185	12	120	82	202
	1	11	110	75	185	13	130	85	215
	1.2	11	110	75	185	14	140	89	229

6.4.4.6 Statistical software

To facilitate the work of statisticians planning trials it is important that statistical software is readily available to calculate sample sizes for pnRCTs. Available software for parallel arm pnRCTs sample size calculations are summarised in Table 6.4 referring to commonly used statistical software packages, sample size software and standalone software.

Table 6.4: Software for sample size calculations in pnRCTs

Software	**Functionality**
PASS	Not aware of any built-in functionality at this time
nQuery	Two group Satterthwaite's t-test (unequal variance) computes power and sample size [165], computed using formulas from Moser et al. [166]
R	Not aware of any built-in functionality at this time
SAS	Not aware of any built-in functionality at this time
Stata	User-written command `clsampsi`: can compute sample size and optimal allocation ratio for pnRCTs for continuous and binary outcomes. Uses the non-central F distribution and Satterthwaite degrees of freedom [127].
SPA-ML	Matlab stand-alone program. Sample size calculations for pnRCTs, and fully nested iRCTs with clustering in both arms. Uses the t-distribution where possible and Hoover [136] degrees of freedom [167].

The two statistical software sepcifically designed with pnRCTs in mind are `clsampsi` Stata program [127, 168] and SPA-ML [167]. The `clsampsi` Stata program [127, 168] was developed to calculate the power or the number of clusters and cluster sizes required to evaluate the difference of means or proportions in the presence of differential clustering effects in each trial arm, including pnRCTs. The paper published in The Stata Journal [127] explains that by default, the program clsampsi calculates power by integrating the non-central F-distribution as described by Moser et al. [166]. It uses a numerical search to find sample sizes, initially starting with estimated number of clusters based on the normal approximation as a starting value and then the number of clusters are increased until sufficient power is reached. This program also 'roughly' approximates the optimum allocation ratio between trial arms for a given power.

SPA-ML (Statistic Power Analysis for Multilevel Design) is written in Matlab and available as stand-alone program for Windows [130, 167]. SPA-ML uses the theoretical sample size formula for pnRCTs from Moerbeek and Teerenstra [130] and Moerbeek and Wong [131] included in this chapter. It is based on mathematical relations between sample size and power: no Monte Carlo simulations are used. The output is provided as text and in graph format. Whenever it was

possible, instead of the presented Normal approximation, the software uses the t-distribution to calculate the required sample size/power (this was found out from email correspondence with the SPA-ML developer hence there is no reference). The degrees of freedom for the t-distribution were taken from the Hoover [136] paper. As a result there will be some small discrepancies between the estimations based on the formulas and obtained with the software SPA-ML, particularly for smaller samples.

Comparison of software

Below shows a comparison of the two sample size software designed specifically with pnRCTs in mind, SPA-ML and `clsampsi` for Stata, for the scenario of 80% power, $\rho = 0.05$, $\delta_s = 0.5$, and $m = 10$. The sample size using approximate asymptotic equations 6.20 and 6.19 gives a total sample size of 154 (9 clusters of size 10 and 64 individuals in control arm); using clsampsi this has an estimated power of 78.4% slightly lower than the approximated 80% using Normal approximation.

Stata `clsampsi` gives total sample size = 166, output from Stata below.

Estimated power/sample size using the Satterthwaite approximate F test for two-sample comparison of means with clustering

Test Ho: mu1 = mu2, where mu1 is the mean in population 1

and mu2 is the mean in population 2

Assumptions: alpha = 0.0500 (two-sided)

	Sample 1	Sample 2
Mean (mu)	0	.5
Total St. Dev.(sd)	1	1
Number of Clusters (k)	10	66
Cluster Size (m)	10	1
Cluster Size Var.(varm)	0	0
Sample Size (N)	100	66
Intra-Cluster Corr. (rho)	.05	0
SD (summary level)	.380789	1
Total Sample Size:		166
Allocation ratio (N2/N1):		.66
Ratio of Number of clusters (k2/k1):		6.6
Ratio of Cluster sizes (m2/m1):		.1
Satterthwaite's degrees of freedom:		32.69
Sample size (ni) for integration:		1000
Estimated power:		0.8027

SPA-ML gives total sample size = 168, output from SPA-ML below.

Individual randomized trial, clustering in experimental condition. Scenario: use cost function and fixed size of experimental clusters to calculate number of control subjects and number of experimental clusters. Outcome variable type: continuous. Tail(s) = two; Type I error rate = 0.05; Desired power = 0.8; Cost ratio = 1; Size of experimental clusters = 10; Intracluster correlation coefficient = 0.05; Standardized effect size = 0.5; Variance ratio = 1;

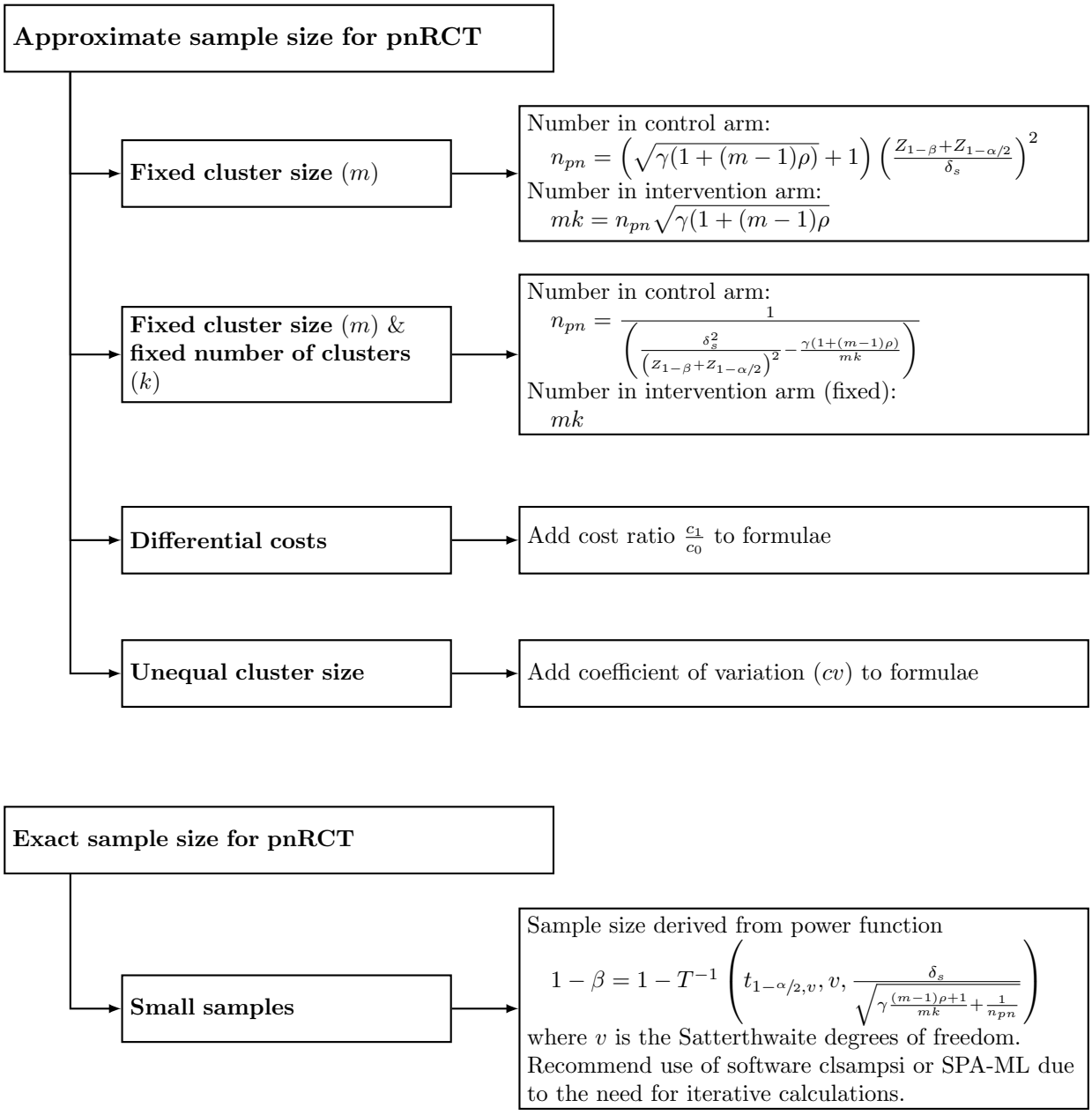
Number of control subjects = 68; Number of experimental clusters = 10; Total sample size = 168; Total costs = 168; Actual power = 0.8013.

There are no alternative designs.

Summary of sample size formulae for pnRCTs

Figure 6.2 presents a flowchart summarising the sample size formulae for pnRCTs.

Figure 6.2: Summary of sample size formulae for pnRCTs



6.4.5 Sample size formulae for within-arm pnRCT

It is typically true that the analysis of a trial should reflect the design, and so a clustered design should be followed by analysis which accounts for clustering. Within-arm pnRCTs involve some of a trial arm having clustered outcomes. An appropriate analysis method which both accounted for the clustering and obtaining an unbiased intervention effect was not identified in chapter 5. When this clustering is dependent upon a post-randomisation outcome/intermediate outcome then it has been shown, in the previous chapter 5, that it is not possible to analyse these types of trials accounting for the clustering and obtaining an unbiased intervention effect. Consequently, the use of such designs is cautioned if a large ICC or large clusters are expected (this will result

in an inflation of the Type I error in the analysis). When within-arm pnRCT designs are used and clustering is non-random the results of chapter 5 suggest ignoring clustering in the analysis such as using a linear regression model.

Sample size calculations are often done from a conservative standpoint. Although no appropriate method was identified in chapter 5 for analysing within-arm pnRCTs (which both accounted for clustering and provided an unbiased effect estimate), in the near future an appropriate method of analysis which accounts for clustering may be found. Inflating sample size to account for within-arm clustering (for example using methods developed for pnRCT sample sizes) would in future allow newly developed appropriate analysis methods to be used.

6.4.6 Inclusion of baseline measures

Trials often include baseline measurements of the outcome of interest. Making use of this measure in the analysis by including both the baseline measure and the intervention group as covariates in the analysis model can improve the precision of the effect estimate of treatment by decreasing the variability (intra-subject variance is reduced). This consequently may allow us to study fewer patients for a given power. When the primary outcome is continuous European Medicines Agency (EMA) guidelines specify that the baseline variable should be included as a covariate in the primary analysis regardless of which outcome is chosen in this scenario (either be the raw outcome variable or the change from baseline) [169].

We would need to know how much the adjustment will reduce the standard deviation of the endpoint. If the correlation between outcome variable and the adjustment variable are known, the reduction in standard deviation can be estimated. With one outcome follow-up measure and one baseline measure the adjusted variance using the baseline measure is given by

$$\text{Variance} = \sigma^2(1 - \rho_f^2) \tag{6.29}$$

where ρ_f^2 is the correlation between the baseline and follow-up measure of the outcome. The above variance estimator can be substituted into the relevant sample size formula. To halve the sample size required the standard deviation needs to be halved, this requires a correlation of 0.87 $\left(\sqrt{1-\rho_f^2} = \frac{1}{2} \rightarrow \rho_f = 0.87\right)$.

6.5 Discussion

This chapter has identified and summarised methods for sample size calculations in pnRCTs with continuous outcomes. Methods have been shown to calculate the number of clusters in the intervention arm when there is a fixed cluster size and to calculate the optimum number of controls, incorporating costs, unequal cluster sizes, inclusion of baseline measures. Asymptotic sample size formulae exists for pnRCTs assuming Normal approximation. Exact sample size methods are also available in specific software SPA-ML and clsampsi. The asymptotic methods have been extended to include the simple correction for coefficient of variation to account for variation in cluster size. This is similar to the commonly used method for cRCTs and simple to implement thus should enable widespread use of this approach as opposed to a potentially more complicated method. However, this will likely have small effect on sample size unless the coefficient of variation is large. Further work investigating the coefficient of variations seen in pnRCTs would be of interest to inform appropriate sample size adjustments.

There is a balance to be met between generalisability, internal validity, and pragmatism when designing trials in which we believe the care provider or group effects have an influence on the outcome. As in cRCTs, precision levels in pnRCTs can typically be improved if more clusters and fewer individuals per cluster are sampled for the study.

However, the extra costs of training and recruiting staff to deliver interventions or running extra groups needs to be considered in parallel with the reduction in sample size that increasing the number of clusters and reducing the cluster size will lead to. Although too few clusters will result in unreliable estimates of the ICC and results with limited generalisability to care providers or groups outside the trial.

Chapter 5 evaluated the effect of clustering in within-arm pnRCTs. No method of analysis evaluated was able to account for this within-arm clustering dependent upon intermediate response. However, taking a conservative standpoint it may be appropriate to still inflate sample size to allow for clustering and in the near future an appropriate method of analysis which accounts for clustering may be found.

Sample size calculations require estimates of various measures. This chapter demonstrates that the ICC is key in any trial with clustering. There is a need for more empirical ICC estimates from iRCTs with clustering, the next chapter 7 provides empirical estimates of ICCs from 15 HTA iRCTs with treatment induced clustering.

6.6 Summary

This chapter describes the sample size formulae for pnRCTs with continuous outcomes. Asymptotic sample size formulae exists for pnRCTs assuming Normal approximation. Exact sample size methods have also been identified and explicitly expressed in familiar sample size terminology in this chapter. The asymptotic methods have been extended to include the simple correction for coefficient of variation to account for variation in cluster size.

These methods can be used in conjunction with recommendations from both the simulation chapters which evaluate analysis methods, chapters 4 and 5, and the following chapter 7 which provides empirical ICC estimates and background information to inform future sample size calculations.

Chapter 7

Review of individually randomised trials with clustering

7.1 Introduction

Previous chapters highlighted that clustering commonly occurs in individually randomised trials in health research. Chapter 6 presented methods to account for clustering of outcomes in the design of iRCTs, these methods require an estimate of the ICC. The correlation of outcomes is increasingly being recognised in both the design and analysis of such trials. However, there is poor transparency of the methods used, how clustering is reported in trial reports, and a limited evidence base of ICCs used and observed for future studies. This chapter explores the extent and reporting of intervention induced clustering in HTA funded iRCTs, adds to the evidence of ICCs used in sample size calculations and empirical ICC estimates, and provides exemplars for design and reporting.

Clustering in iRCTs has been recognised in methodology literature [38, 39, 48], this study investigates how this has translated into application in publicly funded trials and provides further evidence of ICCs in such trials. It was anticipated that the increase in publications related to clustering in iRCTS has raised awareness of this area. CONSORT guidelines are often used by trialists not only when reporting trials but also as a guideline when designing and analysing trials to ensure the relevant criteria will be met when reporting results. Additionaly, the CONSORT extension for nonpharmacologic treatments (CONSORT-NPT) [17, 44] states that where applicable clustering by care provider or centre should be accounted for.

This chapter presents a review of the extent and reporting of clustering in iRCTs with intervention induced clustering and provides evidence of ICCs. HTA reports of individually randomised single or multi-centre RCTs are reviewed and trials with potential intervention induced clustering identified. The NIHR is a major funder of research into the clinical and cost effectiveness of healthcare interventions and tests in the UK, with the NIHR HTA funding the largest programme within this. The HTA programme funds both researcher-led and commissioned health related research including RCTs [170]. Information from the HTA trials are extracted, relating to: trial characteristics, sample size, clustering characteristics, information on ICC, and whether clustering was accounted for in each of the design, statistical analysis, and reporting of the trial. Data completeness in relation to relevant CONSORT items is recorded.

The chapter acknowledges the information provided by various corresponding authors and trial statisticians of the included HTA studies. In particular, Beth Stuart of the University of Southampton for providing a preprint of a manuscript of their work investigating clustering at the general practice level in iRCTs in primary care [171].

7.2 Chapter aims

The main aim of this chapter is to determine the extent and quality of reporting of clustering in iRCTs with intervention induced clustering and provide empirical estimates of ICCs. The specific objectives are:

1. review the extent and reporting of clustering in single and multi-centre individually randomised control trials funded and published by the UK's NIHR HTA Programme;
2. add to the evidence on clustering in iRCTs by reporting ICCs for a number of outcomes;
3. provide exemplars of well-reported iRCTs with clustering, which can be used to enhance adequate trial reporting.

7.3 Background

Investigators are frequently hampered when planning the sample size of trials with clustering due to a lack of prior information on the probable size of the ICC [55]. The ICC is an important component of sample size calculation. To better inform the design of trials with clustering

empirical evidence of ICCs are required, from both trials and observational studies. The ICC will vary depending on the outcome (and follow-up time), population, intervention, and setting and the method of analysis [172]. Therefore, a rich resource of estimates are needed from which to choose from to enable consideration of the planned study population and design

Several studies have detailed ICC estimates in various trial settings for cRCTs, examples include: ICCs from primary care [63]; primary and secondary care implementation studies [172]; maternal and paternal health [173]; low and middle income countries [174]; and school based studies [175]. However, fewer studies exist reporting ICC estimates from iRCTs with clustering, examples include: surgical trials [46], psychotherapy trials [57], and general practice level clustering [171].

A small number of studies have identified and reviewed the extent of clustering in iRCTs in different study areas, including: public health and behavioral health journals [43]; orthopaedic surgery [176]; British Medical Journal (BMJ) [41], and surgical trials [46]. Overall these reviews concluded that clustering in iRCTs is highly prevalent and commonly not accounted for in either the design and/or analysis of the trial, resulting in possibly misleadingly precise effect estimates. Of the 42 iRCTs identified in the review by Lee and Thompson [41], 38 had some form of clustering, 17 (40%) with clustering by health professional imposed by the design of the trial and only six of the 38 (16%) mentioned clustering as an issue. Pals et al. [43] reviewed published articles across four public health and behavioural health journals which identified 34 articles reporting results of iRCTs in which treatment was delivered in groups. Thirty two (94%) used analysis at the individual level, ignoring the group level clustering entirely and six (18%) articles reported size of groups/clusters (between 6-12), and three reported number of groups per trial arm.

As indicated previously, ICC estimates do exist from a small number of studies which collate estimates from iRCTs with clustering specifically. Cook et al. [46] calculated ICCs from ten multi-centre surgical trials: 198 ICC estimates for both centre and surgery level. The median (range) number of centres and surgeons was 19 (8-27) and 49 (16-191), respectively. The largest ICCs came from outcomes such as operation time and length of stay (related to cost), hence, they concluded that clustering was likely to have the most impact on economic evaluation. Baldwin et al. [57] calculated ICCs from psychotherapy trials, mainly behavioural or cognitive behavioural. They included 20 studies and report 495 ICC estimates relating to therapists. The number of therapist sessions ranged between 1 to 23, number of therapists ranged between 2 to 581, and average number of patients per therapist ranged from 2.2 to 51.1. General practice

level clustering in iRCTs was investigated in a recent paper by Stuart et al. [171]. The paper evaluated 17 primary care studies to provide ICC estimates by GP practice, concluding that iRCTs in primary care should also take account of clustering in sample size (particularly when cluster sizes are expected to be large). See Appendix D.1 for links to ICC estimates from other studies.

7.3.1 What ICCs are of interest?

In iRCTs there may be different levels of clustering either related to intervention or centre in a multi-centre trial. Details of different types of clustering were presented in more detail in chapter 2 section 2.7.2. The main focus of this study is to investigate intervention induced clustering, however, where potential centre clustering is also present this is included.

7.3.1.1 Intervention induced ICC

Intervention induced clustering refers to the clustering which occurs due to the nature of the intervention itself. For example, a healthcare provider delivered intervention or a group intervention. For intervention induced clustering it can be realistically assumed that ICCs at baseline will often be zero. An instance where this may not be the case could be when more experienced care providers treat patients with worse baseline outcome measures. Consequently, the baseline ICC will be non-zero as an artefact of the design, patients are effectively sorted and more similar ones put in the same group. It is the intervention induced ICCs at follow-up that are generally are of interest as they will effect the precision of the effect estimate thus used when designing iRCTs with intervention induced clustering.

7.3.1.2 Centre ICC

There may be more than two-levels of hierarchy in iRCT data. Large iRCTs in health research are commonly run across multiple centres, for example multiple geographical regions, NHS hospitals or mental health clinics. Multi-centre studies recruit participants across multiple centres to both achieve the required sample size and to improve generalisability of findings. Participants from the same centre may be expected to have similar outcomes implying a positive correlation and possibly the need to account for the centre based clustering.

A multi-centre iRCT design can result in two or more levels of hierarchy in the data. For

example, the Stretching and Strengthening for Rheumatoid Arthritis of the Hand (SARAH) study was a randomised multi-centre trial to evaluate the clinical and cost-effectiveness of an exercise programme in addition to usual care for Rheumatoid Arthritis. The trial included four levels of data hierarchy from: seventeen NHS trusts in England, comprising 21 rheumatology and therapy departments, 48 hand therapists, and finally individual level variance [53]. However, at times it can be difficult to include multiple levels of variability in the analysis model, adding extra levels of variability can result in non-convergence (particularly in small samples).

Randomisation procedures in multi-centre trials often involve permuted blocks stratified by centre. These are used to ensure similar proportions of intervention assignments across the multiple centres, helpful for practical reasons and/or as centre is expected to be confounded with other prognostic factors. EMA guidelines recommend that stratification variables used in the randomisation procedure are adjusted for in the primary analysis:

> "If centre was used for stratification in a multi-centre trial problems might arise in case of many centres recruiting small numbers of patients ('small centres'). Adjusting for many small centres might be possible but raises analytical problems for which there is no best solution. Analyses either ignoring centres used in the randomisation or adjusting for a large number of small centres might lead to unreliable estimates of the intervention effect and P-values that may be either too large or too small." [169, p.7]

When outcomes are continuous there are two key methods of adjusting for centre in the analysis, using models which use fixed centre effect or random centre effect. Including fixed centre effects can be helpful when the centre effects themselves are of interest. However, where there are a large number of centres relative to number of patients it can be difficult as this can involve estimating a large number of parameters compared to the total sample size.

In the literature there are different interpretations of analysis using fixed or random centre effects. Using fixed centre effects, the results are only applicable to the centres included in the trial. When using random centre effect it has been suggested that results are generalisable to centres not included in the trial. However, Kahan and Morris [177] reason that this interpretation of results using random centre effect assumes the centres in a trial are randomly sampled from the population of centres. It is rarely the case that centres are a true random sample, centres are typically chosen based on their willingness, readiness and ability to participate and recruit participants according to trial protocols. Consequently, it can be argued that both fixed and random centre effects adjustments for centres have the same interpretation. As Kahan and

Morris [177, p.1139] notes "any generalisation to patients or centres outside the trial should be carried out on the basis of external validity, rather than on the basis of a particular statistical model".

Random centre effects have been shown to perform comparably or better than fixed centre effects and were robust to non-Normal centre effects and centre outliers [177]. An additional advantage of random centre effects is that clustering is explicitly modelled and thus the different levels of variability in the data can be investigated [40]. When using random centre effects with a small sample size a degrees or freedom correction is recommended to ensure coverage is maintained at nominal levels. Consequently, random centre effects can offer advantages over fixed centre effects.

7.3.2 Reporting guidance for trials with intervention induced clustering

Clear and consistent reporting of trials is vital for readers to understand the design and analysis and to fully interpret the results. This study aimed to review the reporting of clustering in single and multi-centre iRCTs, this was done in relation to relevant CONSORT items. The following is a description of the identification of relevant CONSORT items with features specific to clustering in single and multi-centre iRCTs.

The CONSORT 2010 statement consists of guidelines for reporting parallel group randomised trials [16]. It provides guidance on reporting RCTs, focussing on parallel group trials and includes a 25-item checklist and a participant flow diagram template. The CONSORT statement has been extended for various specific types of designs and interventions. Two CONSORT extensions specifically relevant for iRCTs with clustering are the CONSORT extension to cluster randomised trials (CONSORT-cluster) [34] and the CONSORT-NPT [17, 44].

Other guidelines have crossovers with the CONSORT statements including the Template for Intervention Description and Replication (TIDieR) checklist and Criteria for Reporting the Development and Evaluation of Complex Interventions (CReDECI 2) in healthcare [178]. CONSORT-NPT items 5, 5a, 5b, 5c, 5d are consistent with the TIDieR checklist.

Table 7.1 presents the relevant reporting items from the CONSORT-cluster and CONSORT-NPT. The CONSORT-NPT statement was originally published in 2008 and recently updated in 2017. In extension to the CONSORT 2010 statement for parallel group randomised trials both the CONSORT-cluster and the CONSORT-NPT explicitly instruct authors to, where applicable,

report details of whether and how clustering by care providers or centres was addressed in the sample size and statistical methods, and results (incorporating into the participant flow diagram). A further results section checklist item exists in the CONSORT-cluster, item 17a. Item 17a instructs authors to report a coefficient of intra-cluster correlation (ICC or k) for each primary outcome. An equivalent checklist item for the results section does not exist in the CONSORT-NPT.

Table 7.1: Relevant CONSORT reporting checklist items

Section/ Topic	**Item no.**	**CONSORT-cluster**	**CONSORT-NPT**
Methods			
Participants	4a	None	When applicable, eligibility criteria for centres and for care providers
Sample size	7a	Method of calculation, number of clusters(s) (and whether equal or unequal cluster sizes are assumed), cluster size, a coefficient of intracluster correlation (ICC or k), and an indication of its uncertainty.	When applicable, details of whether and how the clustering by care providers or centres was addressed
Statistical methods	12a	How clustering was taken into account	When applicable, details of whether and how the clustering by care providers or centres was addressed
Results			
Participant flow (diagram is strongly recommended)	13a	For each group, the numbers of clusters that were randomly assigned, received intended intervention, and were analysed for the primary outcome	The number of care providers or centres performing the intervention in each group and the number of patients treated by each care provider or in each centre
Baseline data	15	Baseline characteristics for the individual and cluster levels as applicable for each group	When applicable, a description of care providers (case volume, qualification, expertise, etc.) and centres (volume) in each group.
Outcomes and estimation	17a	Results at the individual or cluster level as applicable and a coefficient of intracluster correlation (ICC or k) for each primary outcome	None

Figure 7.1 presents the explanation for the extension to CONSORT-cluster extension item 17a. A similar explanation could also be reasoned for reporting the ICC for each primary outcome analysed in trials of non-pharmacological treatments (when applicable). It was unclear from the published literature why such a statement was not included. Donner and Klar [55] state that for trials with more than one level of clustering, ICCs at each level should be reported and also

Figure 7.1: Explanation of Item 17a taken from CONSORT statement cluster extension [34]

> "When reporting the results of a cluster randomised trial, point estimates with confidence intervals should be reported for primary outcomes. Given the impact of the extent of the intracluster correlation on the power of the study, the intracluster correlation coefficient or k statistic, for each outcome being analysed should also be provided. This information will allow readers to assess the appropriateness of the original sample size calculations as well as the magnitude of the clustering for each outcome. Showing both adjusted and unadjusted estimates would provide another indication of the extent of the clustering. Several authors have advocated publishing intracluster correlation coefficients to allow them to inform the development of future cluster trials in similar settings."[34, p.10-11]

argue that a complete report should include estimates of between- and within-cluster variance for each ICC estimate. The CONSORT-NPT first author Isabelle Boutron was contacted by myself to discuss their reasoning for not including this item in the original CONSORT-NPT in 2008 or the update in 2017. Boutron fed-back that it was not added as people felt it was already good to have the clustering/ICC addressed in the methods and it does not apply to all NPT trials. Although similar can be said for the CONSORT-NPT methods items 7a and 12a as these do not apply to all NPT trials. Boutron also stated that it would be a very important point to discuss for further updates. Consequently, this study investigates how many trials are already reporting the ICC for each primary outcome analysed and identifies some examples of good practice.

7.4 Methods

This section describes the methods used to conduct the review, providing rationale for the process undertaken.

7.4.1 Trial identification

Reports of iRCTs published in the NIHR HTA Journal from January 2013 to December 2017 were reviewed. A previous review investigating recruitment and retention in trials funded and published by the UK HTA Programme had been conducted by statisticians at the University of Sheffield up to the end of April 2016 [179]. Access was gained to the data from this review in an Excel file. This was updated to cover up to 31st December 2017 and data extraction undertaken relating to clustering and ICCs for all relevant trials between 2013 and 2017.

Trials published between 2013 and 2017 inclusive were chosen as the interest was in recent reporting to understand current practice. It was envisaged reporting and accounting for clustering in recently published trials will be improved compared to past trials due to the higher prevalence of published papers and specific CONSORT guidelines.

HTA publishes research on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. The NIHR HTA funding is a national peer-reviewed funding programme. Reports are published in the HTA Journal if they have resulted from work for the NIHR HTA Programme and they are of a sufficiently high scientific quality as assessed by the external reviewers and journal editors (`https://www.journalslibrary.nihr.ac.uk/hta/about-the-hta-journal.htm`). HTA Journal trial reports were selected for a number of reasons. They are of high quality and provide detailed trial, design and analysis information. HTA Journal reports are comprehensive, hence, it was anticipated that the reporting of design and analysis including the reporting of ICCs in results sections of these trial reports would be generally more detailed than other journal publications and likely to contain ICC if reported at all. Limiting the review to publicly funded trials published in the HTA Journal identified trials from medicine, surgery and therapy from a range of disease areas.

HTA Journal reports were obtained from the NIHR Journals Library website (http://www. journalslibrary.nihr.ac.uk/hta). Efforts were made to check publicly available sources for additional information about identified studies. If published, the International Standardised Randomised Controlled Trial Number (ISRCTN) was used to check the ISRCTN register of clinical trials for any additional information, a trial website or any previously unobtainable trial reports (http://www.isrctn.com/). Author name, study name and ISRCTN number were used to identify published trial papers and trial protocols where available. The HTA Journal report was used as main source of information if any discrepancies in reporting existed.

7.4.2 Inclusion/exclusion criteria

Trials included in the review were single and multi-centre RCTs that were either fully or partially randomised and where recruitment to the trial was finished. Trials were not restricted to iRCTs with clustering in only one arm. Though this design is common, the aim of this review was to capture iRCTs with potential intervention induced clustering regardless of the the number of trial arms this occurred in. Results would thus be more generalisable and useful to a broader range of trials designs. Nested parallel trials as part of another RCT and trial reports of two

or more parallel RCT were included. Trials with a continuous primary endpoint were included (the focus of this thesis).

During review, trials with the following designs were excluded: cluster randomised trials, trials with non-continuous primary endpoints, adaptive designs, and pilot or feasibility trials. Feasibility studies are not designed to provide robust effect estimates, hence, estimates of the ICC from such trials may not be reliable. Additionally, the reporting of feasibility and pilot studies would follow the CONSORT extension for pilot and feasibility trials [180]. Ethics approval was not required for this study as analysis was based on either published results or existing datasets from previously conducted studies (which had appropriate approvals in place).

7.4.3 Data extraction

7.4.3.1 Trial information

The standardised data extraction form from the original University of Sheffield statistics review was obtained from the authors with approriate permissions. It included detailed trial information from the HTA reports. Data from the original form used in this review were: the trial design, the clinical area, intervention type, type of control, number of arms, single or multi-centre and number of centres, recruitment setting, the number and timing of follow-up visits, and the sample size. Data extraction was updated to include trials published up to 31st December 2017 [179]. Trials with potential intervention induced clustering were identified. For these trials additional data were extracted for the purpose of this review relating to clustering and ICCs, including: potential clustering both by healthcare provider/group intervention and by centre, number of clustered arms, whether clustering was recognised, ICCs used in sample size, evidence source for ICCs, results ICC, and information on number and size of clusters.

ICC estimates (and 95% confidence intervals), where available, were taken from trial reports. When observed ICCs were not reported contact was made with the corresponding author to ask them to either provide this information or the corresponding data for relevant calculations, with follow-ups when no initial response was received. If no response was received by 31st June 2017 no further follow-up was taken, this provided time for analysis and writing-up of results. The ICC can be estimated either from an ANOVA or a mixed effects model. An advantage of using a mixed effects model is that covariates can be included in the model. Including any potential confounders as covariates whilst estimating the ICC represents the kind of analysis

that is undertaken in practice. Where possible the ICC was requested to be from an adjusted mixed effects model to reflect the statistical analysis of a trial.

For both intervention induced clustering and centre based clustering the median ICC, interquartile range (IQR) and range for all trials and all outcomes were calculated and specifically for the primary endpoint (primary outcome at primary follow-up). Where the ICC was reported as $<$ or $\leq$ it was rounded up to that number for calculations (for example ≤ 0.001 rounded to 0.001).

7.4.3.2 Reporting quality

Data were also extracted based on the reporting of relevent CONSORT extension items related to clustering. A seven item checklist was developed based on extension checklist items identified from the CONSORT-cluster [34] and CONSORT-NPT [17, 44] statements in section 7.3.2. Items were identified that deserved special consideration when reporting in iRCT with potential intervention induced clustering. Adherence to the following items was recorded:

1. Methods - When applicable, eligibility criteria for centers and for care providers (CONSORT-NPT 4a);
2. Methods - When applicable, details of whether and how the clustering by care providers or centers was addressed (CONSORT-NPT 7a);
3. Methods - When applicable, details of whether and how the clustering by care providers or centers was addressed (CONSORT-NPT 12a);
4. The number of care providers or centers performing the intervention in each group and the number of patients treated by each care provider or in each center (CONSORT-NPT 13a);
5. As 4 above in diagram form (CONSORT-NPT 13a);
6. Results - When applicable, a description of care providers (case volume, qualification, expertise, etc.) and centers (volume) in each group (CONSORT-NPT 15);
7. Results - Where applicable, to report a coefficient of ICC for each primary outcome (based on CONSORT-cluster 17a)

The CONSORT-cluster item 17a was translated for applicability to CONSORT-NPT: "where applicable, to report a coefficient of ICC for each primary outcome". Reporting adherence of

these items was recorded according to the following system of completeness: "absent", "totally complete", "partially complete", "cannot access" and "not applicable". The number and proportion of studies meeting at least partial, compliance in reporting criteria for each checklist item was calculated. Furthermore, a total measure of the number and proportion of checklist items meeting total and at least partial compliance criteria was calculated.

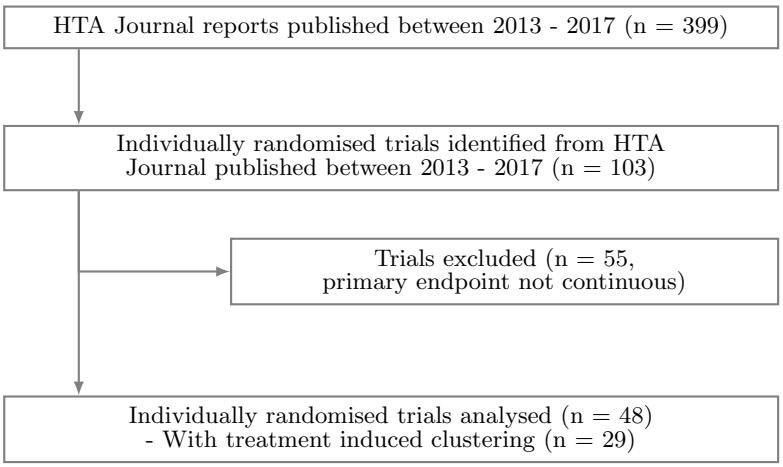
See Appendix D.2 for further details on what the data extraction included.

7.5 Results

7.5.1 Overview

Figure 7.2 details the selection of studies in this review. In total 399 reports were published between January 2013 and December 2017 in the HTA Journal and 103 of these were reports of iRCTs (excluding pilot/feasibility studies). Of the 103 trials, 48 (47%) had a continuous primary endpoint. Twenty nine (60%) of which were categorised as having potential intervention induced clustering. Intervention induced clustering relates to either healthcare provider or group based clustering.

Figure 7.2: Flowchart representing process for a review of trial reports published in the Health Technology Assessment Journal between 2013 and 2017 inclusively.



The 29 trials identified as having potential intervention induced clustering are summarised in the subsequent sections.

7.5.2 Trial characteristics

The characteristics of the 29 trials included in the review are summarised in Table 7.2 and Table 7.3. The majority, 72% (21/29), of trials were parallel group two arm trials. Included trials covered a range of intervetnions: therapy, surgery, complex intervention, and those categorised as other (for example a group weight management programme in the SWAP trial [181]), with therapy the most common intervention type 41% (12/29). No drug intervention trials were identified as having potential intervention induced clustering due to the nature of such trials. Trials were most commonly in the clinical areas of mental health 31% (9/29) and orthopaedics/rheumatology/musculoskeletal 28% (8/29). The trial settings were evenly spread across hospital 28% (8/29) general practice 28% (8/29) and community settings 24% (7/29).

The median number of centres was 17 (IQR 4.5 to 35, range 1 to 94). The median number of intervention induced clusters was 20 (IQR 11 to 70, range 1 to 244), where number of clusters was available in reports for 21 trials explicitly and derived for one trial from the report.

Table 7.2: Summary of studies included

ID	Intervention(s)	Healthcare providers	Number of				
			Participants	Clustered arms	Arms	Centres	Intervention clusters*
1	AESOPS - stepped care intervention for hazardous alcohol users [95]	Step 1 - practice/research nurse. Step 2 - therapist.	529	2	2	51	-
2	AIM - close contact casting vs open surgical reduction and internal fixation for unstable ankle fractures in patients > 60 years [182]	Surgeon	620	2	2	24	100
3	Body psychotherapy (group based) for Schizophrenia patients [183]	Dance movement psychotherapist	356	2	2	5	17
4	Booster - "booster" interventions to sustain increases in physical activity in middle-aged adults in deprived neighbourhoods [184]	Research assistants	282	2	3	1	6
5	BREATHE - self-guided intervention vs. 'face-to-face' physiotherapy for asthma [185]	Physiotherapist	655	1	3	34	1
6	CASPER - low-intensity collaborative care for screen-positive subthreshold depression [186]	Case manager (mental health worker/Improving Access to Psychological Therapies (IAPT) worker)	705	1	2	32	18
7	CASPER plus - low-intensity collaborative care screen-positive major depression [163]	Case manager (mental health worker/IAPT worker)	485	1	2	74	20
8	CHAMP - CBT for health anxiety [187]	Health professional or psychologist	444	1	2	5	11
9	CLASS - foam sclerotherapy vs. endovenous laser ablation vs. surgery for varicose veins [188]	Surgeon	798	3	3	11	-
10	COBRA -behavioural activation (BA) vs. CBT for depression [189]	BA - mental health worker. CBT - psychological therapist.	440	2	2	36	10 & 12
11	eTHoS - stapled haemorrhoidopexy vs. traditional haemorrhoidectomy for haemorrhoids [190]	Surgeon	777	2	2	29	-
12	Families for health - group family programme for overweight children. [13]	Trained facilitators	115	1	2	3	-
13	Getting out the house - outdoor mobility rehabilitation programme for stroke patients [191]	Occupational therapists and physiotherapists	568	1	2	15	29
14	IMPACT - CBT vs. short-term psychoanalytic psychotherapy (STPP) vs. brief psychosocial intervention (BPI) for adolescents with unipolar major depression [192]	CBT - CBT trained staff; STPP - Child Psychotherapist; BPI - therapist.	372	3	3	15	38, 44 & 63

Table 7.2 continued from previous page

ID	Intervention(s)	Healthcare providers	Number of				
			Participants	Clustered arms	Arms	Centres	Intervention clusters*
15	KAT - knee arthroplasty with/without: metal backed tibial component (KATMETAL), mobile bearing between tibial and femoral components (KATMOBILE), patella resurfacing (KATPATELLA) [193]	Surgeon	2352	2	8	34	116
16	OCTET - supported computerised CBT vs. guided self-help vs waiting list for high-intensity CBT for OCD [11]	Psychological wellbeing practitioner	475	2	3	14	93
17	PD REHAB - physiotherapy and occupational therapy for Parkinson's [194]	Physiotherapist and occupational therapist	762	1	2	38	-
18	PEPS - psychoeducation with problem-solving therapy for personality disorder [195]	Mental health nurse/psychology graduates with clinical experience	306	1	2	3	18
19	PhysioDirect - physiotherapist initial assessment and telephone advice, with face-to-face care when necessary for musculoskeletal problems [196]	Physiotherapist	2256	2	2	94	32
20	POWeR+ - brief advice vs. internet-based behavioural intervention with: nurse support face-to-face vs. remote for obesity [197]	Nurse	818	2	3	56	-
21	ProFHER - surgery vs. non-surgical treatment for fracture of the proximal humerus [198]	Surgeon	250	1	2	33	66
22	SARAH - exercise programme for hands and upper limbs for rheumatoid arthritis [53]	Hand therapist	490	1	2	17	48
23	SHEAR - brief advice for excessive alcohol consumption [164]	Clinician	802	1	2	3	79
24	START - manual-based individual therapy for dementia carers [199]	Psychology graduate	260	1	2	4	10
25	STRIDE - CBT for fear of falling in older people [200]	Health-care assistant	415	1	2	3	3
26	SWAP - group weight management programme vs. practice nurse intervention in areas of high social deprivation [181]	Research health psychologist	330	2	2	2	4 & 15
27	TIME-A - music therapy for children with autism	Music therapist [201]	445	3	3	7	-
28	UK DRAFFT - Kirschner-wire fixation vs. locking-plate fixation for fracture of the distal radius [202]	Surgeon	461	2	2	18	244
29	UKUFF - open vs. arthroscopic rotator cuff repair [203]	Surgeon	662	2	2	47	20

* Where more than one number reported this refers to no. clusters in different arms if more than one arm clustered. Some trials with more than one clustered arm reported no. of clusters overall. - : number of clusters not reported.

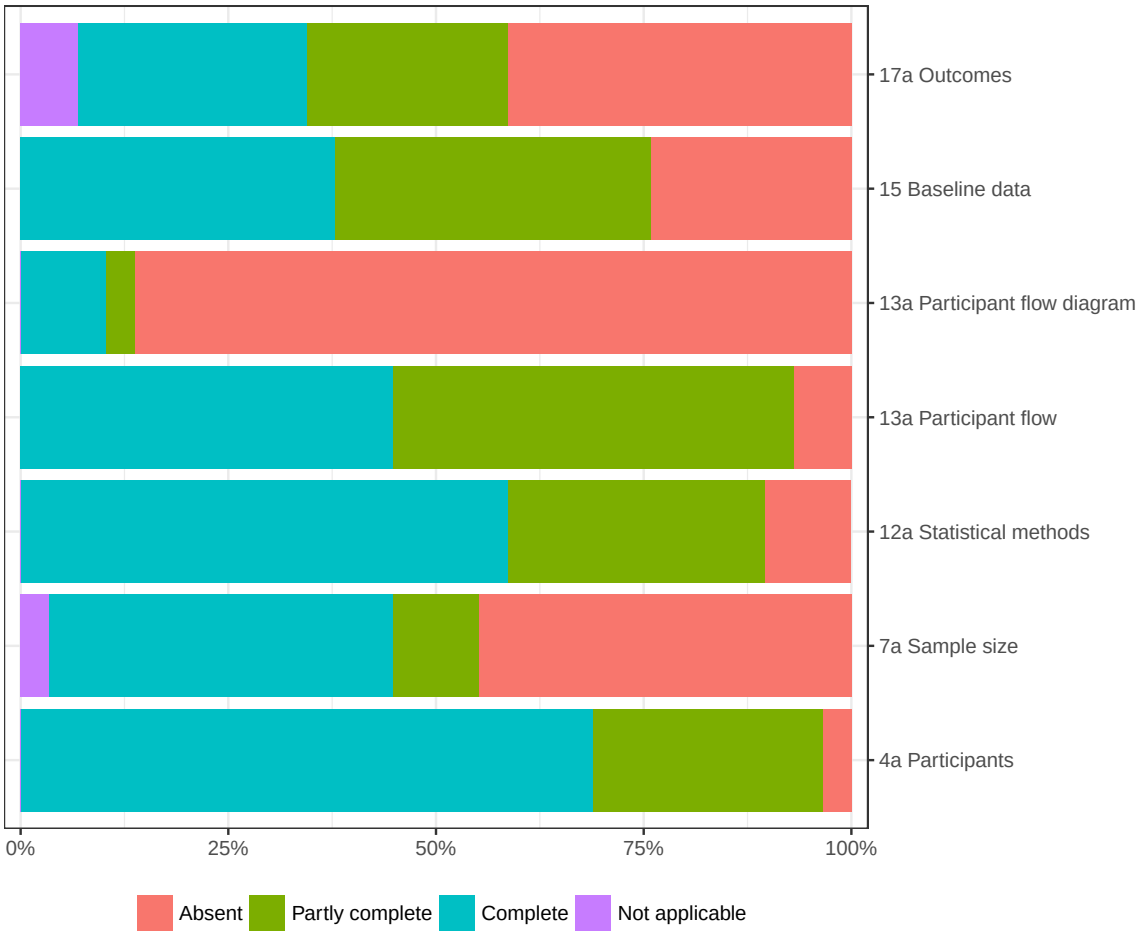
Table 7.3: Trial characteristics

Characteristic	**n (%)**
Trial design	
Parallel	27 (93)
Factorial	1 (3)
Crossover	1 (3)
Study arms	
Two	21 (72)
≥Three	8 (28)
Setting	
Hospital	8 (28)
General practice	8 (28)
Mixed	4 (14)
Community	7 (24)
Other	2 (7)
Therapeutic area	
Mental health	9 (31)
Orthopedics/ Rheumatology/ Musculoskeletal	8 (28)
Obesity	3 (10)
Other	9 (31)
Intervention type	
Therapy	12 (41)
Surgery	5 (17)
Complex intervention	3 (10)
Other	9 (31)
intervention induced clustering	
Care provider	25 (86)
Group & care provider	4 (14)

7.5.3 Reporting of checklist items specific to clustering

Figure 7.3 represents data completeness in relation to CONSORT guidelines and clustering information. As shown in Figure 7.3, suboptimal reporting compliance was observed in items relating to sample size (7a), participant flow diagram (13a) and outcomes (17a). The following proportions exclude items 17a - Outcomes and 13a - Participants flow diagram as these are not explicitly required according to CONSORT-NPT. The proportion of trials meeting at least partially complete reporting (of items 4a, 7a, 12a, 13a, and 15) was 48% (14/29) and meeting complete reporting was 21% (6/29). Regarding the suggested CONSORT-NPT addition item 17a ("were applicable, a coefficient of ICC for each primary outcome"), 24% (7/29) and 28% (8/29) partly and fully reported, respectively. Item 17a is not a checklist item included in CONSORT-NPT at present, it was identified as a key component of reporting for iRCTs with intervention induced clustering. Reporting of item 13a was separately recorded relating to participant flow and the participant flow diagram (which is strongly recommended to report a participants flow diagram in the CONSORT guidelines).

Figure 7.3: CONSORT adherence related to clustering items over all trials included in review



7.5.3.1 Sample size calculations

The trial characteristics related to sample size calculations are summarised in Table 7.4. Among the 29 studies sample size calculations were reported as follows: 52% (15/29) at the individual level, 38% (11/29) took account of ICC at intervention level, 3% (1/29) took account of ICC at centre level, and 3% (1/29) took account of ICC at intervention and centre level. The ProFHER trial [198] was classified as 'other' regarding whether the sample size calculations took account of ICC, the sample size calculation explained that it did not take account of any potential cluster effect as they did not expect there to be many patients treated by individual surgeons. The most common intervention level ICC used in sample size calculations was 0.02 (range between 0.01 to 0.1).

Table 7.4: Trial characteristics relating to sample size calculations

Characteristic	**n (%)**
Sample size calculations	
Reported at individual level	15 (52)
Reported to account for ICC (intervention level)	11 (38)
Reported to account for ICC (centre level)	1 (3)
Reported to account for ICC (centre & intervention level)	1 (3)
Other	1 (3)
Sample size ICC intervention level	
0.01 - 0.025	4 (14)
0.026 - 0.05	5 (17)
$\geq$ 0.06	3 (10)
None	17 (59)
Sample size ICC centre level	
0.04	1 (3)
0.05	1 (3)
None	27 (93)
ICC evidence source	
None reported	10 (34)
Previous trial/s	3 (10)
Previous observational research	1 (3)
Not applicable*	15 (52)

*Sample size calculation did not account for clustering

7.5.3.2 Reasons stated when clustering is not accounted for

This section explains and provides accounts of trials included in this review that did not account for clustering but provided reasoning for this decision. Explanations were identified either through published journal reports or through correspondence with authors and trial statisticians when contacted for more information regarding the ICC.

Three key reasons emerged for not accounting for clustering, including:

1. lots of small clusters;
2. difficult to define clusters;
3. non-convergence mixed effects models with cluster as a random effect.

Some trials provided more than one of these reasons. More detail on each of these reasons is provided in the following paragraphs.

Firstly, a number of the trials either had clusters of size one or very small clusters, hence, the effect of clustering by healthcare provider was expected to be negligible or zero. For instance, the UK DRAFT trial [202] did not include a random effect for surgeon or centre: surgeon because surgeons mainly treat only one person (number of patients treated per surgeon ranged from one to 27) and they used a likelihood ratio test to test for centre effect. However, it is generally recommended not to test for significance of clustering in trials as the test is not powered. Results stated that "any individual surgeon operated only on a small number of patients (n =2 or 3) enrolled in the study; 88% of surgeons (215 out of 244) treated fewer than three study participants. This greatly reduces the likelihood of a surgeon-specific effect on the outcome at any one centre, that is one particularly good or bad surgeon dominating the other surgeons in the study" [202, p.25]. On the opposite end of the spectrum was the BREATHE trial [185] in which all participants involved in the face-to-face physiotherapy arm were treated by a single physiotherapist.

Secondly, was the issue of undefined or hard to define intervention induced clusters (common in primary care and pragmatic trials). This was evident in the PD REHAB trial [194], from further correspondence with the trial statistician they explained that they did not look at the ICC due to the large number of permutations of staff that worked with the patients making it unmanageable. For example, the patient may have seen the same occupational therapist each time, but seen different physiotherapists, or seen by multiple physiotherapists and multiple occupational therapists. The recording of cluster definition can also be an obstacle to accounting for it. The UKUFF trial [203] was undertaken in 19 centres and in many of the centres more than one surgeon participated. The physiotherapy post-surgery was part of routine care, it was not recorded who did it or whether it was all done in the hospital where the surgery was performed. Similarly, the PhysioDirect trial [196] did not collect data related to clustering at the level of the care provider. The corresponding author explained that individual physiotherapists providing

consultations were recorded in the PhysioDirect software held within the NHS, but not within the research data. In addition, participants had contact with more than one provider and given the brief nature of the intervention (assessment and advice) there was not considered to be any likelihood of major intervention-induced clustering affecting the estimates of effect. In the TIME-A trial [201] the NIHR funded part of a larger international trial and data on therapists was not held in this dataset, the methods section specified that clustering was accounted for in the analysis, however, ICC was not reported in the final results.

Thirdly, mixed effects models did not converge when intervention induced clustering was added as a random effect in analysis of some of the trials. For example, the Families for Health trial [13] planned to fit a three-level hierarchical mixed-effects model including a random effect for delivery group and a random effect for family. However, it was stated in the results models comprising the delivery group random effect failed to converge and thus a two-level hierarchical model was used. The AESOPS trial [95] primary analysis compared minimal intervention with stepped care on the primary outcome measure, ADD, at 12 months post randomisation using a mixed model, to account for any variation due to GP practice and the allocated therapist/nurse delivering the intervention. However, the three-level model including the nurse/therapist failed to converge, consequently a two-level mixed model was used with participants nested within GP practice.

Finally, additional explanations included finding the clustering effect had little effect on the results and consequently not including the random effects in the analysis model. The COBRA trial stated in the methods that "although we initially planned to include therapist as a random-effects variable in our models, given the low levels of observed clustering we took a parsimonious approach and fitted our models without inclusion of therapist. We also checked that there was no difference in inference with and without the inclusion of a random-effects therapist term" [189, p.41]. Correspondingly, the results stated "there was evidence of a small, negligible clustering of primary and secondary outcome scores at follow-up across therapists overall and within BA and CBT groups (intracluster correlation coefficient of ≤ 0.04)" [189, p.67]. The SHEAR trial also stated that including the clinician random effect had little effect on the results, and thus it was ignored for the comparison of secondary outcomes [164].

These findings highlight that clustering in iRCTs is not always possible to account for. It may initially be considered during the design stage, however, in the resulting analysis is not possible to include.

7.5.3.3 Empirical ICC estimates

In total there were 221 ICC estimates (39 clustering by centre, 179 clustering related to intervention - 35 group, 1 family and 146 healthcare provider) from 17 studies for which an ICC was either extracted from the original HTA report, further publication or from correspondence with the author. This constituted both primary and secondary outcomes and outcomes reported at multiple follow-ups.

Table 7.5 presents a collated summary of ICC estimates, for all ICC estimates and only those for the primary endpoint, referred to as primary ICC from here on. There were 17 studies providing estimates for 15 intervention induced and eight centre primary ICC estimates. The number of estimates per intervention ranged from 1 to 27 (median 4, IQR 1.5 to 9.5). The median centre based ICC was 0.015 (IQR 0.006 to 0.043). The median healthcare provider ICC was 0.009 (IQR 0.001 to 0.51). The median group based ICC was 0.019 (IQR 0.001 to 0.054). This suggests a small amount of clustering at centre and intervention induced. These are summary estimates, the summaries including all ICCs do not take account of the correlation within studies due to multiple ICC estimates arising from the same study.

A table of all estimates of ICCs from the studies is included in Appendix D.1. An extended table including further information such as sample size, cluster information and status of outcome and follow-up (primary/secondary) has been collected into an Excel sheet, however it was not possible to include all this in the Appendix. This is stored using ORDA - The University of Sheffield Research Data Catalogue and Repository which is hosted on Figshare via `https://figshare.com/s/d2645eb91b3ea9b65e0d`.

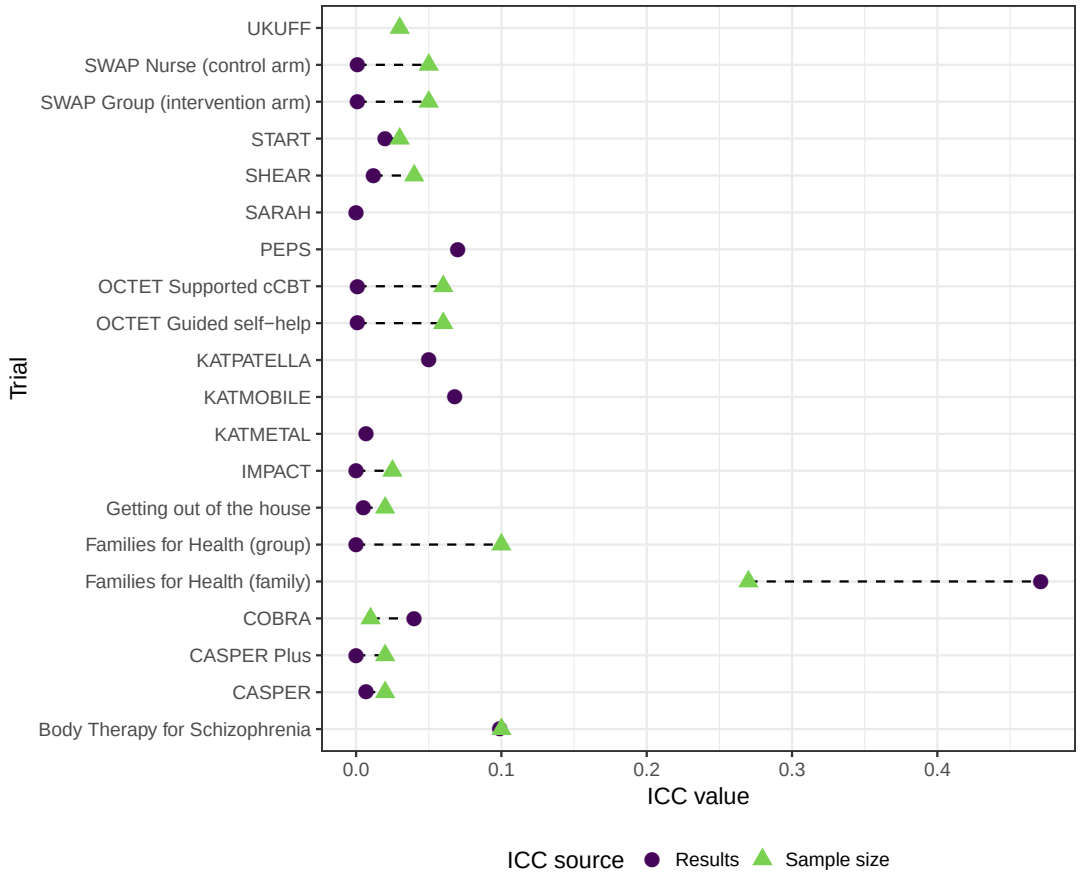
Table 7.5: Summary of ICC estimates (Primary refers to primary endpoint)

Clustering level	**Outcomes**	**n**	**ICC**			
			Median	**IQR**	**Min**	**Max**
Centre	All	39	0.015	(0.006, 0.043)	0.0000	0.4490
Centre	Primary	8	0.014	(0.005, 0.045)	0.0000	0.0730
Healthcare provider	All	146	0.009	(0.001, 0.050)	0.0000	0.678
Healthcare provider	Primary	15	0.007	(0.001, 0.030)	0.000	0.099
Group	All	35	0.019	(0.001, 0.054)	0.000	0.156
Group	Primary	3	0.001	(0.001, 0.036)	0.000	0.070
Family	All/Primary	1	0.471	-	0.471	0.471

Figure 7.4 presents a visual comparison of the ICCs used in sample size matched to the primary

ICC estimate. The majority of ICC estimates were lower than assumed in sample size. Two studies found a higher ICC than assumed in the sample size calculation: the COBRA trial [189] reported an ICC as ≤ 0.04 compared to sample size ICC of 0.01 and the family level ICC in the Families for Health trial was 0.471 compared to sample size ICC of 0.270.

Figure 7.4: Comparison of ICC used in sample size and that found in analysis for intervention induced clustering



7.5.4 Exemplars

One of the aims of this chapter was to explore exemplars of well-reported aspects of iRCTs with clustering, which could be used to enhance adequate trial reporting. None of the publications met full compliance with CONSORT checklist items identified in section 7.3.2. There were, however, some good examples of reporting of iRCTs with potential clustering across the different checklist items. This section seeks to provide exemplars from the HTA studies, for each of the checklist items in turn. This adds to and expands on the examples of adequate reporting provided in the CONSORT-NPT statement [17], now including examples of adequate reporting of checklist item 17a "where applicable, a coefficient of ICC for each primary outcome". A number of tables and figures below have been reproduced from the original HTA reports (with no changes made

and suitable acknowledgement to original reports) adhering to the copyright rules of the 2018 Queen's Printer and Controller of HMSO.

7.5.4.1 Methods - Participants

In the case of reporting eligibility criteria for centres and for care providers (CONSORT-NPT 4a), the CHAMP trial [187] provided a clear description of therapists and the training they received.

> Example: "At each clinic we therefore trained a psychologist, research nurse or equivalent health professional (G-grade or equivalent) to administer the treatment. ...Each therapist attended two workshops at the beginning of the study and received up to 3 months' training from the senior practitioners in the study, sometimes in vivo with two therapists being present in treatment sessions, before taking on the care of patients alone." [187, p.6]

The COBRA trial [189] reported experience and workload of therapists in each trial arm.

> Example: "Ten MHWs provided BA [median 22 participants each (interquartile range 19–25 participants each)] and 12 therapists provided CBT [median 21 participants each (interquartile range 13–23 participants each)]. MHWs had a mean of 18 months' mental health experience (SD 11 months' mental health experience) and CBT therapists had a mean of 22 months' experience (SD 24 months' experience) post CBT qualification. We removed one CBT therapist from the trial in the early stages who did not meet acceptable competency". [189, p.19]

7.5.4.2 Methods - Sample size

A number of studies reported sample size calculations providing details of whether and how clustering by care providers or centres had been addressed. The SWAP trial [181] reported sample size calculations which adhered to CONSORT-NPT, including assumed mean cluster size and ICC.

> Example: "To account for potential clustering effects due to group treatment in the WAP arm, assuming a mean cluster size of 18 and an intracluster correlation coefficient of 0.05, a total of 208 individuals will be required in the WAP arm. The same power can be achieved with 108 in the nurse arm and 216 in the WAP arm, which we increased to 110 in the nurse arm and 220 in the WAP arm to give an allocation ratio between the two arms ($2:1$) that can be expressed in whole numbers. Thus, we required a total of 330 individuals for the entire study." [181, p.20]

The OCTET trial [11] reported sample size calculations with transparency for a crossed-therapist design, therapists deliver interventions in more than one intervention arm.

Example: "The comparison of either supported cCBT or guided self-help is a partially nested design for which the sample size calculation needs to consider the intracluster correlation coefficient (ICC) for therapist. The comparison of supported cCBT with guided self-help is a crossed therapist design, as support for both treatments was delivered by the same therapists. Sample size for crossed therapist design depends on the ICC for therapist for treatment within therapist, which is smaller than the ICC for therapists. Formulae for this calculation are given in Walwyn and Roberts. In the absence of estimates of the two ICCs required for the two calculations, sensitivity of study power to larger values was considered in the calculation below.
Assuming a standard deviation (SD) for the primary outcome (Y-BOCS-OR) at 6 months of 7.3 units, a correlation between baseline Y-BOCS-OR and 6-month Y-BOCS-OR of 0.43, a study with 366 service users followed up to the primary end point has a power $> 80\%$ to detect a difference of 3 Y-BOCS points for each comparison. We were unable to find evidence for a 'clinically important difference'. A reduction of 3 points was agreed based on clinical consensus with the study team. This calculation assumed that supported cCBT and guided self-help were delivered by 24 therapists. It also assumes that the ICC for therapists was 0.06 and an ICC for treatment within therapist was 0.015, which implies that the correlation between the random effect for supported cCBT and guided self-help is 0.75. The design effects, sometimes called the sample size inflation factor, were 1.1225 and 1.06125 for the partially nested and crossed designs, respectively. We considered these values of the ICC to be plausible, but in the event that the ICC for therapist was as large as 0.1 and the ICC for treatment within therapist was 0.05, the power of the trial is still $> 75\%$ for all three comparisons." [11, p.16]

7.5.4.3 Methods - Statistical methods

Details of whether and how the clustering by care providers or centres was addressed in the staistical analysis methods was described in the START trial [199], adjusting for centre and the partially nested design using a mixed-effects model.

Example: "Separate regression analyses were used to estimate group differences in HADS-T score over the short term (using 4- and 8-month follow-ups) and the longer term (using 12- and 24-month follow-ups). In both cases, random-effects models accounted for repeated measurements and therapist clustering in the intervention arm. Adjustments were made for baseline HADS-T scores and centre (on which randomisation was stratified), and also on factors believed from the literature to affect affective symptoms (carer age, sex, carer burden and care recipient neuropsychiatric symptoms)." [199, p.18]

The SWAP trial methods explained clearly that analysis would account for clustering in both arms, with mention of a small sample degrees of freedom correction.

Example: "All analyses accounted for clustering by group in the WAP arm and clustering by nurse in the nurse arm. Each participant has been defined as belonging to a cluster, by which group they belonged to if they were in the intervention arm and by which nurse they were treated by if they were in the control arm. This variable has been included as a random intercept in a mixed-effects regression model. This analysis assumes that the intraclass correlation coefficient is the same between groups in the intervention arm as it is between nurses in the control arm. The Kenward–Roger degree of freedom correction was used for all linear mixed-effects models." [181, p.21]

7.5.4.4 Results - Participant flow

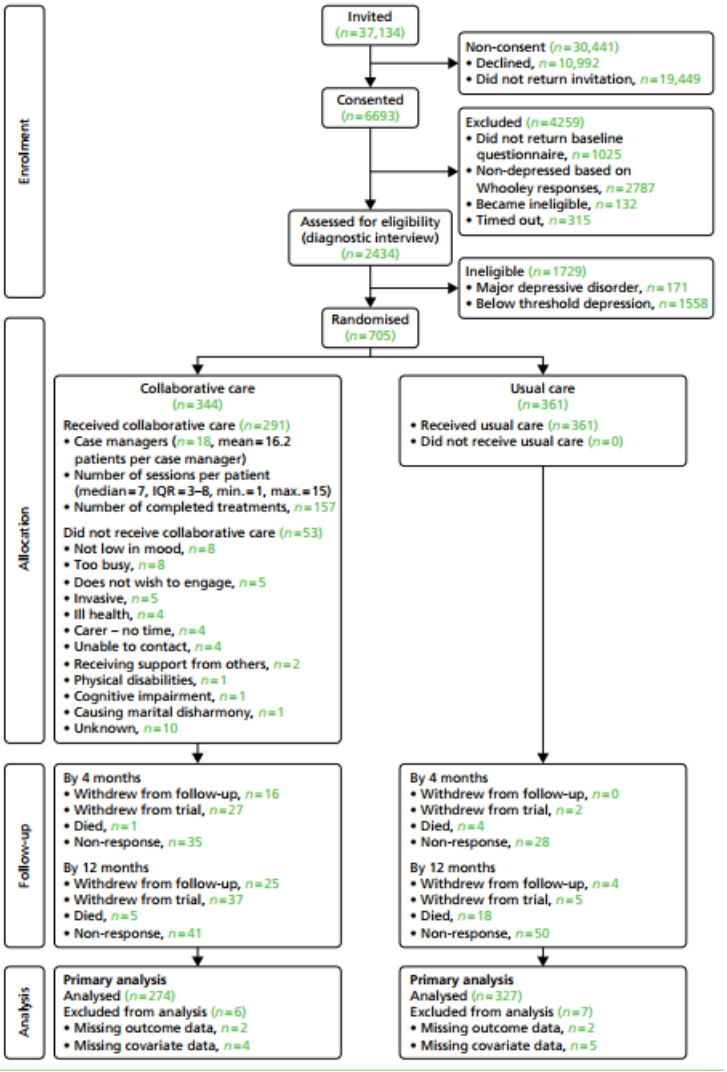
For clarity the results section should report the participant flow, for NPT trials this relates to also reporting the number of care providers or centres performing the intervention in each group and the number of patients treated by each care provider or in each centre. For example, the IMPACT trial [192] provided a description of the care provider workload in both text and a figure of the frequency distribution of number of trial participants for each therapy type seen by a trial therapist.

Example: "All therapists delivering a trial therapy were given a trial therapist identifier. The therapist identifier was missing for 18 (12%) BPI trial treatments, 13 (9%) CBT trial treatments and 2 (1%) STPP trial treatments. A total of 63 therapists delivered BPI, 44 delivered CBT and 38 STPP. For all three modalities, the young person received their trial therapy from a single trial therapist. Figure 6 gives the distribution of the number of young people treated by each therapist for each treatment arm. The number of trial participants seen by a particular therapist ranged from 1 to 15. Forty BPI therapists treated only one young person in the trial, whereas the corresponding figures for CBT and STPP were 19 and 18, respectively. This difference in number of therapists per treatment group is due, in part, to the rather larger number of available BPI compared with CBT or STPP therapists within the 15 NHS CAMHS clinics." [192]

7.5.4.5 Results - Participant flow diagram

Figure 7.5 presents the participant flow diagram used in the CASPER trial report depicting the number of case managers (care providers) performing the intervention in the collaborative care arm. An exemplar related to a surgical trial can be found in the ProFHER HTA trial report (Figure 7) [198].

Figure 7.5: Example of a participant flow diagram depicting iRCT with intervention induced clustering (taken from [186])



7.5.4.6 Results - Baseline data

A clear description of care providers (case volume, qualification, expertise, etc.) and centres (volume) in each group was presented in the report for the UK DRAFFT trial [202]. The surgeon grade, experience by intervention arm was presented in a table and further information regarding operating times etc. presented in figures.

Figure 7.6: Example of a study including results at the individual or cluster level as applicable and an ICC for each primary outcome (taken from Table 9 [186])

TABLE 9 Surgeon grade, experience and methods employed for each operation

Details of surgeons and surgery	K-wire ($n = 230$)	Locking plate ($n = 231$)
Perioperative antibiotic (no : yes)	58 : 164 (71%)	19 : 191 (83%)
Operated wrist (right : left)	101 : 24 (54%)	102 : 124 (54%)
Intraoperative problems (no : yes)	222 : 4 (2%)	219 : 4 (2%)
Surgeon grade, n (%)[a]		
Consultant	60 (26)	71 (31)
Specialist trainee	102 (44)	106 (46)
Staff grade/associate specialist	30 (13)	29 (13)
Other	34 (15)	20 (9)
Surgeon experience (number of prior operations), n (%)[a]		
0	0 (0)	1 (0)
< 5	11 (5)	8 (3)
5–10	16 (7)	23 (10)
11–20	25 (11)	30 (13)
> 20	171 (74)	158 (68)
Wires, n (%) *Number of wires used[a]*		
1	1 (0)	–
2	96 (42)	–
3	105 (46)	–
> 3	5 (2)	–
Wire size[a]		
1.6 mm	187 (81)	–
1.1 mm	1 (0)	–
Other	12 (5)	–
Technique[a]		
Kapandji	54 (23)	–
Interfragmentary	78 (34)	–
Mixed technique	71 (31)	–

7.5.4.7 Results - Outcomes and estimation

Figure 7.7 presents an example of reporting ICCs from the PEPS study [195], which reported an ICC and 95% confidence interval for each primary and secondary outcome at 72 weeks in PEPS arm according to the therapy group, alongside a summary of the size and number of clusters. This provides a clear summary of the level of clustering, it could be further improved by clarifying which models were used to account for this clustering. Another exemplar can be found in Table 8 of the Getting out the House study [191], which reported an ICC for each primary and secondary outcome, however, without precision estimates or number of clusters.

Figure 7.7: Example of study results reporting an ICC for each primary outcome (taken from [195])

Outcome at 72 weeks' follow-up	Estimated intracluster correlation coefficient (95% CI)
SFQ	0.07 (0 to 0.24)
HADS	0.11 (0 to 0.29)
SPSI-R	0.01 (0 to 0.17)
Three main problems	0.01 (0 to 0.15)
Size of cluster in problem-solving group	
As formed on randomisation	
Number of groups	18
Median (25th quartile, 75th quartile)	6.5 (6, 8)
Minimum, maximum	5, 12
Followed up at 72 weeks	
Number of groups	18
Median (25th quartile, 75th quartile)	5 (5, 6)
Minimum, maximum	1, 10

The CASPER trial [186] reported secondary analyses results which adjusted for clustering by case manager, providing ICC and 95% confidence intervals for the primary outcome at each follow-up.

Example: "The average ICC for clustering within case managers was found to be lower than expected (ICC 0.0069, 95% CI 0.0000 to 0.0644, for PHQ-9 scores at 4 months; ICC 0.0072, 95% CI 0.0000 to 0.0676, for PHQ-9 scores at 12 months)." [186, p.36]

7.6 Discussion

7.6.1 Main findings

This study reviews and presents the extent, reporting and evidence of intervention induced clustering (ICC estimates) in single and multi-centre iRCTs funded and published by the UK's NIHR HTA Programme with continuous outcomes. It was found that clustering in iRCTs was often acknowledged and considered in either the design and/or analysis stages. Although, results related to clustering (including number of clusters, range, ICC estimate and 95% confidence interval) were typically not fully reported within the results section.

The reporting of an ICC represents an important contribution to the research literature, both allowing future studies to plan for adequate power and providing evidence of, for example, the role of care provider and centre cluster in the variation in outcome. This chapter has collated and added to the evidence on clustering in iRCTs by reporting ICCs for a number of outcomes that were missed from the original reports. The results of this study have highlighted the importance of recording all elements of the complex intervention to enable clustering to be accounted for, however, it is acknowledged that there are difficulties in this such as access to routine data. It is useful to note that of those who fully reported, the majority had over-estimated the ICC in the sample size. Many trials reported ICC and/or a p-value for the significance of the ICC without a confidence interval. As in other statistical inferences it is good practice to report confidence intervals alongside estimates and p-values. There is generally not enough power to test for statistical significance of ICCs in trials. Consequently, it is of more interest to use a sensitivity analysis for this type of analysis and compare both with/without therapist/cluster effects.

There were various challenges identified related specifically to iRCTs with clustering. Clustering was sometimes acknowledged in the text as a possibility but not directly included in the sample size or analysis methods. This may be for valid reasons, such as expecting very small cluster sizes [198] or trial participants being treat by multiple healthcare providers over the course of an intervention and thus the cluster was difficult to define [194]. Additionally, as shown in chapter 6, sample size methods and corresponding statistical software for iRCTs with clustering are relatively newly developed so awareness may still be limited. Being treated by multiple care providers will also reduce the likelihood of any one care provider having a effect on individual patients outcome. Some trials included a very small number of clusters, and though this may

still result in clustered outcomes it is difficult to account for the clustering in the analysis as estimating an ICC from a small number of clusters can result in bias ICC estimates (as shown in simulation results of chapter 4). Defining therapist/cluster in real life can be challenging as therapists can leave or move roles and be replaced. For example, in the Getting out the House [191] HTA trial two different definitions of therapist were used to evaluate therapist effects: the main therapist was defined as the individual who was the therapist for the initial (assessment) session and an alternative definition of the therapist, for instance for the exercise programme arm, using the therapist who took the second session (first exercise session).

The main overarching factor for not accounting for clustering was the pragmatic nature of the trials included. Many of the interventions had components that were routine practice and not fully documented, which meant that researchers found it hard to define levels of clustering.

7.6.2 Comparison with literature

The extent of potential intervention induced clustering was 59% from trials included in this review. Slightly higher than that found by Lee and Thompson [42] in a review of trials published in the BMJ, 40% had clustering by health professional imposed by the design of the trial.

Reporting has been shown to need improvement across different study designs and therapeutic areas. Samaan et al. [204] found of 50 systematic reviews reporting adherence to reporting guidelines 80.6% reported suboptimal levels (across different clinical areas and study designs). Specifically, Nagendran et al. [205] assessed adherence of RCTs in surgery to the standard CONSORT and CONSORT-NPT guidelines. They identified 54 surgical trials and found of the eight items with less than 30% overall compliance seven were from the CONSORT-NPT extension. These items included: eligibility criteria for centres performing the interventions (24%), how clustering by care providers or centres was addressed as it relates to sample size (6%), how clustering by care providers or centres was addressed as it relates to statistical methods (4%), a description of care providers (case volume, qualification, expertise, etc) and centres (volume) in each group (0%). Adherence to CONSORT-NPT was much poorer than to standard CONSORT, raising awareness of the CONSORT-NPT items and the need to consider and report the role of centres and care providers in trials appears to be a key issue.

The HTA Journal endorses the CONSORT statement. According to the webpage, all RCTs reporting in the HTA Journal are required to submit a CONSORT checklist alongside the report

(reports cannot be reviewed without these forms). Guidelines have been shown to correspond with improved reporting. The CONSORT Statement is endorsed by many medical journals, and CONSORT is part of a wider effort to improve the reporting and quality of health research. Ivers et al. [206] reviewed impact of the 2004 CONSORT-cluster trial extension on the reporting and methodological quality of cluster randomised trials. They identified significant improvements in five of 14 reporting criteria. However, only 18% of the 300 manuscripts [206] reviewed reported an ICC, this was an improvement on previous estimates of 4% and 8%. Of the 29 studies included in this review 24% and 28% partially and fully reported CONSORT-cluster item 17a (related to reporting an ICC where applicable), respectively. This suggests that adding this item to the CONSORT-NPT would not be an undue burden as there are already some studies reporting ICCs at present.

Turning now to the reporting of ICC estimates, estimates of ICCs from iRCTs have been produced for a number of studies and will vary based on the intervention, population and outcome of interest. There are numerous studies reporting ICC estimates relevent for cluster trials [63, 172–175]. However, ICC estimates from cluster trials are related to the unit of randomisation, for example the centre or hospital, and not directly relevant for the ICC estimates for intervention induced clustering. ICC estimates from epidemiological studies and observational research may be more relevant to intervention induced clustering, such as therapist ICCs estimated from observational data [207]. Estimates from observational data may be more precise due to the often large sample sizes, however, estimates may also not be directly relevant to trials in health research. The therapists, conditions, and more manualised interventions used within trials can result in different ICCs to those seen in observational data. ICC estimates from iRCTs of care provider effects include: surgeon ICCs median 0.014 (IQR 0.00 to 0.053 and range 0.000 to 0.514) [46] and therapist ICCs median -0.0255 (IQR -0.114 to 0.078 and range -0.343 to 0.450) [57]. ICCs of centre based effects in multi-centre RCTs include: median ICC of 0.015 (IQR 0.000 to 0.059 and range 0.000 to 0.450) for centres in surgical trials [46] and median ICC of 0.01 (IQR 0.00 to 0.03) for general practice [171]. The ICC estimates collated in this study provide a further resource for trialists planning new studies.

7.6.3 Strengths and limitations

This study adds to the evidence base of observed ICCs in iRCTs. The review includes a range of intervention studies reported within the HTA journal, which includes detailed reports. In

practice, ICC estimates will vary based on the outcome, study design, population, intervention, setting and the method of analysis. Clustering in trials is complex and the exact ICC prior to study completion will never be known. This study does, however, provide further evidence to the potential ICCs that are present in iRCTs and, alongside other studies, can form the basis of future sample size calculations

This study has several limitations. The study was restricted to publicly funded RCTs published as reports in the HTA Journal. Journals adopting CONSORT have been shown to have better reporting standards than others [208] and the limited space available in traditional journal papers (unlike HTA reports) may mean that overall reporting standards in wider published literature are worse than those we found. Data extraction was carried out by a single reviewer for the purpose of this thesis. Some other studies reporting ICCs had access to original data sources for all trials, hence, where able to use consistent methods of ICC estimation across all studies. This study was focussed on investigating the ICCs seen and reported in practice and took a pragmatic viewpoint, extracting empirical ICCs from HTA reports where reported. This resulted in ICC estimates estimated using different methods across different studies (for instance an ANOVA versus mixed effects model). Reporting of the methods or adjustment variables used in the model used to estimate ICCs were not always clear, consequently, it is recommended to clearly report this. ICC estimates calculated with adjustment for important factors such as baseline measures are likely to be lower than unadjusted measures [46].

7.6.4 Implications and future work

The implications of this study are that the amount of clustering in iRCTs may often be small and an unadjusted analysis may result in valid results. However, the true ICC value will not be known during the design stage of the trial. The distribution of the ICCs and full list of ICCs can be used to assist in future sample size calculations and analysis plans. To gain more information on the potential for clustering effect in iRCTs, the clustering level should be at least acknowledged and reported regardless of significance.

There are a number of ways the potential clustering effect can be reduced during the design stage of a trial. When designing studies, using a large number of care providers will likely reduce their clustering effect. For example in the Body Psychotherapy for Schizophrenia patients trial group leaders (psychotherapist or Pilates instructor) were permitted to run a maximum of two groups in an aim to limit the impact of any one group leader on the outcomes [183]. However, practically

this may be a challenge in terms of recruiting, training and coordination of the trial. It is, however, recommended not to use only one healthcare provider or one group where clustering may occur as it will reduce the generalisability of the trial results. Results from such trials cannot be certain that the results will generalise to treatment by a wider range of healthcare providers. Using manualised therapies and interventions will likely reduce the clustering effect by reducing the variation in intervention delivery.

It is recommended, where possible, to use ICC estimates from past studies for sample size calculations. A sensitivity sample size calculation can be undertaken showing power at extremes of ICC values, for instance what would the power of the planned study be at ICC of 0.05 when an ICC of 0.02 was used in the original sample size calculation. When multiple ICC estimates are available a formal meta-analysis would enable good use of available data and achieve greater precision [46, 57].

This study has highlighted that reporting of clustering in iRCTs could be improved. In particular, in relation to the reporting of how clustering was accounted for in the sample size calculations (CONSORT-NPT item 7a), clustering in the results including the participant flow diagram (CONSORT-NPT item 13a) and the reporting of an ICC for each primary outcome (CONSORT-cluster item 17a). Although at present reporting of ICC estimates in results is not included as an item in the CONSORT-NPT. To encourage improvement in the design and reporting of iRCTs with clustering adding a checklist item to CONSORT-NPT is recommended, along the lines of: "when applicable, a coefficient of intracluster correlation for each primary outcome". The framework developed for reporting ICCs in cRCTs [209] is also broadly applicable in iRCTs with clustering (either in the main paper or online supplementary material which are now commonly available in open access journals), identifying three dimensions to consider when reporting an ICC:

- a description of the dataset (including characteristics of the outcome and the intervention);
- how the ICC was calculated;
- and the precision of the ICC.

This has the potential to improve the assumptions about ICCs in iRCTs and raise awareness of the need to account for clustering in both the sample size and analysis in iRCTs with clustering. Finally, in the later stages of this study the CONSORT extension for social and psychological interventions (CONSORT-SPI) [210] was published. This extension only mentions clustering

and ICC in the extension for cRCTs for social and psychological interventions. Clustering in psychological interventions is common [57] and it seems unfortunate that potential clustering is not mentioned in CONSORT-SPI to further raise awareness in this research area.

Reporting guidelines are important to improve the standard of reporting, however, there needs to be concurrent action to increase adherence to relevant guidelines. There is space for improvement in the reporting of ICCs in iRCTs and in future investigations collating ICC estimates (from general and specific study areas). This will, aid the understanding of cluster variability in such trials and enable their use in future study designs with the possibility of meta-analytically combining multiple ICC estimates for sample size calculations.

7.7 Summary

This chapter highlighted the extent of intervention induced clustering in iRCTS with continuous outcomes. For primary endpoints the healthcare provider induced ICC identified was 0.007 (IQR 0.001 to 0.048) and for centre ICC the median was 0.014 (IQR 0.005 to 0.030). The intervention induced ICC used in sample size calculation was typically higher than the empirical ICC estimate from the results, where both were available. Empirical ICC estimates are provided (see Appendix D.3) which can be used in conjunction with sample size formulae in chapter 6.

The exemplars demonstrate some good examples of practice with regards to reporting clustering in iRCTs. Transparency of methods used to design and analyse iRCTs with clustering is important for trialists and to understand generalisability of such trials. To improve transparency an additional CONSORT-NPT item related to reporting ICC where applicable is suggested to bring the results in line with the design and analysis of such trials.

Chapter 8

Discussion

8.1 Introduction

The focus of this thesis was to answer the research question: what elements need to be considered in the design, analysis and reporting of complex intervention trials with continuous outcomes, with a particular focus on proportionate interventions and intervention induced clustering in one trial arm?

The specific aims to address the research question were:

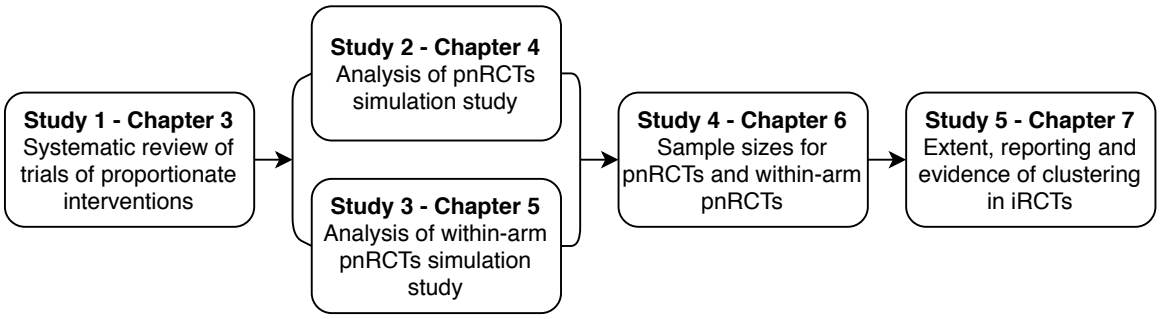
1. To review current practice of how randomised trials of proportionate interventions are designed and analysed and the extent of clustering in such trials.
2. To evaluate commonly used analysis methods for partially nested randomised trials and within-arm partially nested randomised trials to establish which methods are most appropriate and why.
3. To identify and collate a comprehensive summary and resource for sample size methods available for partially nested randomised trials.
4. To determine the extent and quality of reporting of clustering in iRCTs with intervention induced clustering and provide empirical estimates of ICCs.

This chapter summarises the thesis findings and how they address each of the research aims. The results of the research are translated into guidance and recommendations for trialists involved in RCTs of proportionate interventions and RCTs with intervention induced clustering. The chapter concludes with a description of potential future work to build upon this research.

8.2 Main findings and comparison to other work

Figure 8.1 shows an overview of the studies conducted for this thesis.

Figure 8.1: Thesis Overview



8.2.1 Proportionate interventions

The first aim was to review current practice of how randomised trials of proportionate interventions are designed and analysed and the extent of clustering in such trials. This aim was addressed through a systematic review in chapter 3. The proportionate universalism framework enables individuals to receive the care they require and reduces the burden of treatment on an individual whilst reserving resources for those most in need [5]. Proportionate interventions follow this line of thought in aiming to use resources efficiently by providing interventions proportionate to individual need.

The systematic review represented in chapter 3 identified and elaborated on 44 trials of proportionate interventions conducted in various therapeutic areas, the majority in mental health. The trial designs often induced in complex hierarchical data structures, with clustering introduced by both intervention and/or centre, in addition to longitudinal data.

Review findings were summarised into two key sub categories, stepped-care and optimal intervention trials. There appears to be a move towards conducting trials of intervention pathways, with some patients expected to respond to lower intensity interventions thus reducing both the costs to healthcare and the burden of treatment to patients. By evaluating a staged or proportionate intervention as a whole intervention effect using a trial (with randomisation at baseline only) there is an underlying assumption that each of the stages of the intervention is effective. If we were interested in the effect of the individual stages multiple subsequent randomisations would be required possibly leading to the requirement of prohibitively large sample sizes (an issue as many trials struggle to recruit to target [179]). The SMART trial design was used in

a number of the review findings, this has been developed as a method of developing optimal proportionate interventions which are then ideally evaluated in a full confirmatory trial.

Potential intervention induced clustering was identified in the majority, 84%, of the trials of proportionate interventions identified in the review. The intervention induced clustering was typically present at one or more stages of the intervention, often with clustered stages dependent upon outcomes based on previous stages of the intervention. If outcomes are correlated, standard errors will typically be underestimated if this correlation is ignored in the analysis. This clustering was accounted for in the analysis of two of the trials included in review (both published in the HTA Journal), hence appropriate analysis for within-arm partially nested trials was identified as an aim to address in this research.

8.3 Partially nested trials and within-arm partially nested trials

Having highlighted the potential complex hierarchical data structures present in proportionate interventions it prompted the investigation of how to appropriately analyse trials whilst accounting for clustering in the intervention arm. The research presented in chapter 4 and 5 addressed the research aim to evaluate commonly used analysis methods for pnRCTs and within-arm pnRCTs to establish which methods are most appropriate and why. The initial aim was to evaluate methods to analyse pnRCTs (trials with clustering in one arm) and then secondary use these findings to inform evaluation of the more complex partially nested data similar to that seen in the systematic review results in chapter 3, termed within-arm pnRCTs.

Analysis of pnRCTs using a partially nested mixed effects model with the Satterthwaite degrees of freedom correction was shown to be unbiased and maintain confidence interval coverage near to the nominal 95% level. The current findings are broadly consistent with previous simulation studies [48, 122] that demonstrated the partially nested mixed effects model was suitable for analysis under most design scenarios. However, previous research did not assess these methods in a fully systematic manner, the method of classifying the non-clustered controls or performance of models under small ICC ($\rho = 0.01$) which was found in the review of HTA trials in this thesis (chapter 7) and other research [45, 46, 48, 50, 122] commonly occurs in pnRCTs. Unlike findings from Baldwin et al. [122], the Satterthwaite degrees of freedom correction did not fully control the Type I error rate in the simulations.

When ICC was small and/or with very few clusters and small cluster sizes using the partially

nested mixed-effects models 3 and 4 resulted in Type I error rates below the nominal level. These models correctly reflect the design of the trials; however, they can result in conservatism regarding the precision of estimates due to the bias in estimating the variance estimates when there are a small number of clusters. Consequently, using the partially nested mixed effects models with small ICC may indirectly reduce power making it difficult to detect differences between the trial arms when present.

Under homoscedastic individual variances across trial arms the heteroscedastic model is over parametrised, however, it did not result in a substantially lower power than the homoscedastic model. Even under an ICC of zero, no correlation, the heteroscedastic model resulted in a loss in mean statistical power of 2-6% when compared to a linear regression model.

In terms of ICC estimation, when the sample size was small and the true ICC ≤ 0.05, then the ICC was overestimated. ICC estimation improved as sample size increased. Increasing the number of clusters, rather than increasing the cluster size, had a greater increase in power for a fixed total sample size and this will also provide a more accurate estimation of the ICC, similar to cRCTs [40]. Finally the method used to classify the unclustered controls had a negligible impact on results, though using one large cluster for the control arm sped up the model fitting computation time considerably.

Expanding on the work in chapter 4, an example of the more complicated data structures seen in the systematic review and the E-SEE trial were evaluated in chapter 5. Though clustering of partially nested trials has been discussed and evaluated in the literature, to my knowledge no research has directly considered or investigated the more complex multilevel structures of data that are being introduced by trials of proportionate interventions. Schweig and Pane [135] and Roberts [211] evaluated analysis methods for within-arm pnRCTs when the clustering was related to non-compliance. If this non-compliance is random then their methods appeared to hold. However, in the case of proportionate interventions the within-arm clustering is due to the delivery of the intervention and thus clustering is typically non-random.

Simulations of two stage within-arm pnRCTs were undertaken. The intervention comprised a two-stage proportionate intervention with an individual universal stage one intervention and a clustered stage two intervention delivered only to non-responders of stage one. Analysis of this two stage within-arm pnRCT was shown to result in a biased intervention effect if a mixed-effect model, accounting for the clustering of stage two, was used. Neither a linear regression model with cluster robust standard errors or cluster bootstrap standard errors provided coverage of

the effect estimate near the nominal 95% level, although the point estimate was unbiased. This research could not identify a unified approach model for the analysis of within-arm pnRCTs which provided both an overall intervention effect estimate and accounted for clustering. However, if the ICC was ≤ 0.05, then the mean coverage rate of the linear regression model with OLS standard errors was 0.948 (SD 0.006) under the simulation scenarios.

Limited literature or guidance on analysing trials with non-random clustering was found during research for this thesis [52]. However, there needs to be clarity on what effect adjusting for non-random clustering can have on the bias of intervention effect estimates. Statisticians need to exercise caution when adjusting for clustering. For instance, if the aim is to provide an unbiased overall intervention effect and correct precision then statisticians may adjust for clustering, however, this adjusted analysis can result in bias effect estimates.

8.3.1 Sample size methods for partially nested trials

One of the aims of this thesis was to provide a comprehensive summary and resource for sample size methods available for pnRCTs. Several papers [43, 48, 57, 122, 128, 130–133, 136, 137, 162] and a book [130] containing sample size methodology for pnRCTs had been published but none had provided a comprehensive overview and practical resource of the methods available. Chapter 6 summarises and combines relevant sample size methods for pnRCTs with continuous outcomes with links to corresponding software to allow for the design complexities of clustering in trials with continuous outcomes.

The sample size formulae are built up from iRCTs to cRCTs and then to pnRCTs to make the methods clear. The methods have been extended to include the coefficient of variation to account for variable cluster size. This is familiar to the commonly used method for cRCTs thus should enable ease of use of this approach as opposed to a potentially more complicated method. Software for iRCTs with clustering, with options for pnRCTs, were identified and collated.

The practical use of sample formulae require estimates of the relevant input parameters, with the ICC being a key parameter for the design of any trial with potential clustering. With this in mind empirical estimates of ICCs were extracted and presented in chapter 7 and are included in Appendix D.3.

8.3.2 Extent, reporting and evidence of clustering in individually randomised trials

The final aim was to determine the extent and quality of reporting of clustering in iRCTs with intervention induced clustering and provide empirical evidence of the magnitude of ICCs. This was addressed through a review of HTA trials in chapter 7. Of 103 trials published in the HTA journal between 2013 and 2017, 48% had a continuous primary endpoint, of which 59% were categorised as having potential intervention induced clustering. The extent in this review was slightly higher than the 40% of trials which had clustering by health professional found in a review of trials published in the BMJ [42]. This is likely due to the types of interventions funded by the NIHR HTA funding programme in comparison to those published in BMJ, and a possible increase in trials used to evaluate complex interventions.

The review included 29 trials, of these trials clustering was often acknowledged and considered in the design and/or the analysis but to a varying degree. Reporting has been shown to need improvement across different study designs and therapeutic areas [204]. Reporting of CONSORT-NPT checklist items related to clustering were found to be sub-optimal in the review of HTA trials, though improved on the review of surgical trials by Nagendran et al. [205]. This may partly reflect an increased uptake of the CONSORT-NPT, or reflect the increased space available in the HTA journal for reporting compared to traditional journal publications, or reflect improved peer review standards in HTA. It is likely a combination of all three.

Empirical ICC estimates are provided (either from published reports or further correspondence with authors, see Appendix D.3) which can be used as evidence to support assumptions in conjunction with sample size formulae from chapter 6. For primary endpoints the median treatment induced ICC identified was 0.007 (IQR 0.001 to 0.048) and for centre ICC the median was 0.014 (IQR 0.005 to 0.045). These add to the evidence base of empirical ICCs from iRCTs for both care provider effects [46, 57] and centre based effects in multi-centre RCTs [46, 171]. The treatment induced ICC used in sample size calculation was typically higher than the empirical ICC estimate from the results, where both were available. This suggests when clustering was acknowledged in the design, trialsists were actually being relatively conservative in their use of ICC for sample size, this may reflect the uncertainty about the ICC at the design phase and thus conservatism has been used as a precaution.

The examplars in chapter 7 provide examples of good practice when intervention induced clus-

tering is present in RCTs. The importance of transparency is highlighted, both in the design (sample size and analysis methods) and reporting of results. It would be beneficial for the validity of an RCT and for future RCTs to see fully how clustering has been acocunted for and the estimate of clustering effect in the rsults. This could follow the guidelines for reporting ICCs in cRCTs as similar principles apply [209]. To improve transparency an additional CONSORT-NPT item related to reporting ICC where applicable is suggested to bring the results in line with the design and analysis of such trials.

8.4 Strengths and contributions of this research

This research contributes knowledge for the use of those planning and analysing trials in the area of proportionate interventions and with intervention induced clustering.

Robust methodology have been used throughout the thesis. The systematic review, conducted in chapter 3 was implemented using robust methods, including: developing a protocol, scoping search, validation of search terms, data collection form, and following PRISMA guidelines. The methods used to undertake the systematic review also aided the searching undertaken for the review of sample size methodology in chapter 6 and the data collection for chapter 7. The simulation studies in chapters 4 and 5 followed the advice of Burton et al. [144] and Morris et al. [151]. They had a pragmatic focus and covered a range of settings with scenarios covering different cluster sizes, number of clusters, ICCs and under both the null and alternative hypothesis.

This work is likely to have an impact in practice; a strength being its generalisability to commonly employed trial designs. The simulation studies focused on realistic scenarios, and made practical recommendations for trial analysis. The model fitting code for R, SAS and Stata for the analysis methods used in chapters 4 are provided to make the analysis clear. This, along with the findings of previous research, gives strength to the conclusions that the partially nested heteroscedastic mixed effects model is typically the most suitable analysis model for pnRCTs. In addition, it was highlighted that proportionate interventions are initially appealing and come under the remit of proportionate universalism. However, they induce structure that can be difficult to fully accommodate in the analysis.

A final strength is the aim to influence better study design for a new wave of future studies. The collation of sample size methods and ICCs from publicly funded trials provides important

information for future trial designs and providing a broad range of empirical ICC estimates will assist in researchers planning adequately powered trials. The lack of empirical estimates of ICCs directly relevant to iRCTs with intervention induced clustering led to this research. As shown in chapter 6 there are sample size formulae available for pnRCTs thus providing a broad range of empirical ICC estimates will assist in researchers planning adequately powered trials.

Dissemination of results has been considered throughout this PhD. Two manuscripts based on chapters 3 and 4 have been submitted to Trials and BMC Medical research Methodology open access peer reviewed journals [68, 212], the first is currently under review and the latter has been published. In addition, the results of chapters 3, 4 and 5 have been disseminated at a number of statistics and clinical trials conferences, the Meeting of the Society for Clinical Trials in 2018 [213], the joint International Clinical Trials Methodology Conference and Meeting of the Society for Clinical Trials in 2017 [214], and at International Society for Clinical Biostatistics conferences in 2017 and 2016 [215, 216].

8.5 Limitations

There are limits to the generalisability of the thesis findings and some limitations related to resource limitations.

During the initial stages of this project it was considered that the issue of evaluating the effects of different stages of proportionate interventions would be investigated. However, it became clear that randomisation at each stage or some form of fractional factorial design would be required to evaluate each intervention stage of proportionate interventions. There is a large team of researchers investigating these sort of designs in USA thus it did not become the focus of this research. The focus moved towards partial nesting/intervention induced clustering in iRCTs as there was evidence of a lack of awareness of this issue and limited evidence of what to do in such scenarios.

Due to resource limitations it was not possible to supplement the systematic review database searching in chapter 3 with reference list checking or trial registries. The searching and data extraction for both the systematic review and the review of publicly funded trials were typically undertaken by one reviewer (JCa). This is a limitation as it could not be quality-assured during the review process. However, steps were taken to improve the consistency and quality of both reviews and have been explained in more detail in the methods of relevant chapters.

Whilst a broad range of scenarios were considered in the simulation studies, only fixed cluster sizes were used in both studies and only a 50% proportion of within-arm clustering was considered for the within-arm pnRCT study. In practice, cluster size may vary, causing a loss in efficiency when estimating the intervention effect. The proportion that receive subsequent stepped-up intervention stages will vary dependent on the trial, however, the within-arm pnRCT simulation study was able to represent the effect this sort of staged clustering can have in practice.

Trials with clustering in only one arm were considered in the simulation studies and the sample size methodology. Although these designs are common in practice [49], the results are not necessarily generalisable beyond these designs.

It was not possible to gain empirical ICC estimates from all of the HTA trials identified with potential intervention induced clustering. However, a valid reason for the lack of available empirical ICCs was generally provided from correspondence with authors or explanations in HTA reports, either logistical, data limitations, time constraints, or from the nature of the design itself. Where ICC estimates where provided or available they were often for only the primary endpoint and did not extend to other follow-up times or secondary outcomes.

8.6 Implications and recommendations

This section translates the research findings into the implications and a series of recommendations for the design, analysis and reporting of complex intervention trials with continuous outcomes, specifically proportionate interventions and intervention induced clustering in one trial arm. These recommendations are summarised in Figure 8.2 at the end.

8.6.1 Design

In terms of design of trials it is recommended to use specific sample size methods developed for pnRCTs (when clustering is random). Sample size formulae are summarised and collated in chapter 6, providing a useful resource for design of studies. Where available, it is suggested to use prior empirical evidence of probable ICC value in sample size. The fact that this research has shown a median ICC of 0.009 for care provider induced clustering and 0.019 for group induced clustering provides a good starting point for statisticians when discussing the design of a trial. However, the wider generalisability of such findings is cautioned and researchers should aim to consider previous estimates of ICCs calculated from datasets relevant to their study. Sample size

methods should be reported fully and undertaking a sensitivity sample size calculation showing power under different ICC values can contribute to the transparency of the design. Chapter 7 provides some examplars of this linking back to CONSORT-NPT. It was also evident that due to the pragmatic nature of many complex intervention trials there are often hard to define or undefined clusters. Planning during the design phase for better recording of all elements which may be considered part of a complex intervention, even when they are part of routine care, would aid the analysis and improve understanding of the complex intervention delivery and clustering effects. Finally, where proportionate interventions induce clustering based on intermediate outcomes it is recommended to consider the impact this can have on precision of intervention effect estimate (what proportion might be expected to receive this intervention and how the effect of clustering may be reduced) and whether evaluating full intervention effect is suitable.

8.6.2 Analysis

Overarching advice for the analysis of pnRCTs based on this research would be to use a heteroscedastic partially nested mixed effects model in general with a small sample correction, particularly if conservatism and an ICC estimate are desired. However, model choice decision and the requirement or not for conservatism needs to be considered in the context of the specific trial setting. For within-arm pnRCTs with non-random clustering it is recommended to ignore clustering in estimation of overall intervention effect. Taking a pragmatic viewpoint if the cluster effect is small and only some of the intervention arm receive the clustered intervention then the impact of ignoring clustering in the analysis on precision is likely to be minimal. Although, estimation of an ICC from clustered intervention stages would help inform understanding of the intervention. Researchers should be aware of the possible bias in ICC estimation when there are a small number of clusters and true ICCs are small.

8.6.3 Reporting

Trial reports require information on the design and results to enable readers to fully interpret the results. Intervention induced clustering is not always obvious and accounted for, better reporting of trials could feasibly improve awareness of intervention induced clustering. Where applicable, researchers should report their trial findings in accordance with CONSORT-NPT and consider the extension suggested in chapter 7 related to reporting of ICCs for primary outcomes

(as a minimum). It is suggested to add an item to CONSORT-NPT results stating: "when applicable, a coefficient of intracluster correlation for each primary outcome". In addition, full reporting would include: number of clusters, mean (SD) cluster size, ICC and precision of the ICC; adjusted analysis and what method was used to estimate ICC; and when an ICC cannot be estimated during the analysis, a reason should be provided in the trial results reporting. If the reporting of clustering is adopted into routine practice, as is expected from a cRCT, it should improve the interpretation of iRCTs with clustering in the future and improve the evidence base of empirical ICCs for future study design. These results, empirical ICC estimates and exemplars of reporting, will prove useful in the design of trials in the future.

Figure 8.2: Recommendations and considerations for the design, analysis and reporting of complex intervention trials with continuous outcomes, specifically proportionate interventions and intervention induced clustering in one trial arm

Design

- Use specific sample size methods developed for pnRCTs (when clustering is random);
- Where available, use prior empirical evidence of probable ICC value in sample size;
- Report sample size methods fully and undertake sensitivity sample size showing power under different ICC values;
- Transparent and clear recording of all elements which may be considered part of a complex intervention, even when they are part of routine care, including aspects of clustering;
- Where proportionate interventions induce clustering based on intermediate outcomes consider impact on precision of intervention effect estimate and whether evaluating full intervention effect is suitable.

Analysis

- For pnRCTs typically recommended to use heteroscedastic partially nested mixed effects model with small samples degrees of freedom correction;
- For pnRCTs with few clusters, small cluster sizes and small ICC the heteroscedastic partially nested mixed effects model underestimates Type I error rates and there is no optimal model;
- For within-arm pnRCTs with non-random clustering recommended to ignore clustering in estimation of overall intervention effect;
- For within-arm pnRCTs estimation of an ICC from clustered intervention stages is recommended to inform understanding of the intervention.

Reporting

- Report number of clusters, mean (SD) cluster size, ICC and precision of the ICC;
- Report adjusted analysis and what method was used to estimate ICC;
- When ICC cannot be estimated during analysis, provide reasoning in results;
- Add item to CONSORT-NPT - "when applicable, a coefficient of intracluster correlation for each primary outcome".

8.7 Future research

There is potential for further research stemming from this thesis. The area of proportionate or adaptive interventions is a current area of interest as researchers aim to evaluate the effectiveness of intervention pathways which respond to individual need. In addition, the awareness of the different and often multiple levels of clustering present in iRCTs is growing, and this thesis documents only a small portion of the various trial designs and primary outcome measures.

The focus of this thesis was primarily on continuous outcomes, therefore, there is an avenue to further research alternative outcome measures both binary and survival. There has been recent work on the analysis and design of partially nested trials with binary outcomes [217]. This research could be extended to different trial designs, including but not limited to trials with clustering in both arms, more complex crossed-nested trials and investigating the effect of further levels of clustering in within-arm pnRCTs.

Chapter 7 highlighted the variation in cluster size sometimes present in iRCTs with intervention induced clustering. The loss of efficiency and effect of varying cluster size is an area for future research. Candel and Van Breukelen [128] provided sample size adjustments for varying cluster sizes in pnRCTs, they evaluated a number of different scenarios with cv varying from 0.24 to 0.42, suggesting the addition of 11% more clusters to account for the varying cluster size. The review of HTA trials in chapter 7 did suggest that the coefficient of variation in iRCTs with clustering may actually be higher than the range investigated by Candel and Van Breukelen [128], hence, investigating higher variation in cluster size is recommended as an extension to this work.

Work could be extended to investigate the extent of heteroscedasticity in trials with clustering in only one arm. If the ICC is typically quite small, say less than 0.01, we might expect the individual residuals to also be quite similar across trial arms. Empirical work into the extent of heteroscedasticity in partially nested trials would aid statisticians when calculating sample sizes.

Missing data occurs in RCTs through various different mechanisms. Multiple imputation is a commonly used method to handle missing data, it can improve estimation of the precision of parameter estimates by incorporating information from individuals who have missing data (unlike complete case analysis). In cluster trials it is recommended that a multiple imputation model should be multilevel and thus reflect the cluster dependence [218], similar may be true for

iRCTs with clustering, however, further research is required here to provide recommendations.

Various issues relating to reporting and acknowledging of intervention induced clustering have been raised in this thesis. This research has highlighted the gap in reporting guidelines for iRCTs with clustering, findings from chapter 7 advocate adding item 17a to CONSORT-NPT at the next update. In addition, guidance and literature on trials of proportionate or adaptive interventions need to raise awareness to the issue of clustering in such designs as they are being increasingly used. Alongside this it is important to think about how to reduce the effect of clustering as in any study, such as standardising the intervention (though not always suitable) and keeping cluster sizes small.

8.8 Concluding remarks

This thesis has presented an investigation of the elements to be considered in the design, analysis and reporting of complex intervention trials with continuous outcomes, with a particular focus on proportionate interventions and intervention induced clustering in one trial arm. The findings and recommendations will aid trialists designing such trials with intervention induced clustering.

A large amount of research and guidance has been focussed on the design and analysis of cRCTs with continuous outcomes. Now there is raising awareness of the potential effects of clustering in iRCTs, such trials require similar appropriate guidance. This work has provided evidence of the extent of intervention induced clustering both in partially and within-arm partially nested trials. The issue of clustering in trials of both proportionate interventions and the effect such staged clustering can have on increasing Type I error rates has been highlighted. Recommendations based on simulations results are given for the appropriate analysis of both pnRCTs and within-arm pnRCTs. Sample size methods for pnRCTs have been collated and and compared. Relevant reporting related to intervention induced clustering has also been investigated with examplars identified for use to guide future studies.

Bibliography

[1] T. Greenhalgh. *How to implement evidence-based healthcare*. John Wiley & Sons, 2017.

[2] P. Craig, P. Dieppe, S. Macintyre, S. Michie, I. Nazareth, and M. Petticrew. Developing and evaluating complex interventions: new guidance. Medical Research Council. URL: `http://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/`.

[3] National Institute for Health Research (UK). Enhancing Social and Emotional Health in the Early Years. ISRCTN11079129. URL: `http://www.isrctn.com/ISRCTN11079129` (visited on 09/01/2018).

[4] NIHR Public Health Research programme. 13/93 Social and Emotional Wellbeing in Early Years. 2013.

[5] M. G. Marmot, J. Allen, P. Goldblatt, T. Boyce, D. McNeish, M. Grady, I. Geddes, et al. Fair society, healthy lives: Strategic review of health inequalities in England post-2010. The Marmot Review. 2010.

[6] P. Craig, P. Dieppe, S. Macintyre, S. Michie, I. Nazareth, and M. Petticrew. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 337:(2008), p. a1655.

[7] G. Lancaster, M. Campbell, S. Eldridge, A. Farrin, M. Marchant, S. Muller, R. Perera, T. Peters, A. Prevost, and G. Rait. Trials in primary care: statistical issues in the design, conduct and evaluation of complex interventions. *Statistical Methods in Medical Research* 19:4 (2010), p. 349–377.

[8] F. Kühne, R. Ehmcke, M. Härter, and L. Kriston. Conceptual decomposition of complex health care interventions for evidence synthesis: a literature review. *Journal of evaluation in clinical practice* 21:5 (2015), p. 817–823.

[9] M. Petticrew. When are complex interventions 'complex'? When are simple interventions 'simple'? *The European Journal of Public Health* 21:4 (2011), p. 397–398.

[10] D. F. Tate, L. A. Lytle, N. E. Sherwood, D. Haire-Joshu, D. Matheson, S. M. Moore, C. M. Loria, C. Pratt, D. S. Ward, S. H. Belle, et al. Deconstructing interventions: approaches to studying behavior change techniques across obesity interventions. *Translational Behavioral Medicine* 6:2 (2016), p. 236–243.

[11] K. Lovell, P. Bower, J. Gellatly, S. Byford, P. Bee, D. McMillan, et al. Clinical effectiveness, cost-effectiveness and acceptability of low-intensity interventions in the management of obsessive compulsive disorder: the Obsessive Compulsive Treatment Efficacy randomised controlled Trial (OCTET). *Health Technology Assessment* 21:37 (2017).

[12] S. A. Julious, M. J. Horspool, S. Davis, M. Bradburn, P. Norman, N. Shephard, C. L. Cooper, W. H. Smithson, J. Boote, H. Elphick, et al. 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 (2016).

[13] W. Robertson, J. Fleming, A. Kamal, T. Hamborg, K. A. Khan, and F. Griffiths. Randomised controlled trial evaluating the effectiveness and cost-effectiveness of'Families for Health', a family-based childhood obesity treatment intervention delivered in a community setting for ages 6 to 11 years. *Health Technology Assessment* 21:1 (2017).

[14] J. Datta and M. Petticrew. Challenges to evaluating complex interventions: a content analysis of published papers. *BMC Public Health* 13:568 (2013).

[15] M. Campbell, R. Fitzpatrick, A. Haines, A. L. Kinmonth, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ* 321:7262 (2000), p. 694.

[16] K. F. Schulz, D. G. Altman, and D. Moher. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Annals of Internal Medicine* 152:11 (2010), p. 726–732.

[17] I. Boutron, D. G. Altman, D. Moher, K. F. Schulz, and P. Ravaud. CONSORT statement for randomized trials of nonpharmacologic treatments: a 2017 update and a CONSORT extension for nonpharmacologic trial abstracts. *Annals of Internal Medicine* 167:1 (2017), p. 40–47.

[18] NICE. Health inequalities and population health. NICE advice [LGB4]. 2012. URL: `https://www.nice.org.uk/advice/lgb4`.

[19] T. G. Moore, M. McDonald, L. Carlon, and K. O'Rourke. Early childhood development and the social determinants of health inequities. *Health Promotion International*:().

[20] G. Carey, B. Crammond, and E. De Leeuw. Towards health equity: A framework for the application of proportionate universalism. *International Journal for Equity in Health* 14:1 (2015).

[21] G. Thomson, F. Dykes, G. Singh, L. Cawley, and P. Dey. A public health perspective of women's experiences of antenatal care: An exploration of insights from a community consultation. *Midwifery* 29:3 (2013), p. 211–216.

[22] S. Thomsen, D. T. P. Hoa, M. Målqvist, L. Sanneving, D. Saxena, S. Tana, B. Yuan, and P. Byass. Promoting equity to achieve maternal and child health. *Reproductive Health Matters* 19:38 (2011), p. 176–182.

[23] J. Morrison, H. Pikhart, M. Ruiz, and P. Goldblatt. Systematic review of parenting interventions in European countries aiming to reduce social inequalities in children's health and development. *BMC Public Health* 14:1040 (2014).

[24] V. Maharaj, F. Rahman, and L. Adamson. Tackling child health inequalities due to deprivation: using health equity audit to improve and monitor access to a community paediatric service. *Child: Care, Health and Development* 40:2 (2014), p. 223–230.

[25] S. Cowley, K. Whittaker, M. Malone, S. Donetto, A. Grigulis, and J. Maben. Why health visiting? Examining the potential public health benefits from health visiting practice within a universal service: A narrative review of the literature. *International Journal of Nursing Studies* 52:1 (2015), p. 465–480.

[26] J. Welsh, L. Strazdins, L. Ford, S. Friel, K. O'Rourke, S. Carbone, and L. Carlon. Promoting equity in the mental wellbeing of children and young people: a scoping review. *Health Promotion International* 30:suppl 2 (2015), p. ii36–ii76.

[27] J. Benach, D. Malmusi, Y. Yasui, and J. M. Martınez. A new typology of policies to tackle health inequalities and scenarios of impact based on Rose's population approach. *Journal of Epidemiology and Community Health* 67:3 (2013), p. 286–91.

[28] D. Almirall, I. Nahum-Shani, N. E. Sherwood, and S. A. Murphy. Introduction to SMART designs for the development of adaptive interventions: With application to weight loss research. *Translational Behavioral Medicine* 4:3 (2014), p. 260–274.

[29] I. Nahum-Shani, M. Qian, D. Almirall, W. E. Pelham, B. Gnagy, G. A. Fabiano, J. G. Waxmonsky, J. Yu, and S. A. Murphy. Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological Methods* 17:4 (2012).

[30] D. Almirall and A. Chronis-Tuscano. Adaptive interventions in child and adolescent mental health. *Journal of Clinical Child & Adolescent Psychology* 45:4 (2016), p. 383–395.

[31] R. Hayes and L. Moulton. *Cluster Randomised Trials*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis, 2009.

[32] J. Rothwell, C. Cooper, and S. Julious. Quantifying Effect Sizes in Clinical Trials. PhD thesis. University of Sheffield, 2018.

[33] S. A. Julious. *Sample sizes for clinical trials*. CRC Press, 2009.

[34] M. K. Campbell, D. R. Elbourne, and D. G. Altman. CONSORT statement: extension to cluster randomised trials. *BMJ* 328:7441 (2012), p. 702–708.

[35] J. Hutchings, T. Bywater, D. Daley, F. Gardner, C. Whitaker, K. Jones, C. Eames, and R. T. Edwards. Parenting intervention in Sure Start services for children at risk of developing conduct disorder: pragmatic randomised controlled trial. *BMJ* 334:7595 (2007), p. 678.

[36] C. J. Morrell, S. J. Walters, S. Dixon, K. A. Collins, L. M. Brereton, J. Peters, and C. G. Brooker. Cost effectiveness of community leg ulcer clinics: randomised controlled trial. *BMJ* 316:7143 (1998), p. 1487.

[37] M. K. Diener, C. M. Seiler, M. von Frankenberg, K. Rendel, S. Schüle, K. Maschuw, S. Riedl, J. C. Rückert, C. Eckmann, U. Scharlau, et al. Vascular clips versus ligatures in thyroid surgery—results of a multicenter randomized controlled trial (CLIVIT Trial). *Langenbeck's Archives of Surgery* 397:7 (2012), p. 1117–1126.

[38] R. Walwyn and C. Roberts. Therapist variation within randomised trials of psychotherapy: implications for precision, internal and external validity. *Statistical Methods in Medical Research* 19:3 (2010), p. 291–315.

[39] C. Roberts and R. Walwyn. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. *Statistics in Medicine* 32:1 (2013), p. 81–98.

[40] M. J. Campbell and S. J. Walters. *How to design, analyse and report cluster randomised trials in medicine and health related research*. John Wiley & Sons, 2014.

[41] K. J. Lee and S. G. Thompson. Clustering by health professional in individually randomised trials. *BMJ* 330:7483 (2005), p. 142.

[42] K. J. Lee and S. G. Thompson. The use of random effects models to allow for clustering in individually randomized trials. *Clinical Trials* 2:2 (2005), p. 163–173.

[43] S. L. Pals, D. M. Murray, C. M. Alfano, W. R. Shadish, P. J. Hannan, and W. L. Baker. Individually randomized group treatment trials: a critical appraisal of frequently used design and analytic approaches. *American Journal of Public Health* 98:8 (2008), p. 1418–1424.

[44] I. Boutron, D. Moher, D. G. Altman, K. F. Schulz, and P. Ravaud. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Annals of Internal Medicine* 148:4 (2008), p. 295–309.

[45] L. Flight, A. Allison, M. Dimairo, E. Lee, M. Mandelfield, and S. Walters. Recommendations for the analysis of individually randomised controlled trials with clustering in one arm - a case of continuous outcomes. *BMC Medical Research Methodology* 16:165 (2016).

[46] J. A. Cook, T. Bruckner, G. S. MacLennan, and C. M. Seiler. Clustering in surgical trials-database of intracluster correlations. *Trials* 13:2 (2012).

[47] S. Grant, E. Mayo-Wilson, P. Montgomery, G. Macdonald, S. Michie, S. Hopewell, and D. Moher. CONSORT-SPI 2018 Explanation and Elaboration: guidance for reporting social and psychological intervention trials. *Trials* 19:406 (2018).

[48] C. Roberts and S. A. Roberts. Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials* 2:2 (2005), p. 152–162.

[49] D. J. Bauer, S. K. Sterba, and D. D. Hallfors. Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research* 43:2 (2008), p. 210–236.

[50] R. Gossage-Worrall, R. I. Holt, K. Barnard, M. Carey, M. Davies, C. Dickens, Y. Doherty, C. Edwardson, P. French, F. Gaughran, et al. STEPWISE–STructured lifestyle Education for People WIth SchizophrEnia: a study protocol for a randomised controlled trial. *Trials* 17:475 (2016).

[51] S. Eldridge and S. Kerry. *A practical guide to cluster randomised trials in health services research*. John Wiley & Sons, 2012.

[52] S. Landau and T. Chalder. Clustering issues in trials: Cluster randomised trials, therapist effects and group treatments. Presentation at King's Trials Partnership 12th Dec 2012. 2012. URL: `https://www.kcl.ac.uk/ioppn/depts/biostatisticshealthinformatics/ktp/previous-events/2012/december/landau---clustering-issues-in-trials.pdf`.

[53] M. A. Williams, E. M. Williamson, P. J. Heine, V. Nichols, M. J. Glover, and M. Dritsaki. Strengthening And stretching for Rheumatoid Arthritis of the Hand (SARAH). A randomised controlled trial and economic evaluation. *Health Technology Assessment* 19:19 (2015).

[54] B. C. Kahan and T. P. Morris. Assessing potential sources of clustering in individually randomised trials. *BMC Medical Research Methodology* 13:58 (2013).

[55] A. Donner and N. Klar. *Design and analysis of cluster randomization trials in health research*. John Wiley and Sons Ltd, 2000.

[56] S. M. Eldridge, O. C. Ukoumunne, and J. B. Carlin. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. *International Statistical Review* 77:3 (2009), p. 378–394.

[57] S. A. Baldwin, D. M. Murray, W. R. Shadish, S. L. Pals, J. M. Holland, J. S. Abramowitz, G. Andersson, D. C. Atkins, P. Carlbring, K. M. Carroll, et al. Intraclass correlation associated with therapists: estimates and applications in planning psychotherapy research. *Cognitive Behaviour Therapy* 40:1 (2011), p. 15–33.

[58] A. C. Cameron and D. L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources* 50:2 (2015), p. 317–372.

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

[60] P. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002.

[61] C. Leyrat, K. E. Morgan, B. Leurent, and B. C. Kahan. Cluster randomized trials with a small number of clusters: which analyses should be used? *International Journal of Epidemiology*:(2017).

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

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

[64] R. Chu, L. Thabane, J. Ma, A. Holbrook, E. Pullenayegum, and P. J. Devereaux. Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study. *BMC Medical Research Methodology* 11:(2011), p. 21.

[65] NHS Scotland. Proportionate Universalism Briefing. 2014. URL: `http://www.healthscotland.com/uploads/documents/24296-ProportionateUniversalismBriefing.pdf` (visited on 04/29/2016).

[66] P. Hutt and S. Gilmour. Tackling inequalities in general practice. *London: The King's Fund*:(2010), p. 1–37.

[67] D. Lu and I. Tyler. Focus On: A Proportionate Approach to Priority Populations. Public Health Ontario. URL: `https://www.publichealthontario.ca/en/eRepository/Focus_On_Priority_Populations.pdf` (visited on 04/29/2016).

[68] J. Candlish, M. D. Teare, M. Dimairo, L. Flight, L. Mandefield, and S. J. Walters. Appropriate statistical methods for analysing partially nested randomised controlled trials with continuous outcomes: a simulation study. *BMC Medical Research Methodology* 18:105 (2018).

[69] Cohcrane Collaboration. Cochrane Highly Sensitive Search Strategy for identifying randomized trials in MEDLINE: sensitivity- and precision-maximizing version (2008 revision); Ovid format. URL: `http://handbook.cochrane.org/chapter_6/box_6.4.d_cochrane_hsss_2008_sensprec_ovid.htm`.

[70] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine* 151:4 (2009), p. 264–269.

[71] K. Ell, W. Katon, B. Xie, P.-J. Lee, S. Kapetanovic, J. Guterman, and C.-P. Chou. Collaborative care management of major depression among low-income, predominantly Hispanic subjects with diabetes A randomized controlled trial. *Diabetes Care* 33:4 (2010), p. 706–713.

[72] P. van't Veer-Tazelaar, F. Smit, H. van Hout, P. van Oppen, H. van der Horst, A. Beekman, and H. van Marwijk. Cost-effectiveness of a stepped care intervention to prevent depression and anxiety in late life: randomised trial. *The British Journal of Psychiatry* 196:4 (2010), p. 319–325.

[73] A. M. J. Braamse, B. van Meijel, O. Visser, P. van Oppen, A. D. Boenink, C. Eeltink, P. Cuijpers, P. C. Huijgens, A. T. F. Beekman, and J. Dekker. Distress and quality of life after autologous stem cell transplantation: a randomized clinical trial to evaluate the outcome of a web-based stepped care intervention. *BMC Cancer* 10:361 (2010).

[74] V. Patel, H. A. Weiss, N. Chowdhary, S. Naik, S. Pednekar, S. Chatterjee, M. J. De Silva, B. Bhat, R. Araya, M. King, et al. Effectiveness of an intervention led by lay health counsellors for depressive and anxiety disorders in primary care in Goa, India (MANAS): a cluster randomised controlled trial. *The Lancet* 376:9758 (2010), p. 2086–2095.

[75] C. M. Gilliam, G. J. Diefenbach, S. E. Whiting, and D. F. Tolin. Stepped care for obsessive-compulsive disorder: An open trial. *Behaviour Research and Therapy* 48:11 (2010), p. 1144–1149.

[76] F. J. Kay-Lambkin, A. L. Baker, R. McKetin, and N. Lee. Stepping through treatment: Reflections on an adaptive treatment strategy among methamphetamine users with depression. *Drug and Alcohol Review* 29:5 (2010), p. 475–482.

[77] D. Richter, C. Mickel, S. Acharya, P. Brunel, and C. Militaru. Aliskiren-based stepped-care treatment algorithm provides effective blood pressure control. *International Journal of Clinical Practice* 65:5 (2011), p. 613–623.

[78] R. D. Weiss, J. S. Potter, D. A. Fiellin, M. Byrne, H. S. Connery, W. Dickinson, J. Gardin, M. L. Griffin, M. N. Gourevitch, D. L. Haller, et al. Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence: a 2-phase randomized controlled trial. *Archives of General Psychiatry* 68:12 (2011), p. 1238–1246.

[79] J. E. Mitchell, S. Agras, S. Crow, K. Halmi, C. G. Fairburn, S. Bryson, and H. Kraemer. Stepped care and cognitive–behavioural therapy for bulimia nervosa: randomised trial. *The British Journal of Psychiatry* 198:5 (2011), p. 391–397.

[80] W. Seekles, A. van Straten, A. Beekman, H. van Marwijk, and P. Cuijpers. Stepped care treatment for depression and anxiety in primary care. a randomized controlled trial. *Trials* 12:171 (2011).

[81] D. F. Tolin, G. J. Diefenbach, and C. M. Gilliam. Stepped care versus standard cognitive–behavioral therapy for obsessive–compulsive disorder: A preliminary study of efficacy and costs. *Depression and Anxiety* 28:4 (2011), p. 314–323.

[82] A. J. van der Leeden, B. M. van Widenfelt, R. van der Leeden, J. M. Liber, E. M. Utens, and P. D. Treffers. Stepped care cognitive behavioural therapy for children with anxiety disorders: A new treatment approach. *Behavioural and Cognitive Psychotherapy* 39:01 (2011), p. 55–75.

[83] S. R. Apil, E. Hoencamp, P. M. Haffmans Judith, and P. Spinhoven. A stepped care relapse prevention program for depression in older people: a randomized controlled trial. *International Journal of Geriatric Psychiatry* 27:6 (2012), p. 583–591.

[84] J. F. Karp, B. L. Rollman, C. F. Reynolds, J. Q. Morse, F. Lotrich, S. Mazumdar, N. Morone, and D. K. Weiner. Addressing both depression and pain in late life: the methodology of the ADAPT study. *Pain Medicine* 13:3 (2012), p. 405–418.

[85] S. M. Shortreed and E. E. Moodie. Estimating the optimal dynamic antipsychotic treatment regime: evidence from the sequential multiple-assignment randomized Clinical Antipsychotic Trials of Intervention and Effectiveness schizophrenia study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61:4 (2012), p. 577–599.

[86] E. Dozeman, H. W. van Marwijk, D. J. van Schaik, F. Smit, M. L. Stek, H. E. van der Horst, E. T. Bohlmeijer, and A. T. Beekman. Contradictory effects for prevention of depression and anxiety in residents in homes for the elderly: a pragmatic randomized controlled trial. *International Psychogeriatrics* 24:08 (2012), p. 1242–1251.

[87] K. Nordin, R. Rissanen, J. Ahlgren, G. Burell, M.-L. Fjällskog, S. Börjesson, and C. Arving. Design of the study: How can health care help female breast cancer patients reduce their stress symptoms? A randomized intervention study with stepped-care. *BMC Cancer* 12:167 (2012).

[88] J. M. Jakicic, D. F. Tate, W. Lang, K. K. Davis, K. Polzien, A. D. Rickman, K. Erickson, R. H. Neiberg, and E. A. Finkelstein. Effect of a stepped-care intervention approach on weight loss in adults: a randomized clinical trial. *JAMA* 307:24 (2012), p. 2617–2626.

[89] L. Wang, A. Rotnitzky, X. Lin, R. E. Millikan, and P. F. Thall. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association* 107:498 (2012), p. 493–508.

[90] A. M. Pommer, F. Pouwer, J. Denollet, and V. J. Pop. Managing co-morbid depression and anxiety in primary care patients with asthma and/or chronic obstructive pulmonary disease: study protocol for a randomized controlled trial. *Trials* 13:6 (2012).

[91] S. E. Lamb, M. A. Williams, E. M. Williamson, S. Gates, E. J. Withers, and S. Mt-Isa. Managing Injuries of the Neck Trial (MINT): a randomised controlled trial of treatments for whiplash injuries. *Health Technology Assessment* 16:49 (2012).

[92] A.-M. H. Krebber, C. R. Leemans, R. de Bree, A. van Straten, F. Smit, E. F. Smit, A. Becker, G. M. Eeckhout, A. T. Beekman, P. Cuijpers, et al. Stepped care targeting psychological distress in head and neck and lung cancer patients: a randomized clinical trial. *BMC Cancer* 12:173 (2012).

[93] B. Borsari, J. T. P. Hustad, N. R. Mastroleo, T. O. Tevyaw, N. P. Barnett, C. W. Kahler, E. E. Short, and P. M. Monti. Addressing alcohol use and problems in mandated college students: a randomized clinical trial using stepped care. *Journal of Consulting and Clinical Psychology* 80:6 (2012), p. 1062–1074.

[94] J. E. Rose and F. M. Behm. Adapting smoking cessation treatment according to initial response to precessation nicotine patch. *The American Journal of Psychiatry* 170:8 (2013), p. 860–867.

[95] J. Watson, H. Crosby, V. Dale, G. Tober, Q. Wu, J. Lang, et al. AESOPS: a randomised controlled trial of the clinical effectiveness and cost-effectiveness of opportunistic screening and stepped care interventions for older hazardous alcohol users in primary care. *Health Technology Assessment* 17:25 (2013).

[96] D. B. Oosterbaan, M. J. Verbraak, B. Terluin, A. W. Hoogendoorn, W. J. Peyrot, A. Muntingh, and A. J. van Balkom. Collaborative stepped care versus care as usual for common mental disorders: 8-month, cluster randomised controlled trial. *The British Journal of Psychiatry* 203:2 (2013), p. 132–139.

[97] S. E. Van Dijk, A. D. Pols, M. C. Adriaanse, J. E. Bosmans, P. J. Elders, H. W. Van Marwijk, and M. W. Van Tulder. Cost-effectiveness of a stepped-care intervention to prevent major depression in patients with type 2 diabetes mellitus and/or coronary heart disease and subthreshold depression: design of a cluster-randomized controlled trial. *BMC Psychiatry* 13:128 (2013).

[98] C. Arving, I. Thormodsen, G. Brekke, O. Mella, S. Berntsen, and K. Nordin. Early rehabilitation of cancer patients - a randomized controlled intervention study. *BMC Cancer* 13:9 (2013).

[99] S. Mattsson, S. Alfonsson, M. Carlsson, P. Nygren, E. Olsson, and B. Johansson. U-CARE: Internet-based stepped care with interactive support and cognitive behavioral therapy for reduction of anxiety and depressive symptoms in cancer–a clinical trial protocol. *BMC Cancer* 13:414 (2013).

[100] R. A. Carels, D. A. Hoffmann, N. Hinman, J. M. Burmeister, A. Koball, L. Ashrafioun, M. W. Oehlhof, E. Bannon, M. Leroy, and L. Darby. Step-down approach to behavioural weight loss treatment: a pilot of a randomised clinical trial. *Psychology & Health* 28:10 (2013), p. 1121–1134.

[101] H. P. van der Aa, G. H. Van Rens, H. C. Comijs, J. E. Bosmans, T. H. Margrain, and R. M. van Nispen. Stepped-care to prevent depression and anxiety in visually impaired older adults–design of a randomised controlled trial. *BMC Psychiatry* 13:209 (2013).

[102] C. Kasari, A. Kaiser, K. Goods, J. Nietfeld, P. Mathy, R. Landa, S. Murphy, and D. Almirall. Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry* 53:6 (2014), p. 635–646.

[103] A. Muntingh, C. van der Feltz-Cornelis, H. van Marwijk, P. Spinhoven, W. Assendelft, M. de Waal, H. Ader, and A. van Balkom. Effectiveness of collaborative stepped care for anxiety disorders in primary care: a pragmatic cluster randomised controlled trial. *Psychotherapy and Psychosomatics* 83:1 (2014), p. 37–44.

[104] A. M. Kilbourne, D. Almirall, D. Eisenberg, J. Waxmonsky, D. E. Goodrich, J. C. Fortney, J. E. Kirchner, L. I. Solberg, D. Main, M. S. Bauer, et al. Protocol: Adaptive Implementation of Effective Programs Trial (ADEPT): cluster randomized SMART trial comparing a standard versus enhanced implementation strategy to improve outcomes of a mood disorders program. *Implement Science* 9:132 (2014).

[105] K. M. Hamall, T. R. Heard, K. J. Inder, K. M. McGill, and F. Kay-Lambkin. The Child Illness and Resilience Program (CHiRP): a study protocol of a stepped care intervention to improve the resilience and wellbeing of families living with childhood chronic illness. *BMC Psychology* 2:5 (2014).

[106] O. Gureje, B. D. Oladeji, R. Araya, and A. A. Montgomery. A cluster randomized clinical trial of a stepped care intervention for depression in primary care (STEPCARE)-study protocol. *BMC Psychiatry* 15:148 (2015).

[107] C. Stoop, G. Nefs, A. Pommer, V. Pop, and F. Pouwer. Effectiveness of a stepped care intervention for anxiety and depression in people with diabetes, asthma or COPD in primary care: A randomized controlled trial. *Journal of Affective Disorders* 184:(2015), p. 269–276.

[108] H. Stam, J. C. van der Wouden, H. E. van der Horst, and O. R. Maarsingh. Impairment reduction in older dizzy people in primary care: study protocol for a cluster randomised controlled trial. *Trials* 16:313 (2015).

[109] J. Lock, D. Le Grange, W. S. Agras, K. K. Fitzpatrick, B. Jo, E. Accurso, S. Forsberg, K. Anderson, K. Arnow, and M. Stainer. Can adaptive treatment improve outcomes in family-based therapy for adolescents with anorexia nervosa? Feasibility and treatment effects of a multi-site treatment study. *Behaviour Research and Therapy* 73:(2015), p. 90–95.

[110] C. S. Schuurhuizen, A. M. Braamse, A. T. Beekman, H. Bomhof-Roordink, J. E. Bosmans, P. Cuijpers, A. W. Hoogendoorn, I. R. Konings, M. H. van der Linden, E. C. Neefjes, et al. Screening and treatment of psychological distress in patients with metastatic colorectal cancer: study protocol of the TES trial. *BMC Cancer* 15:302 (2015).

[111] T. Haug, T. Nordgreen, L.-G. Öst, G. Kvale, T. Tangen, G. Andersson, P. Carlbring, E. R. Heiervang, and O. E. Havik. Stepped care versus face-to–face cognitive behavior therapy for panic disorder and social anxiety disorder: predictors and moderators of outcome. *Behaviour Research and Therapy* 71:(2015), p. 76–89.

[112] A. Salloum, W. Wang, J. Robst, T. K. Murphy, M. S. Scheeringa, J. A. Cohen, and E. A. Storch. Stepped care versus standard trauma-focused cognitive behavioral therapy for young children. *Journal of Child Psychology and Psychiatry* 57:5 (2015), p. 614–622.

[113] F. Wu, E. B. Laber, I. A. Lipkovich, and E. Severus. Who will benefit from antidepressants in the acute treatment of bipolar depression? A reanalysis of the STEP-BD study by Sachs et al. 2007, using Q-learning. *International Journal of Bipolar Disorders* 3:1 (2015), p. 1–11.

[114] J. T. Painter, J. C. Fortney, A. L. Gifford, D. Rimland, T. Monson, M. C. Rodriguez-Barradas, and J. M. Pyne. Cost-Effectiveness of Collaborative Care for Depression in HIV Clinics. *Journal of Acquired Immune Deficiency Syndromes* 70:4 (2015), p. 377–385.

[115] National Institute for Helath and Care Excellence. Common mental health disorders: Identification and pathways to care. NICE guidelines [CG123]. May 2011. URL: https://www.nice.org.uk/guidance/CG123.

[116] J. Paris. Stepped Care: An Alternative to Routine Extended Treatment for Patients With Borderline Personality Disorder. *Psychiatric Services* 64:10 (2013), p. 1035–7.

[117] B. Borsari, M. Magill, N. R. Mastroleo, J. T. P. Hustad, T. O. O. O. Tevyaw, N. P. Barnett, C. W. Kahler, E. Eaton, and P. M. Monti. Mandated College Students' Response to Sequentially Administered Alcohol Interventions in a Randomized Clinical Trial Using Stepped Care. *Journal of Consulting and Clinical Psychology* 84:2 (2016), p. 103–112.

[118] N. Mitchell, C. Hewitt, J. Adamson, S. Parrott, D. Torgerson, D. Ekers, J. Holmes, H. Lester, D. McMillan, D. Richards, K. Spilsbury, C. Godfrey, and S. Gilbody. A randomised evaluation of CollAborative care and active surveillance for Screen-Positive ElderRs with sub-threshold depression (CASPER): study protocol for a randomized controlled trial. *Trials* 12:225 (2011).

[119] L. M. Collins, S. A. Murphy, and V. Strecher. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. *American Journal of Preventive Medicine* 32:5 (2007), p. S112–S118.

[120] D. Almirall, S. N. Compton, M. Gunlicks-Stoessel, N. Duan, and S. A. Murphy. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Statistics in Medicine* 31:17 (2012), p. 1887–1902.

[121] D. A. Richards, P. Bower, C. Pagel, A. Weaver, M. Utley, J. Cape, S. Pilling, K. Lovell, S. Gilbody, J. Leibowitz, et al. Delivering stepped care: an analysis of implementation in routine practice. *Implementation Science* 7:1 (2012), p. 3.

[122] S. A. Baldwin, D. J. Bauer, E. Stice, and P. Rohde. Evaluating models for partially clustered designs. *Psychological Methods* 16:2 (2011), p. 149–165.

[123] T. C. Hoffmann, P. P. Glasziou, I. Boutron, R. Milne, R. Perera, D. Moher, D. G. Altman, V. Barbour, H. Macdonald, M. Johnston, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 348:(2014), p. g1687.

[124] S. K. Sterba. Partially nested designs in psychotherapy trials: A review of modeling developments. *Psychotherapy Research* 27:4 (2017), p. 425–436.

[125] K. Sprange, G. A. Mountain, J. Brazier, S. P. Cook, C. Craig, D. Hind, S. J. Walters, G. Windle, R. Woods, A. D. Keetharuth, et al. Lifestyle Matters for maintenance of health and wellbeing in people aged 65 years and over: study protocol for a randomised controlled trial. *Trials* 14:302 (2013).

[126] K. Thomas, H. MacPherson, L. Thorpe, J. Brazier, M. Fitter, M. Campbell, M. Roman, S. Walters, and J. Nicholl. Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain. *BMJ* 333:7569 (2006), p. 623.

[127] E. Batistatou, C. Roberts, S. Roberts, et al. Sample size and power calculations for trials and quasi-experimental studies with clustering. *Stata Journal* 14:1 (2014), p. 159–75.

[128] M. Candel and G. J. P. Van Breukelen. Varying cluster sizes in trials with clusters in one treatment arm: Sample size adjustments when testing treatment effects with linear mixed models. *Statistics in Medicine* 28:18 (2009), p. 2307–2324.

[129] M. J. Fazzari, M. Y. Kim, and M. Heo. Sample size determination for three-level randomized clinical trials with randomization at the first or second level. *Journal of Biopharmaceutical Statistics* 24:3 (2014), p. 579–599.

[130] M. Moerbeek and S. Teerenstra. *Power analysis of trials with multilevel data*. CRC Press, 2016.

[131] M. Moerbeek and W. K. Wong. Sample size formulae for trials comparing group and individual treatments in a multilevel model. *Statistics in Medicine* 27:15 (2008), p. 2850–2864.

[132] E. Korendijk. Robustness and optimal design issues for cluster randomized trials. PhD thesis. Utrecht University, 2012.

[133] S. Lohr, P. Z. Schochet, and E. Sanders. Partially Nested Randomized Controlled Trials in Education Research: A Guide to Design and Analysis. NCER 2014-2000. National Center for Education Research (NCER). 2014.

[134] E. A. Sanders. Multilevel analysis methods for partially nested cluster randomized trials. PhD thesis. University of Washington, USA, 2011.

[135] J. D. Schweig and J. F. Pane. Intention-to-treat analysis in partially nested randomized controlled trials with real-world complexity. *International Journal of Research & Method in Education* 39:3 (2016), p. 268–286.

[136] D. R. Hoover. Clinical trials of behavioural interventions with heterogeneous teaching subgroup effects. *Statistics in Medicine* 21:10 (2002), p. 1351–1364.

[137] M. Heo, A. H. Litwin, O. Blackstock, N. Kim, and J. H. Arnsten. Sample size determinations for group-based randomized clinical trials with different levels of data hierarchy between experimental and control arms. *Statistical Methods in Medical Research*:(2014).

[138] M. Bland. Grouping in individually rrandomised trials. Randomised Controlled Trials in the Social Sciences Conference 2009. 2009.

[139] F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2:6 (1946), p. 110–114.

[140] M. G. Kenward and J. H. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*:(1997), p. 983–997.

[141] B. C. Kahan, G. Forbes, Y. Ali, V. Jairath, S. Bremner, M. O. Harhay, R. Hooper, N. Wright, S. M. Eldridge, and C. Leyrat. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials* 17:438 (2016).

[142] StataCorp. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP. 2015.

[143] L. StataCorp. Stata multilevel mixed-effects reference manual. *College Station, TX: StataCorp LP*:(2013).

[144] A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine* 25:24 (2006), p. 4279–4292.

[145] M. K. Smith and A. Marshall. Importance of protocols for simulation studies in clinical drug development. *Statistical Methods in Medical Research* 20:6 (2011), p. 612–622.

[146] H. Wickham. Elegant graphics for data analysis (ggplot2). 2009.

[147] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: `https://www.R-project.org`.

[148] A. Donner and N. Klar. Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology* 49:4 (1996), p. 435–439.

[149] C. J. Maas and J. J. Hox. Robustness issues in multilevel regression analysis. *Statistica Neerlandica* 58:2 (2004), p. 127–137.

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

[151] T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *arXiv preprint arXiv:1712.03198*:(2017). URL: `https://arxiv.org/abs/1712.03198`.

[152] W. Rogers et al. Quantile regression standard errors. *Stata Technical Bulletin* 2:9 (1993).

[153] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap. Monographs on Statistics and Applied Probability 57*. Chapman & Hall, 1993.

[154] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Vol. 1. Cambridge university press, 1997.

[155] T. N. Flynn and T. J. Peters. Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results. *BMC Health Services Research* 4:33 (2004).

[156] J. M. Lachin. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* 2:2 (1981), p. 93–113.

[157] S. Senn. *Cross-over Trials in Medical Research*. Chichester: John Wiley, 1993.

[158] D. Machin, M. Campbell, P. Fayers, and A. Pinol. *Sample size tables for clinical studies*. 3rd ed. John Wiley & Sons, 2008.

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

[160] D. A. Harrison, A. R. Brady, et al. Sample size and power calculations using the noncentral t-distribution. *Stata Journal* 4:(2004), p. 142–153.

[161] E. L. Turner, F. Li, J. A. Gallis, M. Prague, D. M. Murray, et al. Review of Recent Methodological Developments in Group-Randomized Trials: Part 1—Design. *American Journal of Public Health* 107:6 (2017), p. 907–915.

[162] M. J. Candel and G. J. Van Breukelen. Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. *Statistical Methods in Medical Research* 24:5 (2015), p. 557–573.

[163] K. Bosanquet, J. Adamson, K. Atherton, D. Bailey, C. Baxter, J. Beresford-Dent, et al. CollAborative care for Screen-Positive EldeRs with major depression (CASPER plus): a multicentred randomised controlled trial of clinical effectiveness and cost-effectiveness. 21:67 (2017).

[164] M. Crawford, R. Sanatinia, B. Barrett, S. Byford, M. Dean, J. Green, et al. The clinical and cost effectiveness of brief intervention for excessive alcohol consumption among people attending sexual health clinics: a randomised controlled trial (SHEAR). *Health Technology Assessment* 18:30 (2014).

[165] J. Elashoff. nQuery Advisor (Version 7.0 User's Guide. 2007. *Los Angeles, CA: Statistical Solutions Ltd*:(2007).

[166] B. K. Moser, G. R. Stevens, and C. L. Watts. The two-sample t test versus Satterthwaite's approximate F test. *Communications in Statistics-Theory and Methods* 18:11 (1989), p. 3963–3975.

[167] K. Jozwiak, M. Moerbeek, and S. Teerenstra. SPA-ML (Statistic Power Analysis for Multilevel Design). 2014. URL: `http://tinyurl.com/SPAML`.

[168] cluspower Stata program. URL: `http://personalpages.manchester.ac.uk/staff/Chris.Roberts/stata/` (visited on 05/18/2017).

[169] *Guideline on adjustment for baseline covariates*. Tech. rep. Committee for Medicinal Products for Human Use, 2015.

[170] Health Technology Assessment Programme. URL: `https://www.nihr.ac.uk/funding-and-support/funding-for-research-studies/funding-programmes/health-technology-assessment/` (visited on 09/20/2018).

[171] B. Stuart, T. Becque, M. Moore, M. Mullee, and P. Little. Clustering at the general practice level in individually randomised studies in primary care - a review of 10 years of primary care trials:(In Press).

[172] M. Campbell, J. Grimshaw, N. Steen, and C. P. P. in Europe Group (EU BIOMED II Concerted Action). Sample size calculations for cluster randomised trials. *Journal of Health Services Research & Policy* 5:1 (2000), p. 12–16.

[173] M. Taljaard, A. Donner, J. Villar, D. Wojdyla, A. Velazco, V. Bataglia, A. Faundes, A. Langer, A. Narváez, E. Valladares, et al. Intracluster correlation coefficients from the 2005 WHO Global Survey on Maternal and Perinatal Health: implications for implementation research. *Paediatric and Perinatal Epidemiology* 22:2 (2008), p. 117–125.

[174] C. Pagel, A. Prost, S. Lewycka, S. Das, T. Colbourn, R. Mahapatra, K. Azad, A. Costello, and D. Osrin. Intracluster correlation coefficients and coefficients of variation for perinatal outcomes from five cluster-randomised controlled trials in low and middle-income countries: results and methodological implications. *Trials* 12:151 (2011).

[175] J. R. Glassman, S. C. Potter, E. R. Baumler, and K. K. Coyle. Estimates of intraclass correlation coefficients from longitudinal group-randomized trials of adolescent HIV/STI/pregnancy prevention programs. *Health Education & Behavior* 42:4 (2015), p. 545–553.

[176] H. Oltean and J. J. Gagnier. Use of clustering analysis in randomized controlled trials in orthopaedic surgery. *BMC Medical Research Methodology* 15:17 (2015).

[177] B. C. Kahan and T. P. Morris. Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effects? *Statistics in Medicine* 32:7 (2013), p. 1136–1149.

[178] R. Möhler, S. Köpke, and G. Meyer. Criteria for Reporting the Development and Evaluation of Complex Interventions in healthcare: revised guideline (CReDECI 2). *Trials* 16:204 (2015).

[179] S. J. Walters, I. B. dos Anjos Henriques-Cadby, O. Bortolami, L. Flight, D. Hind, R. M. Jacques, C. Knox, B. Nadin, J. Rothwell, M. Surtees, 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* 7:3 (2017), p. e015276.

[180] S. M. Eldridge, C. L. Chan, M. J. Campbell, C. M. Bond, S. Hopewell, L. Thabane, and G. A. Lancaster. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *Pilot and Feasibility Studies* 2:64 (2016).

[181] H. McRobbie, P. Hajek, S. Peerbux, B. Kahan, S. Eldridge, D. Trépel, et al. Tackling obesity in areas of high social deprivation: clinical effectiveness and cost-effectiveness of a task-based weight management group programme a randomised controlled trial and economic evaluation. *Health Technology Assessment* 20:79 (2016).

[182] D. Keene, D. Mistry, J. Nam, L. Tutton, R. Handley, and L. Morgan. The Ankle Injury Management (AIM) trial: a pragmatic, multicentre, equivalence randomised controlled trial and economic evaluation comparing close contact casting with open surgical reduc-

tion and internal fixation in the treatment of unstable ankle fractures in patients aged over 60 years. *Health Technology Assessment* 20:75 (2016).

[183] S. Priebe, M. Savill, T. Wykes, R. Bentall, C. Lauber, U. Reininghaus, et al. Clinical effectiveness and cost-effectiveness of body psychotherapy in the treatment of negative symptoms of schizophrenia: a multicentre randomised controlled trial. *Health Technology Assessment* 20:11 (2016).

[184] E. Goyder, D. Hind, D. Breckon, D. Dimairo, D. Minton, E. Everson-Hock, et al. A randomised controlled trial and cost-effectiveness evaluation of 'booster' interventions to sustain increases in physical activity in middle-aged adults in deprived urban neighbourhoods. *Health Technology Assessment* 18:13 (2014).

[185] M. Thomas, A. Bruton, P. Little, S. Holgate, A. Lee, L. Yardley, et al. A randomised controlled study of the effectiveness of breathing retraining exercises taught by a physiotherapist either by instructional DVD or in face-to-face sessions in the management of asthma in adults. *Health Technology Assessment* 21:53 (2017).

[186] H. Lewis, J. Adamson, K. Atherton, D. Bailey, J. Birtwistle, and K. Bosanquet. CollAborative care and active surveillance for Screen-Positive EldeRs with subthreshold depression (CASPER): a multicentred randomised controlled trial of clinical effectiveness and cost-effectiveness. *Health Technology Assessment* 21:8 (2017).

[187] P. Tyrer, P. Salkovskis, H. Tyrer, D. Wang, M. J. Crawford, S. Dupont, et al. Cognitive-behaviour therapy for health anxiety in medical patients (CHAMP): a randomised controlled trial with outcomes to 5 years. *Health Technology Assessment* 21:50 (2017).

[188] J. Brittenden, S. Cotton, A. Elders, E. Tassie, G. Scotland, and C. Ramsay. Clinical and cost-effectiveness of foam sclerotherapy, endovenous laser ablation and surgery for varicose veins: results from the CLASS trial. *Health Technology Assessment* 19:27 (2015).

[189] D. Richards, S. Rhodes, D. Ekers, D. McMillan, R. Taylor, S. Byford, et al. Cost and Outcome of BehaviouRal Activation (COBRA): a randomised controlled trial of behavioural activation versus cognitive behavioural therapy for depression. *Health Technology Assessment* 21:46 (2017).

[190] A. Watson, J. Cook, J. Hudson, M. Kilonzo, J. Wood, H. Bruhn, et al. A pragmatic multicentre randomised controlled trial comparing stapled haemorrhoidopexy with traditional excisional surgery for haemorrhoidal disease: the eTHoS study. *Health Technology Assessment* 21:70 (2017).

[191] P. Logan, S. Armstrong, A. Avery, D. Barer, G. Barton, J. Darby, et al. Rehabilitation aimed at improving outdoor mobility for people after stroke: a multi-centre randomised controlled study (the Getting out of the House Study). *Health Technology Assessment* 18:29 (2014).

[192] I. Goodyer, S. Reynolds, B. Barrett, S. Byford, B. Dubicka, J. Hill, et al. Cognitive behavioural therapy and short-term psychoanalytic psychotherapy versus brief psychosocial intervention in adolescents with unipolar major depression (IMPACT): a multicentre, pragmatic, observer-blind, randomised controlled trial. *Health Technology Assessment* 21:12 (2017).

[193] D. Murray, G. MacLennan, S. Breeman, H. Dakin, L. Johnston, M. Campbell, et al. A randomised controlled trial of the clinical effectiveness and cost-effectiveness of different knee prostheses: the Knee Arthroplasty Trial (KAT). *Health Technology Assessment* 18:19 (2014).

[194] C. Clarke, S. Patel, N. Ives, C. Rick, R. Woolley, K. Wheatley, et al. Clinical effectiveness and cost-effectiveness of physiotherapy and occupational therapy versus no therapy in mild to moderate Parkinson's disease: a large pragmatic randomised controlled trial (PD REHAB). *Health Technology Assessment* 20:63 (2016).

[195] M. McMurran, M. J. Crawford, J. Reilly, J. Delport, P. McCrone, D. Whitham, et al. Psychoeducation with problem-solving (PEPS) therapy for adults with personality disorder: a pragmatic randomised controlled trial to determine the clinical effectiveness and cost-effectiveness of a manualised intervention to improve social functioning. *Health Technology Assessment* 20:52 (2016).

[196] C. Salisbury, N. Foster, C. Hopper, A. Bishop, S. Hollinghurst, J. Coast, et al. A pragmatic randomised controlled trial of the effectiveness and cost-effectiveness of 'PhysioDirect' telephone assessment and advice services for physiotherapy. *Health Technology Assessment* 17:2 (2013).

[197] P. Little, B. Stuart, R. Hobbs, J. Kelly, E. Smith, K. Bradbury, et al. Randomised controlled trial and economic analysis of an internet-based weight management programme: POWeR+ (Positive Online Weight Reduction). *Health Technology Assessment* 21:4 (2014).

[198] H. Handoll, S. Brealey, A. Rangan, A. Keding, B. Corbacho, L. Jefferson, et al. The ProFHER (PROximal Fracture of the Humerus: Evaluation by Randomisation) trial - a

pragmatic multicentre randomised controlled trial evaluating the clinical effectiveness and cost-effectiveness of surgical compared with non-surgical treatment for proximal fracture of the humerus in adults. *Health Technology Assessment* 19:24 (2015).

[199] G. Livingston, J. Barber, P. Rapaport, M. Knapp, M. Griffin, and R. Romeo. START (STrAtegies for RelaTives) study: a pragmatic randomised controlled trial to determine the clinical effectiveness and cost-effectiveness of a manual-based coping strategy programme in promoting the mental health of carers of people with dementia. *Health Technology Assessmen* 18:61 (2014).

[200] S. Parry, C. Bamford, V. Deary, T. Finch, J. Gray, C. MacDonald, et al. Cognitive behavioural therapy-based intervention to reduce fear of falling in older people: therapy development and randomised controlled trial the Strategies for Increasing Independence, Confidence and Energy (STRIDE) study. *Health Technology Assessment* 20:56 (2016).

[201] M. Crawford, C. Gold, H. Odell-Miller, L. Thana, S. Faber, J. Assmus, et al. International multicentre randomised controlled trial of improvisational music therapy for children with autism spectrum disorder: TIME-A study. *Health Technology Assessment* 21:59 (2017).

[202] M. Costa, J. Achten, C. Plant, N. Parsons, A. Rangan, S. Tubeuf, et al. UK DRAFFT - A Randomised Controlled Trial of Percutaneous Fixation with Kirschner Wires versus Volar Locking-Plate Fixation in the Treatment of Adult Patients with a Dorsally Displaced Fracture of the Distal Radius. *Health Technology Assessment* 19:17 (2015).

[203] A. Carr, C. Cooper, M. Campbell, J. Rees, J. Moser, D. Beard, et al. Clinical effectiveness and cost-effectiveness of open and arthroscopic rotator cuff repair [the UK Rotator Cuff Surgery (UKUFF) randomised trial]. *Health Technology Assessment* 19:80 (2015).

[204] Z. Samaan, L. Mbuagbaw, D. Kosa, V. B. Debono, R. Dillenburg, S. Zhang, V. Fruci, B. Dennis, M. Bawor, and L. Thabane. A systematic scoping review of adherence to reporting guidelines in health care literature. *Journal of Multidisciplinary Healthcare* 6:(2013), p. 169.

[205] M. Nagendran, D. Harding, W. Teo, C. Camm, M. Maruthappu, P. McCulloch, and S. Hopewell. Poor adherence of randomised trials in surgery to CONSORT guidelines for non-pharmacological treatments (NPT): a cross-sectional study. *BMJ Open* 3:12 (2013), p. e003898.

[206] N. Ivers, M. Taljaard, S. Dixon, C. Bennett, A. McRae, J. Taleban, Z. Skea, J. Brehaut, R. Boruch, M. Eccles, 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* 343:(2011), p. d5886.

[207] D. Saxon and M. Barkham. Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psychology* 80:4 (2012), p. 535.

[208] A. C. Plint, D. Moher, A. Morrison, K. Schulz, D. G. Altman, C. Hill, and I. Gaboury. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Medical journal of Australia* 185:5 (2006), p. 263.

[209] M. K. Campbell, J. M. Grimshaw, and D. R. Elbourne. Intracluster correlation coefficients in cluster randomized trials: empirical insights into how should they be reported. *BMC Medical Research Methodology* 4:9 (2004).

[210] P. Montgomery, S. Grant, E. Mayo-Wilson, G. Macdonald, S. Michie, S. Hopewell, and D. Moher. Reporting randomised trials of social and psychological interventions: the CONSORT-SPI 2018 Extension. *Trials* 19:407 (2018).

[211] C. Roberts. Partial Nesting: Some Implications for Analysis and Design. *Unpublished presentation slides*. University of Manchester, 2013.

[212] J. Candlish, M. D. Teare, J. Cohen, and T. Bywater. Statistical design and analysis in trials of proportionate interventions: a systematic review. *Trials*:(Under review).

[213] J. Candlish, M. D. Teare, and J. Cohen. Analysis methods for individually randomised trials with partial clustering of outcomes. *Clinical Trials* 15:(Suppl 2 2018), p. S94–S95.

[214] J. Candlish, M. D. Teare, J. Cohen, M. Dimairo, L. Flight, L. Mandefield, and S. Walters. Methods to analyse partially nested randomised controlled trials. *Trials* 18:Suppl 1 (2017), p. P441.

[215] J. Candlish. Analysis methods for partially nested randomised controlled trials. *38th Annual Conference of the International Society for Clinical Biostatistics*. Vigo, Spain, 2017.

[216] J. Candlish. Design and analysis of proportionate intervention trials: a systematic review. *37th Annual Conference of the International Society for Clinical Biostatistics*. Birmingham, UK, 2016.

[217] C. Roberts, E. Batistatou, and S. A. Roberts. Design and analysis of trials with a partially nested design and a binary outcome measure. *Statistics in Medicine* 35:10 (2016), p. 1616–1636.

[218] A. Schafer. New methods for the analysis of change. Ed. by L. Collins and A. Sayer. Washington, DC: American Psychological Association, 2001. Chap. Multiple imputation with PAN, pp. 355–377.

Appendices

Appendix A

Partially nested randomised trials (chapter 4)

A.1 Supplementary results, figures and tables

Figure A.1: Power when $\theta = 0.2$, by ρ, γ, c and m

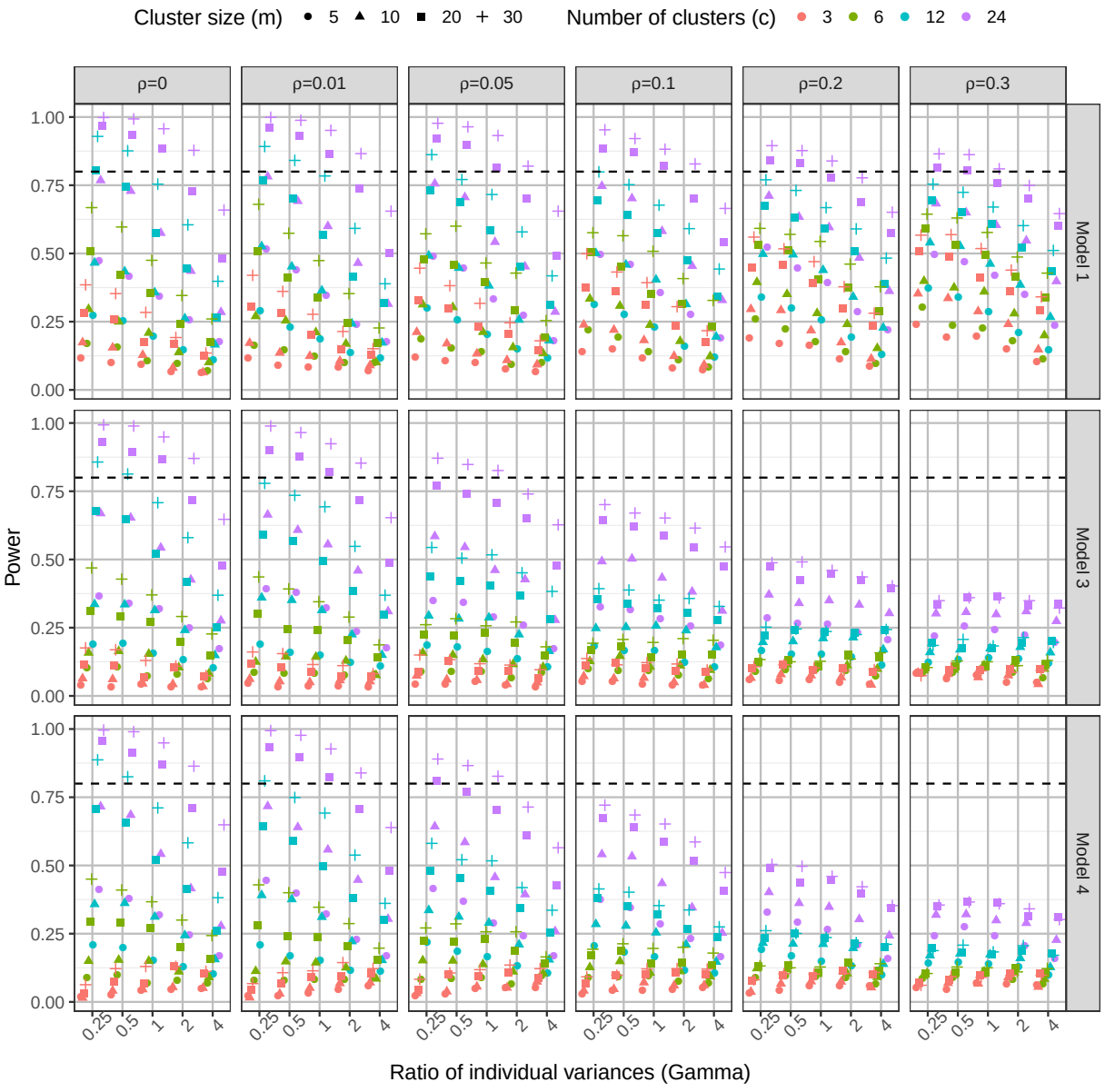


Figure A.2: Bias of between- and within-cluster variance estimates from the heteroscedastic partially nested model (model 4) by ρ, γ, c and m

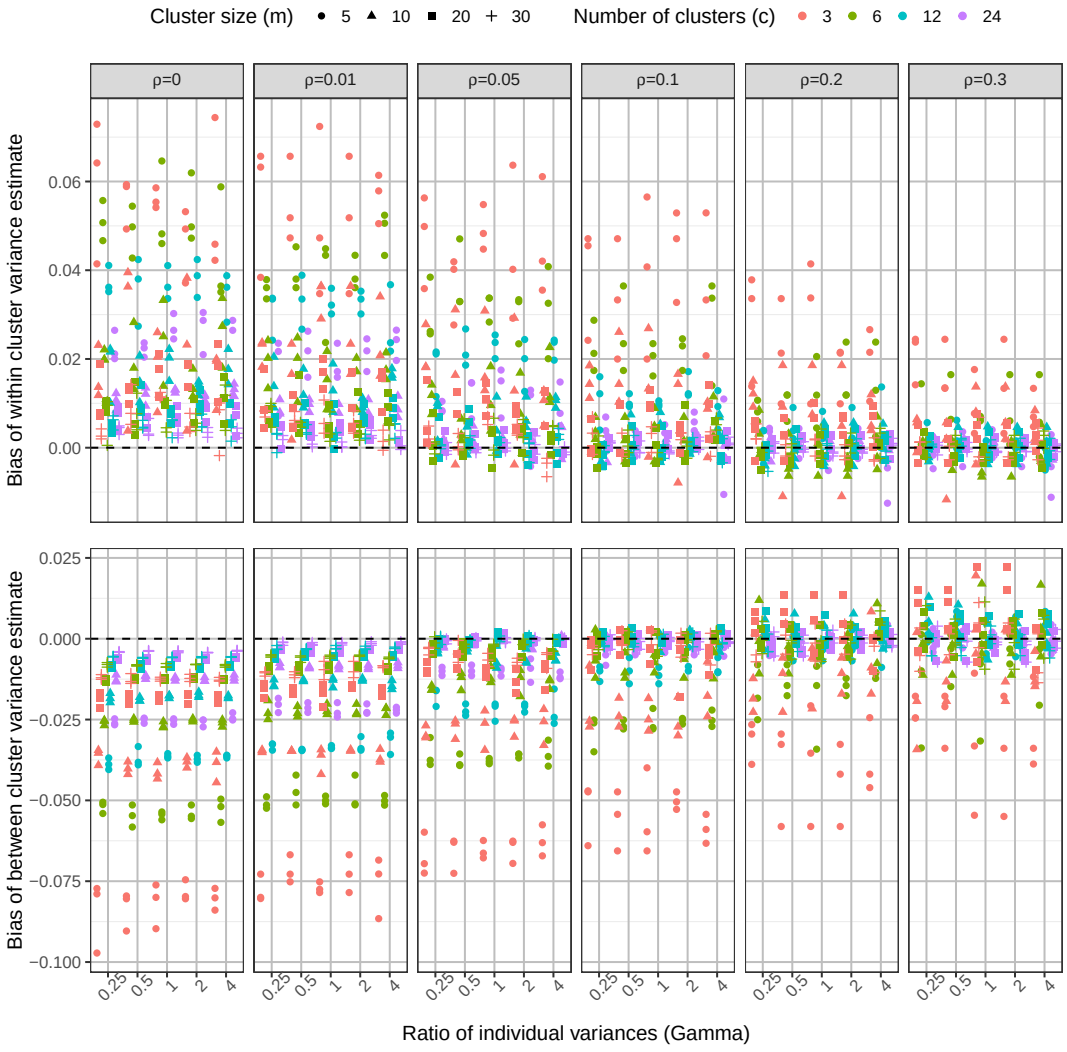


Table A.1: Type I error rate (mean and SD) under null hypothesis by Model, γ and ρ

		Model											
γ	ρ	**1**		**2.1**		**2.2**		**2.3**		**3**		**4**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
0.25	0	0.05	0.01	0.03	0.01	0.00	0.01	0.02	0.01	0.02	0.01	0.02	0.01
	0.01	0.07	0.01	0.04	0.01	0.00	0.01	0.03	0.01	0.03	0.01	0.03	0.02
	0.05	0.12	0.05	0.07	0.03	0.00	0.01	0.04	0.01	0.03	0.01	0.03	0.02
	0.1	0.18	0.08	0.10	0.03	0.00	0.01	0.05	0.01	0.04	0.01	0.04	0.01
	0.2	0.28	0.11	0.15	0.05	0.00	0.01	0.06	0.01	0.05	0.01	0.05	0.01
	0.3	0.35	0.13	0.19	0.06	0.00	0.01	0.06	0.01	0.05	0.01	0.05	0.01
0.5	0	0.05	0.01	0.03	0.01	0.00	0.01	0.03	0.01	0.03	0.01	0.03	0.01
	0.01	0.06	0.01	0.04	0.01	0.00	0.01	0.03	0.01	0.03	0.01	0.03	0.01
	0.05	0.11	0.03	0.06	0.02	0.00	0.01	0.04	0.01	0.04	0.01	0.04	0.01
	0.1	0.17	0.07	0.08	0.03	0.00	0.01	0.05	0.01	0.04	0.01	0.04	0.01
	0.2	0.26	0.11	0.12	0.04	0.00	0.01	0.05	0.01	0.05	0.01	0.05	0.01
	0.3	0.33	0.13	0.16	0.05	0.00	0.02	0.06	0.02	0.05	0.01	0.05	0.01
1	0	0.05	0.01	0.03	0.01	0.00	0.01	0.02	0.01	0.03	0.01	0.03	0.01
	0.01	0.06	0.01	0.03	0.01	0.00	0.01	0.03	0.01	0.03	0.01	0.03	0.01
	0.05	0.09	0.03	0.04	0.01	0.00	0.01	0.03	0.01	0.04	0.01	0.04	0.01
	0.1	0.14	0.05	0.05	0.01	0.00	0.01	0.04	0.01	0.05	0.01	0.05	0.01
	0.2	0.22	0.09	0.07	0.02	0.00	0.01	0.05	0.01	0.05	0.01	0.05	0.01
	0.3	0.29	0.12	0.09	0.02	0.00	0.01	0.05	0.01	0.06	0.01	0.05	0.01
2	0	0.05	0.01	0.01	0.01	0.00	0.01	0.03	0.01	0.04	0.01	0.04	0.01
	0.01	0.06	0.01	0.01	0.01	0.00	0.01	0.03	0.01	0.04	0.01	0.04	0.01
	0.05	0.08	0.02	0.01	0.01	0.00	0.02	0.03	0.02	0.05	0.02	0.04	0.01
	0.1	0.12	0.04	0.01	0.01	0.00	0.02	0.04	0.02	0.06	0.01	0.05	0.01
	0.2	0.18	0.07	0.01	0.01	0.00	0.01	0.04	0.01	0.05	0.01	0.05	0.01
	0.3	0.23	0.10	0.02	0.01	0.00	0.01	0.05	0.01	0.06	0.01	0.06	0.01
4	0	0.05	0.01	0.00	0.00	0.00	0.01	0.02	0.01	0.04	0.01	0.04	0.01
	0.01	0.06	0.01	0.00	0.00	0.00	0.02	0.03	0.02	0.05	0.01	0.05	0.01
	0.05	0.07	0.01	0.00	0.00	0.00	0.02	0.03	0.02	0.05	0.02	0.05	0.01
	0.1	0.09	0.03	0.00	0.00	0.00	0.02	0.03	0.02	0.06	0.02	0.05	0.01
	0.2	0.14	0.06	0.00	0.00	0.00	0.01	0.04	0.01	0.06	0.01	0.05	0.01
	0.3	0.18	0.08	0.00	0.00	0.00	0.01	0.04	0.01	0.06	0.01	0.06	0.01

Table A.2: Power (mean and SD) under alternative hypothesis by Model, γ and ρ

		Model											
γ	ρ	**1**		**2.1**		**2.2**		**2.3**		**3**		**4**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
0.25	0	0.71	0.31	0.67	0.34	0.06	0.18	0.62	0.37	0.62	0.36	0.60	0.38
	0.01	0.71	0.31	0.66	0.34	0.06	0.20	0.60	0.36	0.60	0.36	0.58	0.38
	0.05	0.70	0.30	0.64	0.33	0.03	0.08	0.56	0.35	0.54	0.35	0.54	0.36
	0.1	0.70	0.28	0.62	0.31	0.01	0.02	0.52	0.34	0.49	0.33	0.48	0.35
	0.2	0.70	0.25	0.60	0.29	0.00	0.01	0.45	0.32	0.42	0.32	0.42	0.33
	0.3	0.71	0.23	0.59	0.28	0.00	0.01	0.40	0.30	0.36	0.3	0.36	0.31
0.5	0	0.68	0.32	0.63	0.35	0.03	0.11	0.58	0.37	0.6	0.36	0.59	0.37
	0.01	0.67	0.32	0.62	0.35	0.05	0.16	0.57	0.37	0.58	0.36	0.58	0.37
	0.05	0.67	0.30	0.60	0.33	0.03	0.08	0.53	0.35	0.53	0.34	0.53	0.35
	0.1	0.67	0.29	0.58	0.32	0.01	0.02	0.50	0.34	0.48	0.33	0.48	0.34
	0.2	0.68	0.26	0.56	0.30	0.00	0.00	0.43	0.32	0.41	0.32	0.41	0.32
	0.3	0.68	0.24	0.55	0.29	0.00	0.00	0.39	0.30	0.36	0.30	0.36	0.30
1	0	0.62	0.33	0.55	0.36	0.01	0.02	0.52	0.37	0.56	0.36	0.56	0.36
	0.01	0.62	0.33	0.55	0.36	0.02	0.07	0.51	0.37	0.55	0.36	0.55	0.36
	0.05	0.62	0.32	0.51	0.34	0.02	0.08	0.49	0.36	0.51	0.34	0.51	0.34
	0.1	0.62	0.3	0.48	0.33	0.01	0.02	0.45	0.34	0.46	0.33	0.46	0.33
	0.2	0.64	0.28	0.46	0.32	0.00	0.00	0.41	0.32	0.40	0.32	0.40	0.32
	0.3	0.64	0.25	0.46	0.30	0.00	0.00	0.37	0.3	0.36	0.30	0.36	0.30
2	0	0.54	0.34	0.22	0.29	0.00	0.00	0.44	0.36	0.50	0.35	0.50	0.35
	0.01	0.54	0.33	0.22	0.29	0.00	0.00	0.44	0.36	0.49	0.35	0.50	0.34
	0.05	0.54	0.32	0.22	0.29	0.01	0.04	0.42	0.35	0.47	0.34	0.47	0.33
	0.1	0.55	0.31	0.21	0.28	0.01	0.02	0.40	0.34	0.44	0.33	0.43	0.32
	0.2	0.57	0.29	0.22	0.28	0.00	0.00	0.37	0.32	0.39	0.31	0.38	0.30
	0.3	0.59	0.27	0.25	0.27	0.00	0.00	0.34	0.29	0.35	0.29	0.35	0.28
4	0	0.43	0.32	0.02	0.04	0.00	0.00	0.34	0.33	0.40	0.33	0.42	0.32
	0.01	0.43	0.31	0.02	0.04	0.00	0.00	0.34	0.33	0.40	0.33	0.41	0.32
	0.05	0.44	0.31	0.02	0.04	0.00	0.00	0.33	0.32	0.40	0.32	0.39	0.31
	0.1	0.45	0.31	0.02	0.05	0.00	0.01	0.33	0.32	0.39	0.31	0.37	0.30
	0.2	0.48	0.29	0.04	0.07	0.00	0.00	0.31	0.30	0.36	0.29	0.34	0.28
	0.3	0.50	0.28	0.05	0.10	0.00	0.00	0.30	0.28	0.33	0.28	0.32	0.27

A.2 Stata simulation code for partially nested trials analysis

The following Stata code was used for the simulation study in chapter 4. Code was going to be written in R initially, until it became apparent that the heteroscedastic model with Satterthwaite degrees of freedom correction was not available in R packages at the time of the study. Models 2, 3 and 4 were run for imposed clustering of different types, however, only cluster type 1 (singleton clusters in control group) are shown in simulation code below.

```
/* Simulation Parameters                                                    */
       *Simulation for Yij = alpha +theta*tij + ujtij+(1-tij)rij + tijeij
       *normally distributed continuous outcome with cluster (j) and
       *residual (i) variability

/* Function parameters                                    */
       *theta*: difference in outcome between treatment and control
       *gam*: ratio of individual varaince between control and treatment arm
       *rho*: ICC in clustered treatment arm
       *clusterSize*: Size of each cluster
       *nClusters*: Number of clusters in treatment arm
       *total trial sample size = 2*nClusters*clusterSize
       *nIterations*: number of iterations for simulation

/* Variable Specification                                          */
       *y*: is the dependent variable;
       *cluster* is one of the following clustering options
       *cluster1* is the cluster indicator and treats the control group as clusters of
            size 1;
       *cluster2* is the cluster indicator and treats the control group as one cluster
           ;
       *cluster3* is the cluster indicator and treats the control group as random
           clusters;
       *tx* is a binary indicator for treatment (intervention group = 1, control group
            = 0).

cap log close
clear
set more off
capture log

/* 1_simulation.do takes args simul_No nIterations theta rho gam clusterSize nClusters
    .
Use a master file to run 1_simulation.do file with various scenarios given by args
for example, do do\1_simulation.do 1 1000 0.2 0 1 5 12                          */

/* args assigns the first command line argument to the local macro simula_No, second
    argument to nIterations and so on.*/
args simul_No nIterations theta rho gam clusterSize nClusters

/* Set working directory */
cd "working directory"

local initial_seed = c(seed)

/* Log file */
log using log\simulation‘simul_No’.smcl, smcl replace
```

```
/*Set up file for output*/
* Create a temporary file that will store the output from the simulations
tempname memhold
tempfile results
qui postfile ‘memhold’ Theta gam rho clusterSize nClusters Model Treat SE P_value
    CI_ll CI_ul Cover Ind_cont_var Ind_clust_var Btw_clust_var ICC Converge using ‘
    results’, replace

forvalues sim = 1/‘nIterations’{
* monitor iterative process of simulations by displaying a dot after every 100
    simulations
 if int(‘sim’/100) ==‘sim’/100{
              di as text "." _cont
              }

* quietly performs command but suppress terminal output
quietly{

* Create an empty dataset where the number of observations are number of clusters x 2
local n2 = ‘nClusters’*2

set obs ‘n2’
gen cluster = _n
gen clustZ = rnormal(0,1)

* Expand so have ’clusterSize’ observations per cluster
expand ‘clusterSize’
bysort cluster: generate clustID = _n

* Generate treatment indicator
gen tx = 0
replace tx = 1 if _n< = _N/2

* Generate clustering in control arm
bysort tx: gene txID=_n
*cluster1 = singletone clusters in control arm
gen cluster1 = .
replace cluster1 = (txID + ‘nClusters’) if (tx == 0)
replace cluster1 = cluster if (tx == 1)

* cluster2 = one large cluster in control arm
gen cluster2 = .
replace cluster2 = 0 if (tx == 0)
replace cluster2 = cluster if (tx == 1)

* cluster3 = pseudo clusters in control arm
gen cluster3 = cluster

* Generate random variables
gen indZ = rnormal(0,1)

* Generate individual and cluster residuals
gen ClustRes_j = clustZ*sqrt(‘rho’)
gen ClustInd_ij = indZ*sqrt(1-‘rho’)
gen ContInd_ij = indZ*sqrt(‘gam’)*sqrt(1-‘rho’)

drop clustZ indZ txID

* Generate outcome
gen y = ‘theta’*tx + tx*ClustRes_j + tx*ClustInd_ij + (1-tx)*ContInd_ij
```

```
/* Fit models */
        * Model 1: ignoring clustering
        capture {
                regress y tx
                *define model number as first column
                if _rc == 0 local converge = 1
                        local m 1
                        * retrieve the contents of the outputs (estimates)
                        mat define M1 = r(table)
                        local b M1[1,1] //estimate of the treatment effect
                        local se M1[2,1] //standard error of the treatment effect
                        local p M1[4,1] //p value
                        local ll M1[5,1] //ll 95%CI
                        local ul M1[6,1] //ul 95%CI
                        local cv inrange(‘theta’,‘ll’,‘ul’) //coverage indicator
                        local ind_var = e(rmse)
                        post ‘memhold’ (‘theta’) (‘gam’) (‘rho’) (‘clusterSize’) (‘
                            nClusters’) (‘m’) (‘b’) (‘se’) (‘p’) (‘ll’) (‘ul’) (‘cv’) (‘
                            ind_var’) (.) (.) (.) (‘converge’)
                }

                if _rc != 0 {
                        local converge = 0
                        local m 1
                        post ‘memhold’ (‘theta’) (‘gam’) (‘rho’) (‘clusterSize’) (‘
                            nClusters’) (‘m’) (.) (.) (.) (.) (.) (.) (.) (.) (.) (.) (‘
                            converge’)
                        }

        * Model 2: fully clustered random effects
        * Cluster 1
        capture{
                mixed y tx || cluster1: , reml nolog dfmethod(sat) emiterate(10)
                if _rc == 0 local converge = 1
                        local m 2.1 //define model number
                        * retrieve the contents of the outputs (estimates)
                        mat define M21 = r(table)
                        mat list M21
                        local b M21[1,1] //estimate of the treatment effect
                        local se M21[2,1] //standard error of the treatment effect
                        local p M21[4,1] //p value
                        local ll M21[5,1] //ll 95%CI
                        local ul M21[6,1] //ul 95%CI
                        local cv inrange(‘theta’,‘ll’,‘ul’) //coverage indicator
                        local ind_var = exp(M21[1,4])^2 //individual variance
                        local btw_clust_var = exp(M21[1,3])^2 //between cluster variance
                            (clustered arm)
                        estat icc
                        local estrho r(icc2) //ICC
                        *post the results of each model output from each simulation
                        post ‘memhold’ (‘theta’) (‘gam’) (‘rho’) (‘clusterSize’) (‘
                            nClusters’) (‘m’) (‘b’) (‘se’) (‘p’) (‘ll’) (‘ul’) (‘cv’) (‘
                            ind_var’) (‘ind_var’) (‘btw_clust_var’) (‘estrho’) (‘converge
                            ’)
                }

                if _rc != 0 {
                        local converge = 0
                        local m 2.1
```

```
                post 'memhold' ('theta') ('gam') ('rho') ('clusterSize') ('
                    nClusters') ('m') (.) (.) (.) (.) (.) (.) (.) (.) (.) (.) ('
                    converge')
                }

* Model 3: partial clustering homoscedastic
* Cluster 1
capture {
        mixed y tx || cluster1: tx , nocons reml nolog dfmethod(sat) emiterate
            (20)
        if _rc == 0 local converge = 1
                *define model number
                local m 3.1
                * retrieve the contents of the outputs (estimates)
                mat define M31 = r(table)
                local b M31[1,1] //estimate of the treatment effect
                local se M31[2,1] //standard error of the treatment effect
                local p M31[4,1] //p value
                local ll M31[5,1] //ll 95%CI
                local ul M31[6,1] //ul 95%CI
                local cv inrange('theta','ll','ul') //coverage indicator
                local btw_clust_var = exp(M31[1,3])^2 //between cluster variance
                    (clustered arm)
                local ind_var = exp(M31[1,4])^2 //individual variance (clustered
                    arm)
                local estrho 'btw_clust_var'/('btw_clust_var'+'ind_var') // ICC
                *post the results of each model output from each simulation
                post 'memhold' ('theta') ('gam') ('rho') ('clusterSize') ('
                    nClusters') ('m') ('b') ('se') ('p') ('ll') ('ul') ('cv') ('
                    ind_var') ('ind_var') ('btw_clust_var') ('estrho') ('converge
                    ')
        }
        if _rc != 0 {
                local converge = 0
                local m 3.1
                post 'memhold' ('theta') ('gam') ('rho') ('clusterSize') ('
                    nClusters') ('m') (.) (.) (.) (.) (.) (.) (.) (.) (.) (.) ('
                    converge')
                }

* Model 4: partial clustering heteroskedastic
* Cluster 1
capture {
        mixed y tx || cluster1:tx, nocons reml nolog residuals(independent, by(
            tx)) dfmethod(sat) emiterate(10)
        if _rc == 0 local converge = 1
                *define model number
                local m 4.1
                * retrieve the contents of the outputs (estimates)
                mat define M41 = r(table)
                local b M41[1,1] //estimate of the treatment effect
                local se M41[2,1] //standard error of the treatment effect
                local p M41[4,1] //p value
                local ll M41[5,1] //ll 95%CI
                local ul M41[6,1] //ul 95%CI
                local cv inrange('theta','ll','ul') //coverage indicator
                local btw_clust_var = exp(M41[1,3])^2 //between cluster variance
                    (clustered arm)
                local ind_clust_var = (exp(M41[1,4]+ M41[1,5]))^2 //individual
                    variance (clustered arm)
```

```
                        local ind_cont_var = (exp(M41[1,4]))^2 //individual variance (
                            control arm)
                        local estrho `btw_clust_var'/(`btw_clust_var'+`ind_clust_var') //
                             ICC
                        *post the results of each model output from each simulation
                        post `memhold' (`theta') (`gam') (`rho') (`clusterSize') (`
                            nClusters') (`m') (`b') (`se') (`p') (`ll') (`ul') (`cv') (`
                            ind_cont_var') (`ind_clust_var') (`btw_clust_var') (`estrho')
                             (`converge')
                }
                if _rc != 0 {
                        local converge = 0
                        local m 4.1
                        post `memhold' (`theta') (`gam') (`rho') (`clusterSize') (`
                            nClusters') (`m') (.) (.) (.) (.) (.) (.) (.) (.) (.) (.) (`
                            converge')
                        }

}
        clear
}

qui postclose `memhold'

qui use `results', clear

/* After all all simulations are run and the N-obsrvation dataset of results is
    created, code ends with */
note: File results`simul_No'
note: Ran simluation `0'
note: Time taken (minutes) `tmins'
note: Seed was `initial_seed'

save dta\results`simul_No', replace

/* Closed and convert log to html */
log close
translate log\simulation`simul_No'.smcl log\simulation`simul_No'.log, replace
```

A.3 Publication

Candlish et al. BMC Medical Research Methodology (2018) 18:105
https://doi.org/10.1186/s12874-018-0559-x

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Appropriate statistical methods for analysing partially nested randomised controlled trials with continuous outcomes: a simulation study

Jane Candlish*, M. Dawn Teare, Munyaradzi Dimairo, Laura Flight, Laura Mandefield and Stephen J. Walters

Abstract

Background: In individually randomised trials we might expect interventions delivered in groups or by care providers to result in clustering of outcomes for participants treated in the same group or by the same care provider. In partially nested randomised controlled trials (pnRCTs) this clustering only occurs in one trial arm, commonly the intervention arm. It is important to measure and account for between-cluster variability in trial design and analysis. We compare analysis approaches for pnRCTs with continuous outcomes, investigating the impact on statistical inference of cluster sizes, coding of the non-clustered arm, intracluster correlation coefficient (ICCs), and differential variance between intervention and control arm, and provide recommendations for analysis.

Methods: We performed a simulation study assessing the performance of six analysis approaches for a two-arm pnRCT with a continuous outcome. These include: linear regression model; fully clustered mixed-effects model with singleton clusters in control arm; fully clustered mixed-effects model with one large cluster in control arm; fully clustered mixed-effects model with pseudo clusters in control arm; partially nested homoscedastic mixed effects model, and partially nested heteroscedastic mixed effects model. We varied the cluster size, number of clusters, ICC, and individual variance between the two trial arms.

Results: All models provided unbiased intervention effect estimates. In the partially nested mixed-effects models, methods for classifying the non-clustered control arm had negligible impact. Failure to account for even small ICCs resulted in inflated Type I error rates and over-coverage of confidence intervals. Fully clustered mixed effects models provided poor control of the Type I error rates and biased ICC estimates. The heteroscedastic partially nested mixed-effects model maintained relatively good control of Type I error rates, unbiased ICC estimation, and did not noticeably reduce power even with homoscedastic individual variances across arms.

Conclusions: In general, we recommend the use of a heteroscedastic partially nested mixed-effects model, which models the clustering in only one arm, for continuous outcomes similar to those generated under the scenarios of our simulations study. However, with few clusters (3–6), small cluster sizes (5–10), and small ICC (≤0.05) this model underestimates Type I error rates and there is no optimal model.

Keywords: Clustering, Randomised controlled trial, Partially nested, Partially clustered, Therapist effects, Individually randomised group treatment, Individually randomised cluster trial, Intervention studies

* Correspondence: jane.candlish@sheffield.ac.uk
School of Health and Related Research (ScHARR), University of Sheffield, 30 Regent Street, S1 4DA, Sheffield, UK

Background

Randomised controlled trials (RCTs) are often categorised into two types: individually randomised controlled trials (iRCTs) where participants are individually randomised to treatment arms to receive one of the investigative treatments; and cluster randomised controlled trials (cRCTs) where groups of participants (clusters) are randomised to treatment arms. We may expect outcomes for participants within the same cluster to be more similar than those from different clusters. The similarity can arise due to participants in the same cluster receiving care from the same health provider or interacting with one another [1]. The implications of clustering in cRCTs are widely acknowledged [1, 2]. Clustering results in a reduction in statistical efficiency in cRCTs and if ignored standard errors and p-values for intervention effects are typically underestimated.

Clustering can also occur in iRCTs. For instance, clustering of participants outcomes due to receiving treatment as part of a group-based parenting intervention [3], treatment in specialist clinics for the treatment of venous leg ulcers [4], or participants under the care of a surgeon for comparison for hemostasis in elective benign thyroid surgery [5]. The care provider or group dynamics may play a role in the causal pathway of the intervention effect. We might expect correlated outcomes between individuals either in the same group or receiving treatment from the same care provider.

Standard sample size and analysis methods for iRCTs rely on the assumption of independence between participants, which is violated when clustering is present. The 'clustering effect' is commonly quantified using the intracluster correlation coefficient (ICC). The ICC measures the extent to which outcomes from participants within the same cluster are correlated to one another [1]. When designing and analysing iRCTs with clustering we need to consider implications of the potential lack of independence. Ignoring clustering in the analysis can lead to over precise results and consequently incorrect conclusions [1]. Clustering of any form results in a reduction in the effective sample size, hence, there is a reduction in the power to detect an intervention effect if it truly exists.

In addition to obtaining sufficient power and accurate results, accounting for clustering enables us to estimate the ICC. ICCs are often important for the interpretation of trial results; we may be directly interested in the intervention group or care provider effects. ICCs are also key when calculating sample sizes for RCTs with clustering, in order to maintain power [1].

An increasingly applied design in healthcare and education research is a partially nested randomised controlled trial (pnRCT) [6], where participants are individually randomised to trial arms and clustering of outcomes occurs in only one arm of the trial [7] (also termed partially clustered trials). The STEPWISE trial is an example of a pnRCT, assessing a structured lifestyle education programme aimed at supporting weight loss for adults with schizophrenia and first episode psychosis in a community mental health setting. Individuals were randomised to either an intervention arm of group-based lifestyle education sessions or a control arm receiving usual care at the individual level [8].

Specific statistical methods need to be used for analysing pnRCTs. Consequently, there has been a considerable growth in the methodology literature (particularly in the fields of psychotherapy and educational research) in the past few decades both proposing and reviewing statistical methods for pnRCTs.

Table 1 presents a summary of relevant literature on the analysis of pnRCTs. This expands on the literature search by Flight et al. [9] summarising models for the analysis of pnRCTs. Sample size calculations for pnRCTs have been addressed elsewhere [10–14]. Analysis methods for pnRCTs have mainly focussed on using mixed-effects models, individual-level models which account for the hierarchical structure of the data [6, 7, 9, 15–19]. These models allow us to control for baseline covariates and represent the different levels in the data, including cluster, individual, and repeated measures (where applicable). In addition to accounting for the clustering, we may expect the variance of the individual errors to differ between trial arms in pnRCTs, termed heteroscedastic variance [7]. When a clustered intervention arm is compared to a non-clustered control arm the between-cluster variation in the intervention arm is not present in the control arm. The clustered intervention may result in a decrease or increase of the individual level variability.

In this study, we use a series of simulations to evaluate the statistical analysis models for two-arm parallel pnRCTs with continuous outcomes, assessing a range of scenarios including the effect of cluster size and the number of clusters. In theory, the mixed-effects models can be formulated so that they do not model clustering in the control arm, however, when running these models in statistical software packages it is necessary to impose some form of clustering in the control arm. The literature identified in Table 1 highlighted that research to date is lacking in addressing the best way to treat the non-clustered control arm when fitting the models in statistical software, using scenarios of relevance in the field of public health with small clusters and small ICCs [9], and evaluating the effect of the variance ratio of the residuals on the model fit. In pnRCTs we may have small numbers of clusters [9], thus we evaluate the impact of the number of clusters on statistical inference and if statistical inference remains valid using mixed-effects models.

Table 1 Summary of relevant literature on analysis of pnRCTs

Paper	Relevant themes	Range of values[a]	Findings
Schweig & Pane [16]	Describe and compare models for pnRCTs with non-compliance using a simulation study.	Simulation for two levels of clustering, exact cluster sizes (m) unclear in paper, $c_{school} = 37$, $c_{class} = 177$, $\lambda_B = 2, 8$, $\rho_{school} = 0.005, 0.05, 0.15$, $\rho_{class} = 0.0004, 0.10, 0.25$, and $\theta = 0.087$.	Clustering and non-compliance may have a substantial impact on statistical inference about intention-to-treat effects. Provide methods that may accommodate pnRCT with non-compliance, recommend using complier average causal effect estimate (CACE) and scaling by the proportion of compliers. No mention of degrees of freedom, we have assumed they used default degrees of freedom method available in R lme packages.
Flight et al. [9]	Compare models applied to four examples of pnRCTs. Compare three different methods for classifying the non-clustered control arm in pnRCTs, including: singleton clusters, one large cluster and pseudo clusters.	Examples with $\{m, c\} = \{36, 8; 24, 7; 14, 8; 5, 6\}$, and estimated $\rho = < 0.0001, 0.02, 0.007$.	Recommend use of the heteroscedastic model, recommendations based only on re-analysis of case studies. Methods for classifying the non-clustered control arm in pnRCTs had a large impact in fully clustered mixed effects models and no measurable impact in partially nested mixed-effects models. ICCs in four examples were small.
Sterba [27]	Review of modelling developments for pnRCTs, focused on those particularly relevant to psychotherapy trials.		Recommend the inclusion of cluster variability in analysis model as it provides insight into treatment process (rather than treating it as a nuisance). Annotated Mplus commands for models
Lohr, Schochet & Sanders [19]	Report presenting a guide to design and analysis issues for pnRCTs in education research, using example trials. Discussion of degrees of freedom issue in Appendix.		Guidance document, defines pnRCT in context of education research and show methods to analyse these using SAS. Provide SAS commands for model fitting in examples.
Korendijk [18]	Compare models for pnRCTs using simulation study, investigate mis-specification for the estimation of the parameters and their standard errors.	Simulation study with $m = 5$, $c = 10, 30, 50, 100$, $\rho = 0.05, 0.1, 0.2$, $\lambda_A = 1$, $d = 0.3$.	All models perform comparably with respect to fixed effect estimates. Recommend use of partially nested mixed-effects model. Simulations were under null and ICC always greater than zero. No mention of degrees of freedom, we have we assumed default degrees of freedom used from MLwiN software, and homoscedasticity was assumed for ndividual variances between the two arms.
Sanders [17]	Compare models for pnRCTs using simulation study in terms of Type I error and power	Simulation study with $\{m, c\} = \{2, 10; 4, 4; 5, 4; 10, 2\}$, $\rho = 0, 0.1, 0.2, 0.3, 0.4, 0.5$, $\lambda_A = 1$, and $\omega^2 = 0, 0.01, 0.059, 0.138$.	Type I error rate increased as ICC increased, Satterthwaite degrees of freedom had better control than Kenward-Roger degrees of freedom. Found using mixed-effects model for pnRCT when ICC is zero likely leads to never detecting intervention effects, observed Type I error rates nearly non-existent under all scenarios with ICC equal to zero. Recommend to evaluate if ICC is significantly different from zero prior to selecting analysis method. Homoscedas ticity was assumed for individual variances between the two arms.
Baldwin et al. [15]	Compare analysis models for pnRCT simulation study, comparing three degrees of freedom calculations, and a pnRCT example.	Simulation for $m = 5, 15, 30$, $c = 2, 4, 8, 16$, $\rho = 0, 0.05, 0.1, 0.15, 0.3$, $\lambda_B = 0.25, 1, 4$, and $d = 0, 0.5$.	Recommend pnRCTs take account of heteroscedasticity. Satterthwaite and Kenward-Roger degrees of freedom control Type I error rate. The heteroscedas tic model provides an unbiased estimate and little reduction in power compared to the homoscedastic model. Argue that using a partially nested mixed-effects model only a problem for statistical inference when the number of clusters is small. The number of clusters has greater impact on power in

Table 1 Summary of relevant literature on analysis of pnRCTs *(Continued)*

Paper	Relevant themes	Range of values[a]	Findings
			pnRCTs. At least eight, preferably 16 clusters, to maintain Type I error rate.
Bauer et al. [6]	Review of RCTs to ascertain the prevalence of pnRCTS in four public health and clinical research journals. Analysis models for pnRCTs extended to include pre-test measures as covariates, individual and group level covariates, and example of pnRCT	Example with clustering in one arm $c = 41$, $m = 9$, and estimated $\rho = 0.02$.	Out of 94 RCTs, 32% were pnRCTs, 40% iRCTs and 27% cRCT. None used methods specific to pnRCTs. Example pnRCT data could be analysed using mixed-effects models. Argue pnRCTs "often increase external validity at the expense of internal validity" (p.20).
Roberts & Roberts [7]	Examine the case of pnRCTs, heterogeneity, comparison of analysis methods for simulation study and present an example.	Simulation for m = 6, $c = 8$, $\rho = 0$, 0.1, 0.2, 0.3, $\lambda_A = 0.5$, 0.75, 1, 1.25, 1.5, 1.75, 2 and $d = 0$.	Recommend pnRCTs take account of heteroscedasticity. Satterthwaite unequal variances t-test gave robust to heterosce dasticity. The heteroscedastic model gives slightly inflated test size for large ρ: suggest Satterthwaite degrees of freedom as a solution.
Lee & Thompson [28]	Describe analysis models for iRCTs with clustering and apply to two examples (using Bayesian approach)		Show that ignoring clustering may underestimate uncertainty, leading to incorrect conclusions.
Hoover [34]	Statistical tests for RCTs with clustering that differ across trial arms.	Example with clustering in both arms with $m = 7 - 12$, $c = 3$.	Provide an adjustment for the independent samples t-test for pnRCTs. Statistical impact of heterogeneity effect increases as the cluster size increases, and as heterogeneity increases. The test does not allow for the inclusion of covariates, multiple treatments, baseline measures, or non-normally distributed outcomes.

[a] m = cluster size, c = number of clusters, ρ= ICC, d = standardised effect size, ω^2= Omega Squared effect size percent of variability accounted for by treatment condition, λ_A= ratio of total variance in control arm compared to clustered, λ_B= ratio of individual variance in control arm compared to clustered arm. Ordered by year of publication

We evaluate and provide recommendations for the most appropriate analysis methods for pnRCTs, including:

1) where mixed-effects models are necessary,
2) methods of specifying the clusters in the non-clustered arm when fitting a model and the impact these have on the analysis,
3) the impact of cluster sizes and the number of clusters on statistical inference and,
4) the impact of heteroscedastic individual variance between trial arms on statistical inference.

Methods

Methods for analysis of partially nested trials

In this section, we present the main modelling approaches currently available and used for pnRCTs, including ignoring clustering altogether, imposing clustering in the non-clustered control arm, and explicitly modelling the partially nested design by modelling clustering only in the intervention arm.

It is possible to account for clustering by including each cluster as a fixed effect in a standard regression model, in addition to a fixed effect representing the intervention effect. Although this method is simple to implement, it is not recommended. Firstly, it does not reflect the study design of a pnRCT and may require a large number of fixed effects to be fitted lowering the degrees of freedom [9]. Secondly, if clusters are of size one there is insufficient information to estimate both the intervention effect and the cluster effect for each cluster. Finally, it will generally underestimate the intervention effect variability as the cluster level variability is removed.

Table 2 presents a summary of the models for the analysis of pnRCTs using findings from the literature search by Flight et al. [9]. We define: y as a continuous outcome, i is the individual participant indicator, j is the cluster indicator, t is the intervention indicator (0 = control, 1 = intervention), θ is the intervention effect, β_0 is an intercept term. Error terms are defined depending on the model procedure, represented using ϵ, u, and r, where u represents the between cluster variation and ϵ and r represent individual level variation.

Model 1 (Table 2) is the linear regression model which ignores the clustering and uses analysis for non-clustered trials, assuming independence between individuals regardless of whether they are in the same cluster. This infers that the conditional variance of y in both the intervention and control arms is equal. If the outcomes of individuals in the same cluster are correlated, the independence assumption

Table 2 Models for the analysis of pnRCTs

Model description		Statistical model	Heteroscedastic residuals
Model 1	Linear regression (ignore clustering)	$y_i = \beta_0 + \theta t_i + \epsilon_i$ • $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ the individual level variation	No
Model 2	Fully clustered (impose clustering)	$y_{ij} = \beta_0 + \theta t_{ij} + u_j + \epsilon_{ij}$ • $u_j \sim N(0, \sigma_u^2)$ a random effects term representing between cluster variation • $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ the individual level variation	No
Model 3	Partially nested homoscedastic	$y_{ij} = \beta_0 + \theta t_{ij} + u_j t_{ij} + \epsilon_{ij}$ • $u_j \sim N(0, \sigma_u^2)$ a random effects term representing between-cluster variation in clustered arm • $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ the individual level variation	No
Model 4	Partially nested heteroscedastic	$y_{ij} = \beta_0 + \theta t_{ij} + u_j t_{ij} + r_{ij}(1 - t_{ij}) + \epsilon_{ij} t_{ij}$ • $u_j \sim N(0, \sigma_u^2)$ a random effects term representing between cluster-variation in clustered arm • $r_{ij} \sim N(0, \sigma_r^2)$ the individual level variation in the non-clustered control arm. • $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ the individual level variation in the clustered arm	Yes

is violated and we underestimate uncertainty around intervention effects when using the linear regression model above.

Model 2 (Table 2) is the fully clustered mixed-effects model which includes the cluster as a random effect; this includes variability at both the individual and cluster level. The mixed-effects model with imposed clustering of the control arm requires the estimation of a random cluster effect for both intervention and control groups. Some options for the imposed clustering in the control arm are given in Table 3. The variance of the control arm and intervention arm are assumed to be the same (homoscedastic). When the variance is believed to differ between control and intervention arm model 2 is not appropriate as it does not account for heteroscedasticity. Models 3 and 4 (Table 2) apply the cluster effect to the clustered arm only [7, 10, 11, 14], we term these the partially nested models.

Individuals in the non-clustered control arm are assumed independent. This accurately reflects the design of the study with clustering only in one arm. In the partially nested homoscedastic model, we apply the random effect u_j to the clustered intervention arm only; between-cluster variability is only present for the intervention arm. Model 3 is homoscedastic as the variance of the individual errors, ϵ_{ij}, between arms is the same. In practice, the variance of the individual errors may differ between trial arms [7]. Therefore, model 3 is extended to a partially nested heteroscedastic model, model 4, this allows for differing individual errors between intervention and control arms but does not constrain the form of heteroscedasticity.

Table 3 Options for imposing clustering in the non-clustered control arm

Option	Cluster	Intervention
1	$j = 0$	$j = 1, \ldots, c$
2	$j = 1, \ldots, l$	$j = l + 1, \ldots, c$
3	$j = 1, \ldots, k$	$j = k + 1, \ldots, c$

Imposed clustering in the control arm

Regardless of whether or not the model assumes clustering in one (models 3 and 4) or both arms (model 2), within the statistical software package a decision must be made about how to code the cluster indicators in the control arm. One method is to impose clusters for all individuals, including those in the control arm, and use analysis for cRCTs with clustering in both arms.

Table 3 represents the different options for imposing clustering, j, in the control arm, l is the number of individuals in the control arm and k is the number of arbitrary clusters in the control group. Option one treats the control group as one single cluster; option two treats each individual in the control arm as their own cluster of size one (singleton clusters) giving $j=l$ clusters in the control arm. ICC estimation can be problematic with options one and two, in theory, it is not possible to estimate between-cluster variability in option one, or estimate within cluster variability in the control group using option two [20]. Option three imposes artificial pseudo-random clusters in the control group to overcome the problem of estimating within or between-cluster variability. The number of arbitrary clusters, k, needs to be considered. We chose it to be equal across treatment arms. In addition, option three will likely result in a lower ICC estimation due to the assumed independence of control participants.

In our simulation study, the fully clustered model 2 is parametrised using the imposed clustering from Table 3. The models are:

- Model 2.1 fully clustered mixed-effects model with singleton clusters in the control arm;

- Model 2.2 fully clustered mixed-effects model with one large cluster in the control arm;
- Model 2.3 fully clustered mixed-effects model with pseudo clusters in the control arm.

Flight et al. [9] investigated the effect of the different methods of imposing clustering in the control arm presented in Table 3 in four pnRCT case-studies. The four case-studies covered trials evaluating the effect of: specialist leg ulcer clinics (clustered by clinic), acupuncture for low back pain (clustered by acupuncturist), postnatal support in the community (clustered by community support worker), and telephone befriending for maintaining quality of life in older people (clustered by volunteer facilitator). Little difference was found between the different methods for the fully clustered mixed-effects models and there was no difference between the different methods for the partially nested mixed-effects models.

We anticipated that the method of imposing the clustering in the control arm does not affect the results of the methods which model clustering in only one arm, however, this evaluated in the simulation study.

Degrees of freedom for fixed effect estimates

In the mixed-effects models above we wish to carry out significance tests for the intervention effect. In addition to the correct choice of model, the test statistics and degrees of freedom in mixed-effects models also need to be considered. For large sample sizes in mixed-effects models, the test statistics for fixed effects can be assumed Normally distributed. However, for small samples, the t-distribution is generally used as an approximation of the distribution of the test statistic. Estimating the degrees of freedom for the t-distribution is unclear for pnRCTs and will affect both the significance test and the confidence intervals of the intervention effect estimate.

Comparison of degrees of freedom correction methods has been undertaken for cRCTs and pnRCTs with small numbers of clusters [15, 21]. The Satterthwaite small-sample degrees of freedom correction takes into account the variance structure of the data, for pnRCTs, it has been shown to be superior to the between-within method for maintaining Type I error rates (and comparable to the Kenward-Roger method) [15]. Following these results, the Satterthwaite approximation was used to calculate degrees of freedom (using dfmethod() option for mixed, available in Stata 14 onwards [22]).

Simulation study

Overview

We performed a simulation study to evaluate the statistical analysis models for pnRCTs presented in Table 2, and the imposed clustering of the control arm in Table 3 [23]. All models were fitted using a restricted maximum likelihood procedure (REML). All simulations were done in Stata [22], graphs produced using ggplot2 [24] in R [25]. See Additional file 1 for simulation code.

Data-generating mechanism

We simulated data to replicate a two-arm parallel pnRCT trial with a non-clustered control arm and a clustered intervention arm (randomised 1:1) and a continuous outcome. We simulated data under various design scenarios and under both the null ($\theta = 0$) and alternative hypothesis ($\theta = A$, where $A \neq 0$).

Data were simulated from the following model with the intercept set to zero and group allocation denoted by t ($t = 0$ for control, $t = 1$ for intervention arm):

- For the intervention arm $(t = 1)$ $y_{ij} = \theta + u_j\sqrt{\rho} + z_{ij}\sqrt{1-\rho}$
- For the control arm $(t = 0)$ $y_{ij} = z_{ij}\sqrt{\gamma(1-\rho)}$

Where $u_j \sim N(0, 1)$ and $z_{ij} \sim N(0, 1)$. This simulates an ICC of ρ and a ratio of individual level variance between the non-clustered control arm and the clustered intervention arm of γ. If $\gamma = 1$, there is homoscedasticity between the individual level variance in the control and intervention arms. Full simulation study steps, including the data generation process and modelling, are presented in Fig. 1.

Simulation scenarios are presented in Table 4. We varied: the intervention effect, ICC, cluster size, number of clusters, and ratio of individual variance between the trial arms. If $\gamma = 0.25$ then individual variance in the control arm is one quarter that in the intervention arm and if $\gamma = 4$ then individual variance in the control arm is four times that in the intervention arm.

Simulation values were chosen based on literature on pnRCTs [7, 9, 15, 17, 18, 26–28], as well as extending these to more extreme cases of γ and ρ that may occur in rare instances. Reporting of ICCs in iRCTs with clustering is limited at present and it is plausible that ICCs in pnRCTs differ from those of cRCTs. Current evidence suggests ICCs in iRCTs with clustering in either one or both arms are generally small and often less than 0.05 [7–9, 29], hence the choice to include a small ICC $\rho = 0.01$ in the simulations with ICCs of 0.2 or more occurring only in rare instances. We were unaware of specific literature on the evidence of heteroscedasticity, however, from the authors experience of working on trials it was expected γ to typically stay within the range of 0.5–2. The number of clusters in the intervention group was 3, 6, 12 or 24. These figures reflect the small numbers of clusters recruited in many pnRCTs and, coupled with the cluster sizes of 5, 10, 20 or 30, they allowed alternative combinations of cluster size and number of clusters

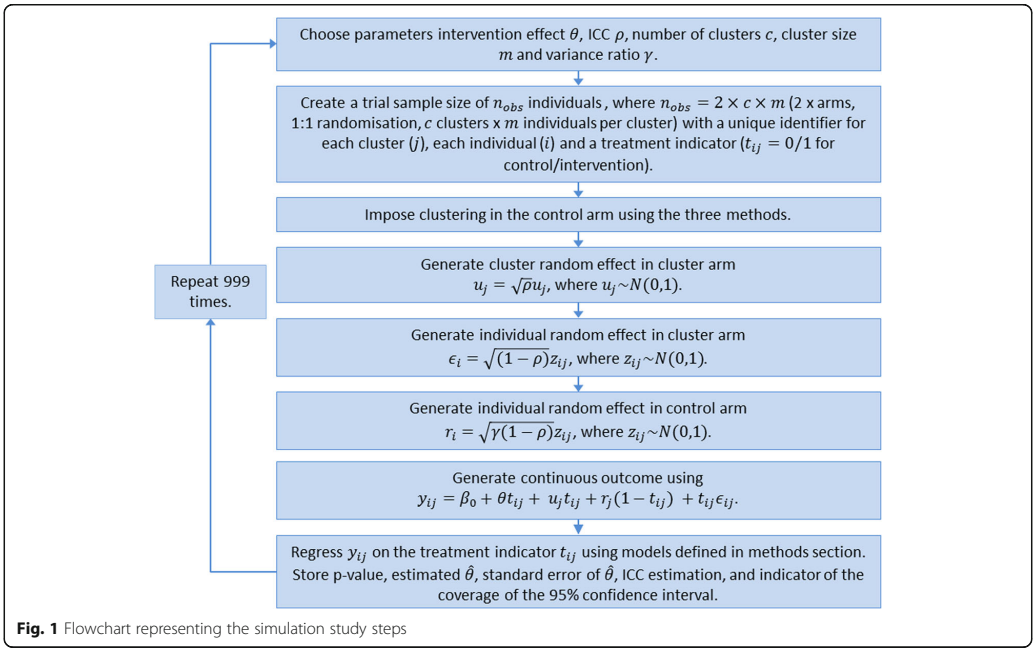


Fig. 1 Flowchart representing the simulation study steps

to be investigated for a given total trial sample size. Figure 2 provides a graphical example of the simulated partially nested trial data.

Methods

Each simulated dataset was analysed using the models described in Table 2.

Estimand

Our estimands of interest are the REML estimate of the intervention effect θ and the model estimate of the ICC ρ.

Table 4 Simulation input scenario values (total 1440 scenarios)

Variable	Notation	Values
Number of clusters	c	3, 6, 12, 24
Cluster size	m	5, 10, 20, 30
Intervention effect	θ	0, 0.2, 0.5
ICC	ρ	0, 0.01, 0.05, 0.1, 0.2[a], 0.3[a]
Ratio of individual variance between control and cluster trial arms	γ	0.25[a], 0.5, 1, 2, 4[a]

[a]Considered extreme values to occur in rare scenarios

Performance measures

We used the following performance measures: bias, mean square error (MSE), and coverage of 95% confidence intervals for $\hat{\theta}$, Type I error rate and power (calculated as the proportion of simulated results with a statistically significant intervention effect at the 5% level when the null or alternative hypothesis were true, for Type I error and power respectively) and where applicable, model estimated ICC. See Additional file 2 for more detail on performance measures. For each of the 1440 scenarios 1000 datasets were generated; a 5% significance level and 95% confidence interval based on 1000 simulations has a Monte Carlo error of 0.7%.

Results

Model convergence was generally satisfactory for all models with models converging 95–100% of the time across the different scenarios.

Imposed clustering in the control arm

Methods for imposing clustering in the control arm, presented in Table 3, had a negligible impact on the performance measures of the partially nested mixed-effects models (models 3 and 4). Under the simulation scenarios, the differences in the p-value, confidence intervals

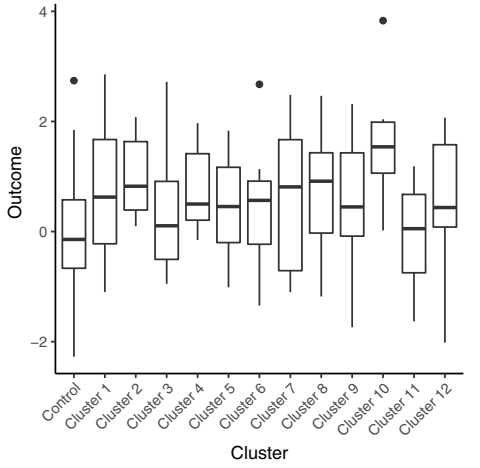


Fig. 2 Example of simulated partially nested trial dataset, $\rho = 0.1$, $\gamma = 1$, $c = 12$, and $m = 10$

and estimated ICC between the methods were only present at four decimal places. Model fitting was considerably faster (around four to five times faster) using either one large cluster or the pseudo clusters compared to the singleton clusters, however, this will likely be immaterial when fitting only a small number of models.

Methods for imposing clustering in the control arm had a large impact on the performance measures of the fully clustered mixed-effects models (models 2.1, 2.2, and 2.3). Specific results for each performance measure are presented in the following sections.

Results are reported only for the partially nested mixed-effects models (models 3 and 4) with the non-clustered controls classified as one large cluster, as other methods were comparable. All three methods for classifying the non-clustered control arms for the fully clustered mixed-effects model (models 2.1, 2.2, and 2.3) are reported.

Bias

The bias of the intervention effect estimate was not affected by the analysis model used, individual variances (γ) or the ICC (ρ). The maximum absolute bias of the intervention effect was |0.057|, |0.043|, and |0.053| for $\theta = 0$, 0.2 and 0.5, respectively.

Mean square error

The models produced unbiased estimators with no difference in the observed MSE between the different models. The MSE of the intervention effect estimate had a mean of 0.051 (standard deviation (SD) 0.056) and maximum of 0.346.

Type I error

Plots of the mean Type I error rates split by model, the ratio of individual variances (γ) and the ICC (ρ) are presented in Fig. 3. As would be expected the linear regression model which ignores clustering had inflated Type I error rates, with Type I error rate affected by ICC (ρ), the ratio of individual variances (γ), number of clusters (c), and cluster size (m). Although the inflation was minimal when ICC $\rho = 0.01$, the mean Type I error was 0.061 (SD 0.010). When cluster size $m \leq 10$ and ICC $\rho = 0.01$ the mean Type I error rate was 0.056 (SD 0.007).

Model 2, the fully clustered models with imposed clustering in the control arm resulted in biased Type I error rates. Imposing clustering as singleton clusters (model 2.1) led to Type I error rates which were largely affected by the ratio of individual variances (γ) and ICC (ρ). Imposing one large cluster in the control arm (model 2.2) resulted in Type I error rates of zero (due to the Satterthwaite degrees of freedom correction resulting in large degrees of freedom when imposing one large cluster in the control arm). Imposing pseudo clusters in the control arm of the same size as the intervention arm (model 2.3) provided relatively good control of Type I error rates, mean Type I error of 0.039 (SD 0.018), but was affected slightly by both the ratio of individual variances (γ) and ICC (ρ).

Both the homoscedastic and heteroscedastic partially nested models (models 3 and 4) provided good control of Type I error rates (model 3: mean Type I error 0.045 (SD 0.016) and model 4: mean Type I error 0.044 (SD 0.014)) with little difference present between the two models.

For more detailed comparison Fig. 4 presents the Type I error rates for the linear regression model (model 1), the homoscedastic (model 3) and the heteroscedastic (model 4) partially nested models by ICC (ρ), the ratio of individual variances (γ), number of clusters (c), and cluster size (m). Higher ICC values resulted in higher Type I error rates in each model. The impact of ignoring clustering (model 1) depends on both ICC (ρ), cluster size (m), and number of clusters (c). Larger number of clusters (c) resulted in better control of Type I error rates for the partially nested models. When ICC $\rho = 0$, the Type I error rates of the partially nested models (models 3 and 4) were reduced from the nominal level. This is due to the cluster variance components being estimated when they are not actually present in the data. When the ICC was small ($\rho \leq 0.05$) and the individual variance in the control arm smaller than that in the intervention arm ($\gamma < 1$), the Type I error rates of partially nested models were reduced from the nominal 5% level. When ICC was large ($\rho = 0.3$) the partially nested models generally resulted in inflated Type I error rates. As ICC increased Type I error rates increased, with the

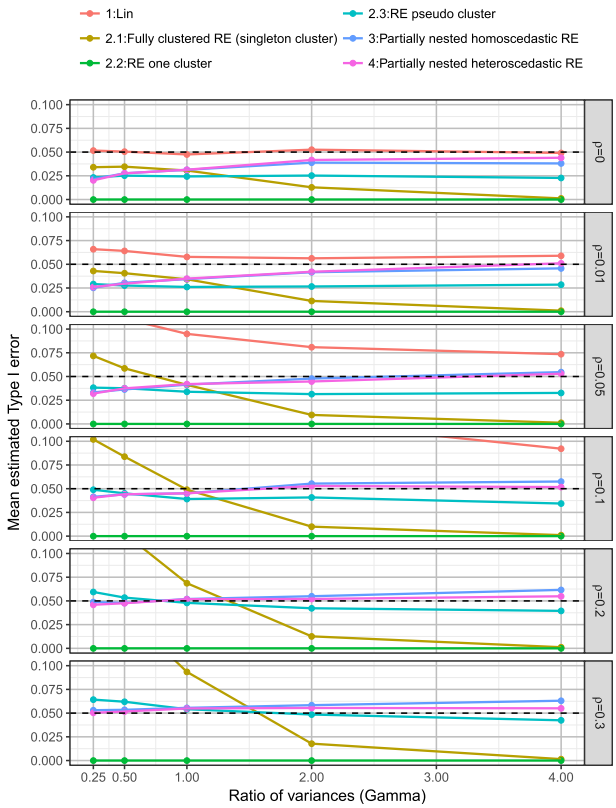


Fig. 3 Mean Type I error rate by γ and ρ over all scenarios, for each model

partially nested models 3 and 4 only reaching above the nominal Type I error rate of 5% on average when ICC $\rho \geq 0.2$.

The Satterthwaite correction used in Stata mixed (dfmethod(sat)) did not fully correct the Type I error rates to the nominal 5% level, even with the use of the heteroscedastic model 4. The heteroscedastic model 4 did have slightly improved control of Type I error rates than the homoscedastic model 3.

Coverage

Plots of the mean coverage of the 95% confidence intervals of the intervention effect estimate split by model, ICC (ρ) and the ratio of individual variances (γ) are presented in Fig. 5 under the alterative hypothesis. The linear regression model (model 1) resulted in under coverage when ICC was small ($\rho \leq 0.05$) and the coverage rates decrease as ICC (ρ) increases. The fully clustered models with imposed clustering in the control arm resulted in both over and under coverage dependent on the direction of the variance ratio and the method of imposed clustering. Imposing clustering as singleton clusters (model 2.1) resulted in coverage rates largely affected by ratio of individual variances (γ). Imposing one large cluster in the control arm (model 2.2) resulted in over coverage, due to the reduced Type I error rates of zero caused by the Satterthwaite degrees of freedom correction. Imposing pseudo clusters in the control arm (model 2.3) provided mean coverage rates of 0.961 (SD 0.018).

Both the homoscedastic and heteroscedastic partially nested models (models 3 and 4) provided good control of coverage rates of 95% confidence intervals (model 3: mean coverage rate 0.956 (SD 0.014) and model 4: mean coverage rate 0.956 (SD 0.014)) with little difference between the two models. In the simulations over coverage of the 95% confidence intervals for the heteroscedastic model 4 occurred when ICC $\rho \leq 0.05$, except when the

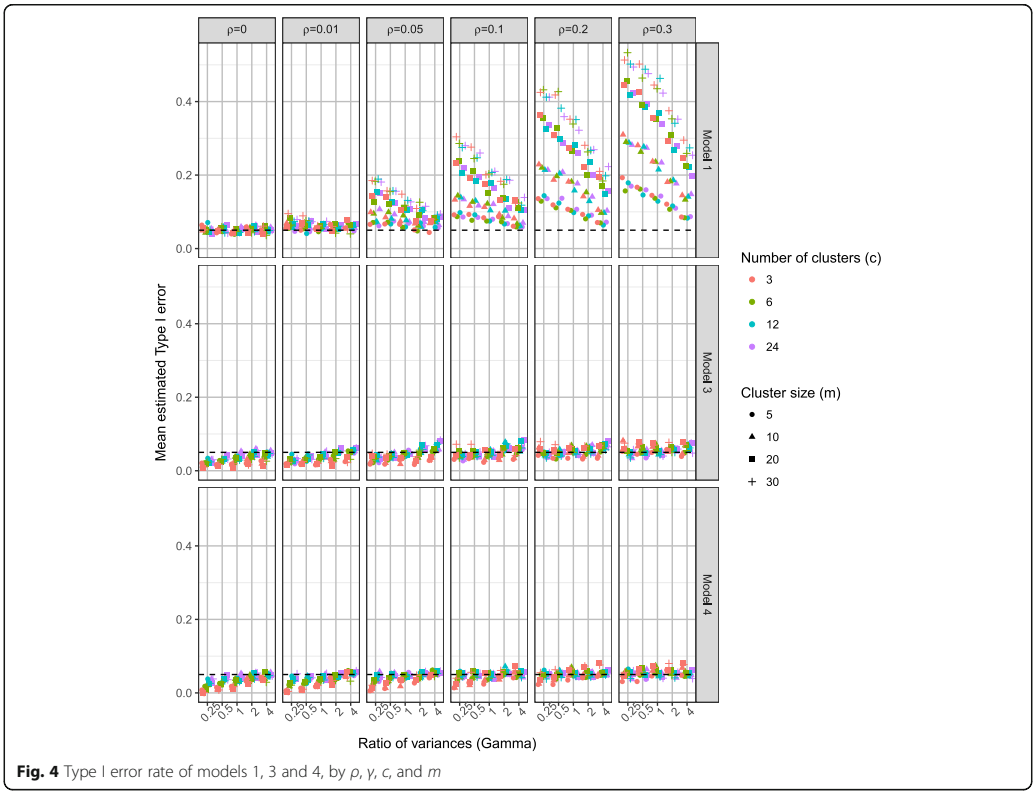


Fig. 4 Type I error rate of models 1, 3 and 4, by ρ, γ, c, and m

ratio of individual variances $\gamma = 4$. Hence, the results were generally conservative when ICC was small. Under coverage of the 95% confidence intervals for the heteroscedastic model 4 only occurred for large ICC (ρ) and ratio of individual variances (γ).

Power

Increasing the number of clusters as opposed to increasing the cluster size had a bigger impact on power with a fixed total sample size. Fig. 6 shows the power of the linear regression model (model 1), the homoscedastic (model 3) and the heteroscedastic (model 4) partially nested models when intervention effect $\theta = 0.5$ by ICC (ρ), the ratio of individual variances (γ), number of clusters (c), and cluster size (m) (see Additional file 2 for when $\theta = 0.2$). Under the simulation scenarios conducted, 12 or more clusters and cluster sizes of ten or more were generally needed for a power greater than 80%. Using three or six clusters rarely gave power over 80%, only for ICC $\rho \leq 0.05$ and relatively large cluster sizes $m \geq 20$, did power go over 80%.

For ICC $\rho \leq 0.05$, which is commonly assumed when planning complex intervention trials in healthcare, power of 80% was generally achieved with: 24 clusters of any size, 12 clusters of size ten or more, and six clusters of size 20 or more (120 in each arm).

Under a ratio of individual variances $\gamma = 1$ the total residual variance in both trial arms is equal to one, hence, the intervention effect (θ) we simulated is the standardised intervention effect. Figure 7 shows the power of models 1, 3 and 4 under homoscedastic individual variances ($\gamma = 1$). The heteroscedastic model 4 is over-parameterised in the case of the ratio of individual variances $\gamma = 1$, however, it did not result in a substantially lower power than the homoscedastic model.

Table 5 presents the power of model 4 and model 1 under ICC $\rho = 0$, model 4 is over-parametrised here. There is a loss in mean statistical power which ranged between 1.7 to 6.3%.

ICC

Figure 8 presents the mean estimated ICC across the fully clustered and partially nested mixed effect models,

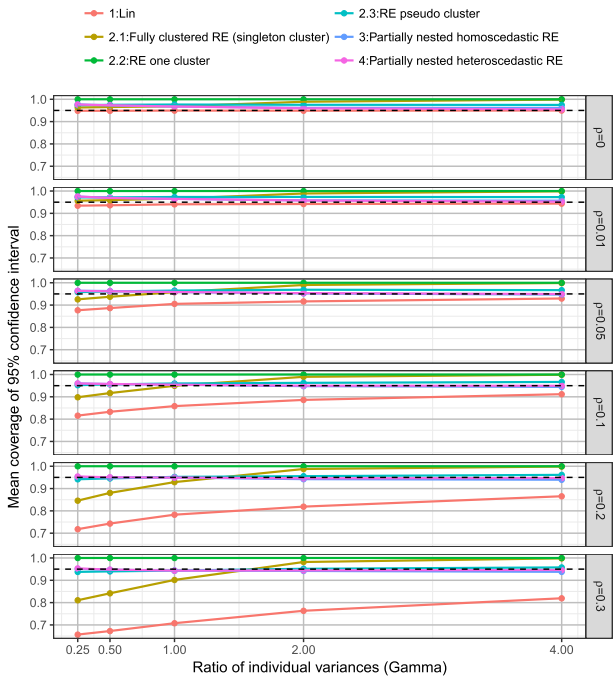


Fig. 5 Mean coverage of 95% confidence interval, by ρ and γ over all scenarios

by the ratio of individual variances (γ) and ICC (ρ). ICC estimation was consistent under the heteroscedastic partially nested model (model 4). The homoscedastic partially nested model (model 3) resulted in biased ICC, with the direction of bias dependent upon the ratio of individual variances (γ).

Figure 9 presents the ICC for the homoscedastic (model 3) and heteroscedastic (model 4) partially nested models by the ratio of individual variances (γ), ICC (ρ), number of clusters (c), and cluster size (m). The ICC estimation from the homoscedastic model was highly affected by γ. The ICC estimation from the heteroscedastic model was not affected by γ. Using the heteroscedastic model, there was a slight positive bias in the ICC estimation when ICC $\rho \leq 0.05$, and when ICC $\rho \geq 0.2$ there was slight negative bias in the ICC estimation. For example, when ICC $\rho = 0.0$ the mean ICC estimation was 0.028 (SD 0.018), and when ICC $\rho = 0.05$ the mean estimation was 0.060 (SD 0.014). As expected ICC estimation improved as sample size increased. The ICC estimation was only consistent across all values of ICC (ρ) when there were 24 clusters, regardless of cluster size. For an accurate estimate of ICC when true ICC $\rho = 0.05$, under the simulation scenarios we required cluster sizes (m) of 20 or 30 or at least six clusters of size ten or 24 clusters of size five.

Summary of results

Simulation results are summarised in Table 6 presenting the performance of the simple linear regression model (model 1), homoscedastic partially nested mixed effects model (model 3) and heteroscedastic partially nested mixed effects model (model 4) under different design scenarios. Results from the fully clustered mixed-effects models (model 2) are excluded from Table 6 as we do not recommend these in any scenario regardless of the method used to impose clustering in the control arm. None of the fully clustered mixed-effects models provided full control of the Type I error rates and the partially nested mixed effects models always outperformed them.

Discussion

In this study, we have investigated six modelling strategies for the analysis of pnRCTs with a continuous outcome. Our simulation study showed that when analysing pnRCTs the use of the heteroscedastic partially nested mixed-effects model for normally distributed outcome

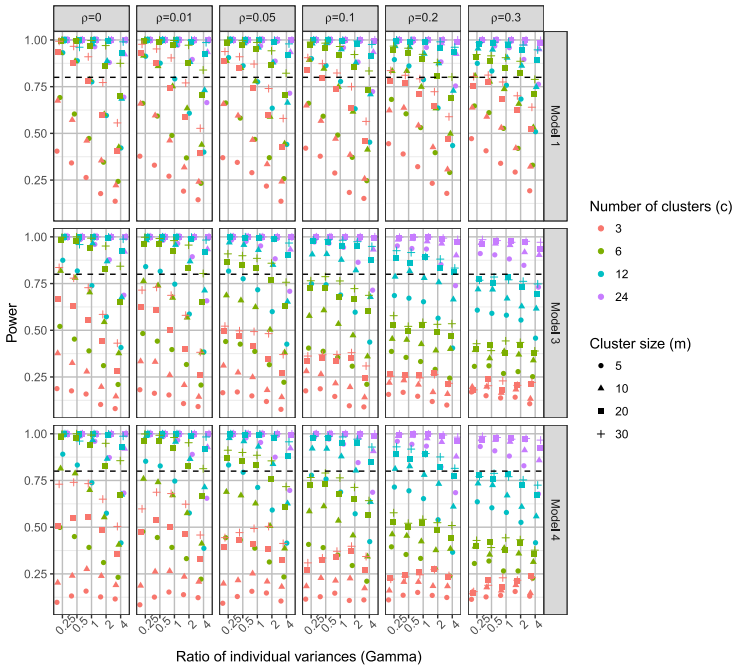


Fig. 6 Power when $\theta = 0.5$, by ρ, γ, c, and m

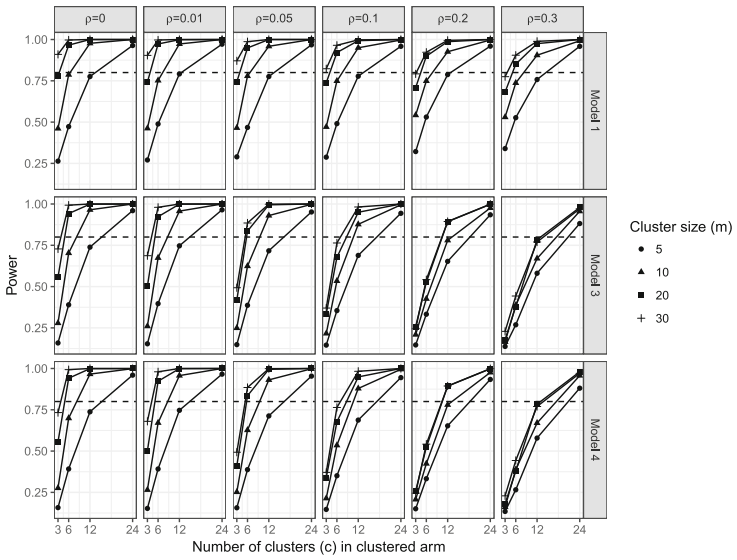


Fig. 7 Power with standardised intervention effect of 0.5 ($\theta = 0.5$ and $\gamma = 1$)

Table 5 Mean and SD of power of model 4 versus model 1 under $\rho = 0$ over all scenarios

Intervention effect (θ)	Model	Power Mean (SD)
0	1	0.050 (0.007)
	4	0.033 (0.014)
0.2	1	0.388 (0.276)
	4	0.327 (0.286)
0.5	1	0.803 (0.254)
	4	0.740 (0.298)

data (using Satterthwaite degrees of freedom) in general provides: unbiased effect estimates; maintains relatively good control of Type I error rates; and did not noticeably cause a reduction in power even with homoscedastic individual variances across arms. The heteroscedastic partially nested model takes account of the between-cluster variance (if present) and therefore provides valid inferences for the intervention effect. Additional file 2 presents model-fitting code for R, Stata and SAS. When using the partially nested mixed-effects model, the method of classifying the non-clustered controls had a negligible impact on statistical inference under the simulation scenarios, agreeing with findings from analysis of four example pnRCTs by Flight et al. [9].

Our findings were broadly similar to those of Baldwin et al. [15]. However, they did not assess the method of classifying the non-clustered controls or performance of models under small ICC ($\rho = 0.01$, lowest value used in our study) which commonly occur in pnRCTs [7–9, 26, 29]. Unlike findings from Baldwin et al. [15], the Satterthwaite degrees of freedom correction did not fully control the Type I error rate in our simulations. The largest discrepancy from the nominal level occurred when the ICC was small, ratio of individual variances <1, and under small sample sizes.

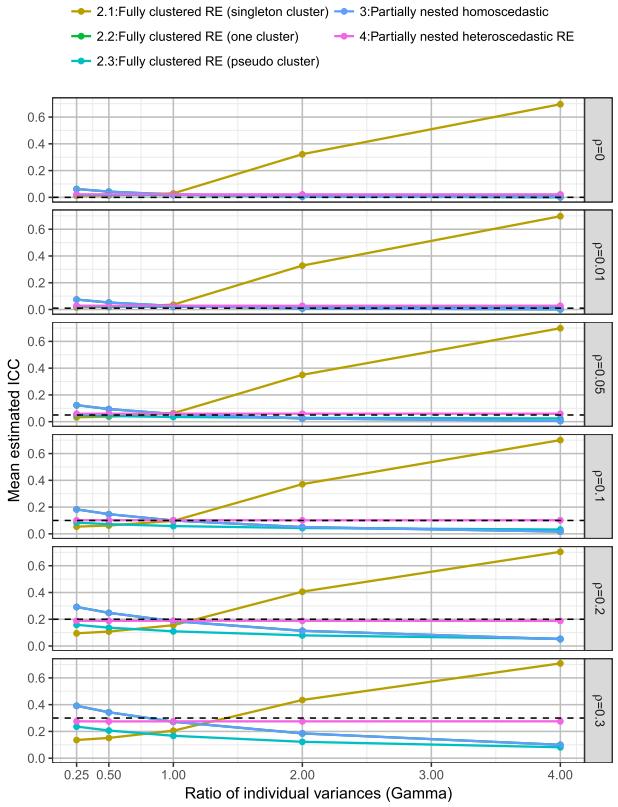


Fig. 8 Mean estimated ICC by γ and ρ over all scenarios, for each model

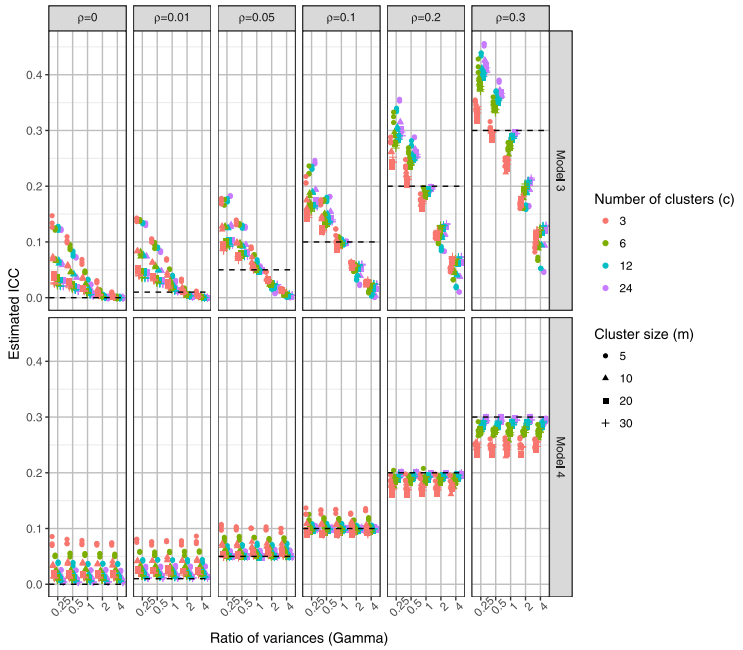


Fig. 9 ICC estimation of heteroscedastic partially nested model, by γ, ρ, m and c

We have illustrated that using a naïve linear regression model, which ignores clustering in pnRCTs, gave inflated Type I error rates and resulted in under coverage of confidence intervals when clustering of outcomes was present. When ICC $0.01 \leq \rho \leq 0.05$, which we believe is common in pnRCTs [9], ignoring clustering led to largely inflated Type I error rates using the linear regression model. A low ICC may still have a large impact, particularly when cluster sizes are large.

When ICC was small and/or with very few clusters and small cluster sizes using the partially nested mixed-effects models 3 and 4 resulted in underestimated Type I error rates. These models correctly reflect the design of the trials; however, they can result in conservativism regarding the precision of estimates due to the bias in estimating the variance estimates when we have a small number of clusters. Consequently, using the partially nested mixed effects models with small ICC may make it difficult to detect differences between the trial arms when present.

Sanders [17] recommend evaluating whether ICC is significantly different from zero prior to selecting an analysis method. We caution such significance testing for ICC, and similarly testing for heteroscedasticity [7]. These tests will generally lack power in a pnRCT and it is not the statistical significance of the ICC that matters but impact of the magnitude on inference. In general, we recommend the use of the partially nested models when analysing pnRCT trials, particularly if conservatism and an ICC estimate are desired. However, the choice of model and the requirement or not for conservatism needs to be considered in the context of the specific trial setting.

Similar to cRCTs [1], in a pnRCT increasing the number of clusters rather than increasing the cluster size has a greater increase in power for a fixed total sample size. Our simulation results showed that this will also provide a more accurate estimation of the ICC. When the number of clusters is small, for example, three clusters in the intervention arm, the ICC estimation will likely be upwardly biased. With six clusters in the intervention arm, the ICC estimate was relatively unbiased once the true ICC ≥0.1. The ICC estimation became consistent regardless of cluster size or true ICC only once there were 24 clusters in the simulation scenarios. This reflects findings from previous research that to reliably estimate the size of clustering effects a large number of clusters are required [30].

This study investigated the case of analysing partially nested trials under complete compliance. Non-compliance in the clustered arm of a pnRCT may occur when some participants randomised to a particular treatment group or

Table 6 Summary of simulation results by different models split by ρ, m, and c averaged over all γ

ICC (ρ)	Cluster size (m)	Number of clusters (c)	Mean (SD) Model 1	Model 3		Model 4	
			Type I error	Type I error	ICC	Type I error	ICC
0	5-10	3-6	0.049 (0.007)	0.025 (0.009)	0.047 (0.043)	0.026 (0.013)	0.047 (0.02)
		12-24	0.052 (0.007)	0.040 (0.010)	0.035 (0.04)	0.042 (0.009)	0.023 (0.01)
	20-30	3-6	0.050 (0.007)	0.023 (0.011)	0.014 (0.012)	0.024 (0.014)	0.013 (0.004)
		12-24	0.050 (0.007)	0.038 (0.010)	0.01 (0.011)	0.04 (0.009)	0.006 (0.002)
0.01	5-10	3-6	0.058 (0.007)	0.028 (0.010)	0.052 (0.043)	0.03 (0.016)	0.052 (0.017)
		12-24	0.055 (0.006)	0.041 (0.011)	0.041 (0.044)	0.043 (0.007)	0.029 (0.01)
	20-30	3-6	0.064 (0.015)	0.029 (0.010)	0.021 (0.016)	0.029 (0.016)	0.019 (0.003)
		12-24	0.066 (0.008)	0.044 (0.012)	0.017 (0.016)	0.046 (0.008)	0.013 (0.001)
0.05	5-10	3-6	0.072 (0.016)	0.031 (0.011)	0.077 (0.057)	0.031 (0.016)	0.079 (0.016)
		12-24	0.071 (0.012)	0.047 (0.012)	0.067 (0.061)	0.048 (0.008)	0.058 (0.007)
	20-30	3-6	0.120 (0.035)	0.041 (0.008)	0.051 (0.031)	0.039 (0.011)	0.052 (0.002)
		12-24	0.123 (0.032)	0.052 (0.017)	0.050 (0.036)	0.050 (0.006)	0.050 (0.001)
0.1	5-10	3-6	0.093 (0.024)	0.037 (0.007)	0.108 (0.068)	0.037 (0.012)	0.114 (0.011)
		12-24	0.092 (0.025)	0.050 (0.013)	0.103 (0.082)	0.050 (0.008)	0.100 (0.004)
	20-30	3-6	0.192 (0.058)	0.053 (0.011)	0.09 (0.046)	0.050 (0.012)	0.093 (0.004)
		12-24	0.185 (0.055)	0.055 (0.015)	0.097 (0.056)	0.050 (0.007)	0.099 (0.002)
0.2	5-10	3-6	0.136 (0.049)	0.047 (0.012)	0.174 (0.091)	0.044 (0.012)	0.187 (0.008)
		12-24	0.135 (0.047)	0.054 (0.011)	0.183 (0.113)	0.051 (0.005)	0.193 (0.004)
	20-30	3-6	0.301 (0.087)	0.06 (0.009)	0.169 (0.072)	0.057 (0.012)	0.177 (0.011)
		12-24	0.286 (0.077)	0.051 (0.011)	0.188 (0.084)	0.049 (0.006)	0.196 (0.003)
0.3	5-10	3-6	0.181 (0.068)	0.056 (0.011)	0.242 (0.108)	0.053 (0.01)	0.262 (0.012)
		12-24	0.177 (0.065)	0.054 (0.012)	0.268 (0.135)	0.050 (0.007)	0.288 (0.006)
	20-30	3-6	0.383 (0.092)	0.065 (0.010)	0.245 (0.090)	0.061 (0.011)	0.258 (0.017)
		12-24	0.368 (0.094)	0.051 (0.009)	0.278 (0.105)	0.050 (0.007)	0.292 (0.005)

*Model 1: simple linear regression; Model 3: homoscedastic partially nested mixed effects model; Model 4: heteroscedastic partially nested mixed effects model. Green highlighted ≤ than expected, red highlighted > than expected

care provider do not attend any sessions or receive treatment as part of different treatment group or care provider intended at randomisation. Consequently, non-complier outcomes may be assumed independent if they do not receive the clustered intervention. Schweig and Pane [16] describe and compare models for pnRCTs with non-compliance using a simulation study. They argue that an unbiased intention-to-treat (ITT) estimate under non-compliance on a pnRCT may be obtained using a Complier Average Causal Effects (CACE) model. This method involves estimating the treatment effect for compliers and scaling this CACE effect estimate by the proportion of compliers to provide an ITT effect estimate. The issues posed by non-compliance warrant further investigation, considering a broader range of scenarios and investigating the degrees of freedom corrections for valid statistical inferences.

The design and analysis of trials with clustering in one arm needs to take account of heterogeneity and ICC to have a sufficiently powered sample size and accurate intervention effect. We strongly recommend the reporting of ICCs in trials results papers. The framework developed for cRCTs is also broadly applicable in iRCTs with clustering, identifying three dimensions to consider when reporting an ICC: a description of the dataset (including characteristics of the outcome and the intervention); how the ICC was calculated; and the precision of the ICC [31]. This has the potential to improve the assumptions about ICCs in iRCTs, adhere to CONSORT reporting guidelines for RCTs of nonpharmacological interventions [32], and raise awareness of the need to account for clustering in both the sample size and analysis in iRCTs with clustering.

A wide variety of terminology are used in iRCTs with clustering in one arm, including partially nested, partially clustered, multi-level, and individually randomized group intervention. A more consistent use of terminology would reduce confusion, improve reporting and make finding relevant ICCs from previous trials easier. We suggest the terminology partially nested randomised trial (pnRCT) to describe an iRCT with clustering in one arm.

Limitations

All the mixed-effects models assume that the cluster level means follow a Normal distribution. This may not be a valid assumption, for example, when we have a small number of clusters.

In the simulations, we have used fixed cluster sizes. In practice, cluster size may vary, causing a loss in efficiency when estimating the intervention effect. A simulation study by Candel and Van Breukelen [10] found the efficiency loss in the intervention effect estimate was rarely more than 10%, requiring recruitment of 11% more

clusters for the intervention arm and 11% more individuals for the control arm. The loss of efficiency in the intercept variance reached to 15%, requiring 19% more clusters in the clustered arm, and no additional recruitment in the control arm. Additionally, it has been shown in cluster trials if the coefficient of variation in cluster size is small, less than 0.23, then the correction on sample size is negligible [33]. It should be noted that cluster sizes are likely to be more similar in group administered interventions compared to trials which impose clustering by being treated by the same care provider [7].

Throughout the simulations we assumed there was no effect of clustering in the control arm, this may not strictly be true in practice. In healthcare intervention trials, a commonly used control intervention is 'care as usual'. This type of control may induce some form of low-level clustering, for instance, treatment by a healthcare practitioner. If the same practitioner treats numerous individuals, we can assume, in the same sense as we have done for the intervention arm that these individuals are clustered and include this in the modelling procedure. However, this information is often not available in trial data and is not unique to pnRCTs.

Partially nested trials pose a number of challenges, in particular, the issue of internal validity [6]. The grouping of individuals as part of the delivery of a treatment may affect the outcome. However, taking a pragmatic viewpoint, we consider the grouping as part of the treatment as a whole if this is reflective of treatment delivery in real-world practice. In addition, if the ungrouped controls are the true comparison in real life a pnRCT design will provide external validity.

Conclusion

Partially nested RCTs are increasingly used in complex intervention research. Ignoring clustering can lead to inflations of the Type I error rates, even for small ICCs. When analysing a pnRCT with continuous outcomes we recommend the use of a heteroscedastic partially nested mixed-effects model with corrected degrees of freedoms such as using the Satterhwaite method, for outcomes similar to those generated under the scenarios of our simulations study. The method used for classifying the non-clustered controls had a negligible impact on the results using the partially nested mixed-effects model. The model is easy to implement in standard statistical software and does not cause a notable reduction in power under homoscedastic variances. With few clusters, small cluster sizes and small ICC, the partially nested model underestimated Type I error rates and gave largely inflated ICC estimates, hence, for such designs there is no optimal model and we need to be cautious in model interpretation. Finally, to aid the design and prior selection of an appropriate analysis plan for pnRCTs, we strongly recommend the reporting of estimated ICC when publishing trials results.

Additional files

Additional file 1: Example Stata code used to run the simulations described in the manuscript text. (DOCX 16 kb)

Additional file 2: Additional details including: model fitting code for Stata, R, and SAS for the homoscedastic and heteroscedastic partially nested models; performance measures; and results tables. (DOCX 189 kb)

Abbreviations

cRCT: Cluster randomised controlled trials; ICC: Intracluster correlation coefficient; iRCT: Individually randomised controlled trials; pnRCT: Partially nested randomised controlled trials

Funding

JC was funded by the University of Sheffield Harry Worthington PhD Scholarship. MDT, SJW, MD, and LM were funded by the University of Sheffield. This is a summary of independent research funded by the National Institute for Health Research (NIHR)'s Doctoral Research Fellowship funding LF (DRF-2015-08-013). The views expressed are those of the author and not necessarily those of the NHS, the NIHR, the Department of Health or the University of Sheffield.

Availability of data and materials

All data used was simulated, simulations code is available in Additional file 1.

Authors' contributions

JC designed and implemented the simulations and drafted and edited the manuscript. MD, LF, and LM provided assistance and consultation in running the simulations. MDT, MD, LF, LM, and SJW provided input into the simulation conception and design and revised the manuscript critically. All authors approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 15 February 2018 Accepted: 18 September 2018
Published online: 11 October 2018

References

1. Campbell MJ, Walters SJ. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Chichester: Statistics in Practice, Wiley; 2014.
2. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345:e5661.
3. Hutchings J, Gardner F, Bywater T, Daley D, Whitaker C, Jones K, et al. Parenting intervention in sure start services for children at risk of developing conduct disorder: pragmatic randomised controlled trial. BMJ. 2007;334(7595):678.
4. Morrell CJ, Walters SJ, Dixon S, Collins KA, Brereton LM, Peters J, et al. Cost effectiveness of community leg ulcer clinics: randomised controlled trial. BMJ. 1998;316(7143):1487–91.
5. Diener MK, Seiler CM, von Frankenberg M, Rendel K, Schüle S, Maschuw K, et al. Vascular clips versus ligatures in thyroid surgery—results of a multicenter randomized controlled trial (CLIVIT trial). Langenbeck's Arch Surg. 2012;397(7):1117–26.

6. Bauer DJ, Sterba SK, Hallfors DD. Evaluating group-based interventions when control participants are ungrouped. Multivariate Behav Res. 2008; 43(2):210–36.
7. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. Clin Trials. 2005;2(2):152–62.
8. Gossage-Worrall R, Holt RIG, Barnard K, Carey ME, Davies MJ, Dickens C, et al. STEPWISE – STructured lifestyle education for people WIth SchizophrEnia: a study protocol for a randomised controlled trial. Trails. 2016;17(1):475.
9. Flight L, Allison A, Dimairo M, Lee E, Mandefield L, Walters SJ. Recommendations for the analysis of individually randomised controlled trials with clustering in one arm – a case of continuous outcomes. BMC Med Res Methodol. 2016;16(1):165.
10. Candel MJJM, Van Breukelen GJP. Varying cluster sizes in trials with clusters in one treatment arm: sample size adjustments when testing treatment effects with linear mixed models. Stat Med. 2009;28(18):2307–24.
11. Fazzari MJ, Kim MY, Heo M. Sample size determination for three-level randomized clinical trials with randomization at the first or second level. J Biopharm Stat. 2014;24(3):579–99.
12. Batistatou EE, Roberts C, Roberts S. Sample size and power calculations for trials and quasi-experimental studies with clustering. Stata J. 2014;14(1):159–75.
13. Moerbeek M, Wong WK. Sample size formulae for trials comparing group and individual treatments in a multilevel model. Stat Med. 2008;27(15):2850–64.
14. Moerbeek M, Teerenstra S. Power analysis of trials with multilevel data. Boca Raton: Chapman & Hall; 2016.
15. Baldwin SA, Bauer DJ, Stice E, Rohde P. Evaluating models for partially clustered designs. Psychol Methods. 2011;16(2):149–65.
16. Schweig JD, Pane JF. Intention-to-treat analysis in partially nested randomized controlled trials with real-world complexity. Int J Res Method Educ. 2016;39(3):268–86.
17. Sanders E. Multilevel Analysis Methods for Partially Nested Cluster Randomized Trials. University of Washington; 2011. Available online at: https://eric.ed.gov/?id=ED529306. Accessed 04 Sep 2018.
18. Korendijk EJH. Robustness and optimal design issues for cluster randomized trials. Utrecht University; 2012. Available online at: https://dspace.library.uu.nl/handle/1874/240965. Accessed 04 Sep 2018.
19. Lohr S, Schochet PZ, Sanders E. Partially Nested Randomized Controlled Trials in Education Research: A Guide to Design and Analysis. National Center for Education Research (NCER). Washington, DC; 2014. Available online at: https://ies.ed.gov/ncer/pubs/20142000/pdf/20142000.pdf. Accessed 04 Sep 2018.
20. Bland M. Grouping in individually randomised trials. In: 4th Annual Conference on Randomised Controlled Trials in the Social Sciences. York; 2009. Available online at: https://www-users.york.ac.uk/~mb55/talks/individ.pdf. Accessed 4 Sept 2018.
21. Kahan BC, Forbes G, Ali Y, Jairath V, Bremner S, Harhay MO, et al. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. Trials. 2016;17(1):438.
22. StataCorp. Stata Statistical Software: Release 14. College Station: StataCorp LP; 2015. p. 2015.
23. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006;25(24):4279–92.
24. Wickham H. ggplot2: elegant graphics for data analysis: New York: Springer-Verlag; 2016.
25. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017. http://www.R-project.org/
26. Baldwin SA, Murray DM, Shadish WR, Pals SL, Holland JM, Abramowitz JS, et al. Intraclass correlation associated with therapists: estimates and applications in planning psychotherapy research. Cogn Behav Ther. 2011;40(1):15–33.
27. Sterba SK. Partially nested designs in psychotherapy trials: a review of modeling developments. Psychother Res. 2017;27(4):425–36.
28. Lee KJ, Thompson SG. The use of random effects models to allow for clustering in individually randomized trials. Clin Trials. 2005;2(2):163–73.
29. Cook JA, Bruckner T, Maclennan GS, Seiler CM. Clustering in surgical trials : database of intra-cluster correlations. Trials. 2012;13:2.
30. Maas CJM, Hox JJ. Robustness issues in multilevel regression analysis. Stat Neerl. 2004;58(2):127–37.
31. Campbell MK, Grimshaw JM, Elbourne DR, Pocock S, Campbell M, Grimshaw J, et al. Intracluster correlation coefficients in cluster randomized trials: empirical insights into how should they be reported. BMC Med Res Methodol. 2004;4:9.
32. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P, Grp C. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Intern Med. 2008;148(4):295–309.
33. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. Int J Epidemiol. 2006;35(5):1292–300.
34. Hoover DR. Clinical trials of behavioural interventions with heterogeneous teaching subgroup effects. Stat Med. 2002;21(10):1351–64.

Appendix B

Within-arm partially nested randomised trials (chapter 5)

B.1 Supplementary results, figures and tables

B.1.1 Linear regression models (OLS, cluster robust and cluster bootstrap)

Table B.1: Type 1 error mean and standard deviation by ρ

	Model: linear regression with					
ICC	**OLS SEs**		**Cluster robust SEs**		**Cluster bootstrap SEs**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
0.01	0.051	0.005	0.006	0.006	0.076	0.010
0.05	0.057	0.008	0.006	0.008	0.080	0.009
0.1	0.064	0.009	0.007	0.007	0.080	0.009
0.2	0.077	0.021	0.009	0.009	0.080	0.008

Table B.2: Coverage rates of 95% confidence intervals mean and standard deviation by ρ

	Model: linear regression with					
ICC	**OLS SEs**		**Cluster robust SEs**		**Cluster bootstrap SEs**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
0.01	0.951	0.007	0.989	0.009	0.935	0.009
0.05	0.947	0.005	0.988	0.009	0.936	0.006
0.1	0.937	0.010	0.985	0.011	0.935	0.010
0.2	0.920	0.019	0.981	0.010	0.931	0.009

B.2 Stata simulation code for within-arm partially nested trials analysis

The following Stata code was used for the simulation study in chapter 5.

```
/* Simulation Parameters                                          */
       *Simulation for Yij = alpha + theta1*tij + theta2*t2ij + ujt2ij + eij
       *normally distributed continuous outcome with cluster (j) and
       *residual (i) variability

/* Function parameters                          */
       *theta1*: stage one intervention effect
       *theta2*: stage two intervention effect
       *rho*: ICC in clustered stage two
       *clusterSize*: Size of each cluster
       *nClusters*: Number of clusters in stage two
       *nIterations*: number of iterations for simulation

/* Variable Specification                                 */
       *y*: is the dependent variable;
       *cluster* is one of the following clustering options
       *cluster1* is the cluster indicator and treats the unclustered as clusters of
           size one

cap log close
drop _all
clear
set more off
capture log

/* Value of parameters: args assigns the first command line argument to the local
    macro simula_No, the second argument to nIterations and so on. */

args simulNo nIterations theta1 theta2 sigmau sigmae rho clusterSize nClusters

/* Set working directory */
cd "working directory"

local initial_seed = c(seed)

/* Log file for debugging and working out where mistakes are*/
log using log\simulation‘simulNo’.smcl, smcl replace

tempname memhold
tempfile results

qui postfile ‘memhold’ simICC theta1 theta2 sigmau sigmar sigmae rho clusterSize
    nClusters Model TreatEst SE p_value NormalCI_ll NormalCI_ul NormalCI_cv BCCI_ll
    BCCI_ul BCCI_cv ind_var converge using ‘results’, replace

forvalues sim = 1/‘nIterations’{
 * monitor iterative process of simulations by displaying a dot after every 100
     simulations
 if int(‘sim’/100) ==‘sim’/100{
              di as text "." _cont
              }
quietly{
```

```
local n2 = round((‘nClusters’*‘clusterSize’)/.5,2)*2
set obs ‘n2’
di ‘n2’
local n2 = _N

local sigmar = 1-‘sigmau’

* Generate treatment indicator
gen treat1 = 0
replace treat1 = 1 if _n< = _N/2

* Generate unique patient ID
gen patID = _n

* generate individual residuals control arm
gen e = rnormal(0,sqrt(‘sigmae’))

* generate individual residuals int arm
gen r = rnormal(0,sqrt(‘sigmar’))

* generate outcome after follow-up 1
gen y1 = ‘theta1’*treat1 + r*treat1 + e*(1-treat1)

* generate treat2
gen treat2 = 0
replace treat2 = 1 if y1<‘theta1’ & treat1 ==1

* generate cluster ID for those who get treat2 and singleton cluster ID for others
bysort treat2: egen cluster1 = seq(), b(‘clusterSize’)
summarize cluster1 if treat2 == 1
scal nClusters_treat2 = r(max)

*Number of clusters who receive intervention treat2-
replace cluster1 = (_n + nClusters_treat2) if (treat2==0)

* Generate cluster residual
sort cluster1
by cluster1: gen _rcl = rnormal(0,sqrt(‘sigmau’)) if _n==1
by cluster1: egen u = max(_rcl)

* generate follow-up 2
gen y2 = y1 + ‘theta2’*treat2 + u*treat2

* Calculate ICC of simulated data
tabstat u, by(treat2) stat(v) save
mat define I1 = r(Stat2)
local ClustRes_var I1[1,1]
tabstat r, by(treat2) stat(v) save
mat define I2 = r(Stat2)
local ClustInd_var I2[1,1]
local simICC =‘ClustRes_var’/(‘ClustRes_var’+‘ClustInd_var’)

*--------------------------
/*     FIT MODELS     */
*--------------------------
/* Notes on standard errors in regress: vce(vcetype) specifies the type of standard
    error reported, which includes types that are derived from asymptotic theory (ols)
    , that are robust to some kinds of misspecification (robust), that allow for
    intragroup correlation (cluster clustvar), and that use bootstrap or jackknife
```

```
    methods (bootstrap, jackknife); see [R] vce option */

* Model: Simple linear regression model ignoring clustering for follow-up 1 outcome
       capture {
              regress y2 treat1
              *define model number as first column
              if _rc == 0 local converge = 1
                     local m 1.0
                  * retrieve the contents of the outputs (estimates)
                     mat define M10 = r(table)
                     local b M10[1,1] //estimate of the treatment effect
                     local se M10[2,1] //standard error of the treatment effect
                     local p M10[4,1] //p value
                     local ll M10[5,1] //ll 95%CI
                     local ul M10[6,1] //ul 95%CI
                     local cv inrange(‘theta1’,‘ll’,‘ul’) //coverage indicator
                     local ind_var = e(rmse)
                     post ‘memhold’ (‘simICC’) (‘theta1’) (‘theta2’) (‘sigmau’) (‘
                         sigmar’) (‘sigmae’) (‘rho’) (‘clusterSize’) (‘nClusters’) (‘m
                         ’) (‘b’) (‘se’) (‘p’) (‘ll’) (‘ul’) (‘cv’) (.) (.) (.) (‘
                         ind_var’) (‘converge’)
              }

              if _rc != 0 {
                     local converge = 0
                     local m 1.0
                     post ‘memhold’ (‘simICC’) (‘theta1’) (‘theta2’) (‘sigmau’) (‘
                         sigmar’) (‘sigmae’) (‘rho’) (‘clusterSize’) (‘nClusters’) (‘m
                         ’) (.) (.) (.) (.) (.) (.) (.) (.) (.) (.) (‘converge’)
                     }


* Model: Simple linear regression model with cluster robust Std errors
* cluster() indicates that observations are clustered into groups and that
    observations may be
* correlated within these clusters, but independent between the clusters. By including
     the cluster
* option we also imply the robust option. The standard errors now take account that
    observations within
* clusters are correlated. Using this procedure, clusters are bootstrapped and the
    resampled clusters
* are kept intact.
       capture {
              regress y2 treat1, cluster(cluster1)
              *define model number as first column
              if _rc == 0 local converge = 1
                     local m 2
                  * retrieve the contents of the outputs (estimates)
                     mat define M21 = r(table)
                     local b M21[1,1] //estimate of the treatment effect
                     local se M21[2,1] //standard error of the treatment effect
                     local p M21[4,1] //p value
                     local ll M21[5,1] //ll 95%CI
                     local ul M21[6,1] //ul 95%CI
                     local cv inrange(‘theta1’,‘ll’,‘ul’) //coverage indicator
                     local ind_var = e(rmse)
                     post ‘memhold’ (‘simICC’) (‘theta1’) (‘theta2’) (‘sigmau’) (‘
                         sigmar’) (‘sigmae’) (‘rho’) (‘clusterSize’) (‘nClusters’) (‘m
                         ’) (‘b’) (‘se’) (‘p’) (‘ll’) (‘ul’) (‘cv’) (.) (.) (.) (‘
                         ind_var’) (‘converge’)
                     }
```

```
            if _rc != 0 {
                   local converge = 0
                   local m 2
                   post ‘memhold’ (‘simICC’) (‘theta1’) (‘theta2’) (‘sigmau’) (‘
                       sigmar’) (‘sigmae’) (‘rho’) (‘clusterSize’) (‘nClusters’) (‘m
                       ’) (.) (.) (.) (.) (.) (.) (.) (.) (.) (.) (‘converge’)
                   }

* Model: boostrap SEs, cluster1, strata treat2
       capture {
       bootstrap _b[treat1], cluster(cluster1) strata(treat2) reps(1000): regress y2
           treat1
               *define model number as first column
               if _rc == 0 local converge = 1
                   local m 3
                   mat define M3 = r(table)
                   local b M3[1,1] //estimate of the treatment effect
                   local se M3[2,1] //standard error of the treatment effect
                   local p M3[4,1] //p value
                   estat bootstrap, all
                   mat define N3 = e(ci_normal)
                   local n_ll N3[1,1] //normal ll 95%CI
                   local n_ul N3[2,1] //normal ul 95%CI
                   mat define B3 = e(ci_bc)
                   local bc_ll B3[1,1] //bias-corrected confidence interval ll 95%CI
                   local bc_ul B3[2,1] //bias-corrected confidence interval ul 95%CI
                   local n_cv inrange(‘theta1’,‘n_ll’,‘n_ul’) //coverage indicator
                   local bc_cv inrange(‘theta1’,‘bc_ll’,‘bc_ul’) //coverage
                       indicator
                   mat define V3 =e(V)
                   local ind_var = V3[1,1]
                   di ‘b’ ‘se’ ‘p’
                   post ‘memhold’ (‘simICC’) (‘theta1’) (‘theta2’) (‘sigmau’) (‘
                       sigmar’) (‘sigmae’) (‘rho’) (‘clusterSize’) (‘nClusters’) (‘m
                       ’) (‘b’) (‘se’) (‘p’) (‘n_ll’) (‘n_ul’) (‘n_cv’) (‘bc_ll’) (‘
                       bc_ul’) (‘bc_cv’) (‘ind_var’) (‘converge’)
                   }

               if _rc != 0 {
                   local converge = 0
                   local m 3
                   post ‘memhold’ (‘simICC’) (‘theta1’) (‘theta2’) (‘sigmau’) (‘
                       sigmar’) (‘sigmae’) (‘rho’) (‘clusterSize’) (‘nClusters’) (‘m
                       ’) (.) (.) (.) (.) (.) (.) (.) (.) (.) (.) (‘converge’)
                   }
}
clear
}

qui postclose ‘memhold’

qui use ‘results’, clear

/*
After all all simulations are run and the N-obsrvation
dataset of results is created, the code ends with
*/
note: File results‘simul_No’
note: Ran simluation ‘0’
note: Seed was ‘initial_seed’
```

```
save dta\results‘simulNo’, replace

/* Closed and convert log to html           */
log close
translate log\simulation‘simulNo’.smcl log\simulation‘simul_No’.log, replace
```

Appendix C

Sample size methods for partially nested trials (chapter 6)

C.1 List of included papers

Turner, E. L., Li, F., Gallis, J. A., Prague, M., and Murray, D. M. (2017). Review of recent methodological developments in group-randomized trials: part 1-design. American journal of public health, 107(6), 907-915.

Candel, M. J., and Van Breukelen, G. J. (2015). Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. Statistical methods in medical research, 24(5), 557-573.

Moerbeek, M. and Teerenstra, S. (2015). Power analysis of trials with multilevel data. Chapman and Hall/CRC.

Batistatou, E., Roberts, C., and Roberts, S. (2014). Sample size and power calculations for trials and quasi-experimental studies with clustering. The Stata Journal, 1, 159-75.

Heo, M., Litwin, A. H., Blackstock, O., Kim, N., and Arnsten, J. H. (2017). Sample size determinations for group-based randomized clinical trials with different levels of data hierarchy between experimental and control arms. Statistical methods in medical research, 26(1), 399-413.

Lohr, S., Schochet, P. Z., and Sanders, E. (2014). Partially Nested Randomized Controlled Trials in Education Research: A Guide to Design and Analysis. NCER 2014-2000. National Center for Education Research.

Roberts, C., and Walwyn, R. (2013). Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. Statistics in medicine, 32(1), 81-98.

Korendijk, E. (2012). Robustness and optimal design issues for cluster randomized trials (Doctoral dissertation, Utrecht University).

Baldwin, S. A., Murray, D. M., Shadish, W. R., Pals, S. L., Holland, J. M., Abramowitz, J. S., ... and Christensen, A. (2011). Intraclass correlation associated with therapists: estimates and applications in planning psychotherapy research. Cognitive Behaviour Therapy, 40(1), 15-33.

Baldwin, S. A., Bauer, D. J., Stice, E., and Rohde, P. (2011). Evaluating models for partially clustered designs. Psychological Methods, 16(2), 149.

Walwyn, R. E. (2010). Therapist variation within meta-analyses of psychotherapy trials. Manchester, UK: University of Manchester.

Candel, M. J., and Van Breukelen, G. J. (2009). Varying cluster sizes in trials with clusters in one treatment arm: sample size adjustments when testing treatment effects with linear mixed models. Statistics in medicine, 28(18), 2307-2324.

Moerbeek, M., and Wong, W. K. (2008). Sample size formulae for trials comparing group and individual treatments in a multilevel model. Statistics in Medicine, 27(15), 2850-2864.

Pals, S. L., Murray, D. M., Alfano, C. M., Shadish, W. R., Hannan, P. J., and Baker, W. L. (2008). Individually randomized group treatment trials: a critical appraisal of frequently used design and analytic approaches. American journal of public health, 98(8), 1418-1424.

Roberts, C., and Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. Clinical Trials, 2(2), 152-162.

Hoover, D. R. (2002). Clinical trials of behavioural interventions with heterogeneous teaching subgroup effects. Statistics in medicine, 21(10), 1351-1364.

C.2 Sample size comparison

Table C.1 provides a comparison of empirical sample size calculated using respective formulae from Heo et al. [137] and Moerbeek and Wong [131].

Table C.1: Comparison of pnRCT sample size methods from Heo et al. [137] and [131]. Assuming a standardised intervention effect $\delta_s = 0.5$ and homoscedastic variances $\gamma = 1$.

ICC	m	**Heo [137] method**				**Moerbeek [131] method**			
		k	mk	n_{pn}	**Total**	k	mk	n_{pn}	**Total**
0.01	10	7	70	70	140	7	70	65	135
0.05	10	8	80	80	160	9	90	70	160
0.1	10	9	90	90	180	11	110	75	185
0.01	5	13	65	65	130	14	70	64	134
0.05	5	14	70	70	140	15	75	66	141
0.1	5	15	75	75	150	17	85	69	154
0.2	10	12	120	120	240	15	150	84	234
0.4	10	17	170	170	340	22	220	99	319
0.6	10	22	220	220	440	29	290	111	401
0.2	5	17	85	85	170	20	100	74	174
0.4	5	21	105	105	210	27	135	83	218
0.6	5	24	120	120	240	34	170	90	260

Appendix D

Information on empirical ICCs and data extraction (chapter 7)

D.1 ICC estimates from other studies

Links to ICC estimates from other studies that may be used alongside those presented in this Appendix:

- Baldwin et al. [57] contact author for ICC estimates from psychotherapy trials (therapist ICCs) (on-line link from paper is no longer in use);
- Cook et al. [46] database of ICC estimates in surgery trials (centre and surgeon) available via downloadable excel spreadsheet from webpage `https://www.abdn.ac.uk/hsru/what-we-do/tools/index.php#panel177`;
- Stuart et al. [171] database of ICCs from primary care trials (GP centre ICCs), at the time of this thesis submission the paper was still under revisions .

D.2 Data extraction variables

The extraction variables included for the purpose of the HTA audit in chapter 7 are listed below. This is useful to show the data extracted but is not vital for the main body of the thesis.

Paper details:

- Database ID
- Data extraction date
- Auditor
- Publication Year, Volume, Issue
- Study
- Lead Author
- ISRCTN

Trial details:

- Trial type and design
- Number of arms
- Clinical area
- Setting
- Type of primary endpoint
- Primary endpoint
- Primary follow-up
- Control type
- Geographical region
- Recruitment centre type
- Recruitment total

Treatment induced clustering:

- Treatment clustering
- Number clustered arms
- Clustering intervention

- Clustering control
- Multi-centre
- Number centres
- Treatment induced clustering recognised
- Centre clustering recognised
- Sample size accounted for clustering?
- Sample size ICC treatment induced
- Sample size ICC centre
- Is sample size inflated in all arms?
- Evidence source for ICC
- Information on number of treatment clusters overall (if not reported per arm)
- Information on number of treatment clusters per arm (average, SD, range)
- Free text column regarding notes on clustering analysis and ICC

CONSORT adherence relating to items:

- CONSORT-NPT 4a, 7a, 12a, 13a (participant flow and diagram recorded separately), 15
- CONSORT-cluster 17a
- Free text column regarding whether and for what item it could be used for an exemplar.

D.3 Empirical ICC estimates

Table D.1 shows ICC values from the HTA iRCT identified in chapter 7. An extended table including further information such as sample size, cluster information and status of outcome and follow-up (primary/secondary) has been collected into an Excel sheet It was not possible to include all this in the Appendix. I aim to include the excel file as an online supplementary file to be submitted alongside a publication based on findings from chapter 7. For the purpose of this thesis submission it is presently stored using ORDA - The University of Sheffield Research

Data Catalogue and Repository which is hosted on Figshare via `https://figshare.com/s/d2645eb91b3ea9b65e0d`.

Table D.1: ICC values from HTA studies

ID	Trial	Outcome with follow-up	Cluster	ICC	95% CI		Clinical area
1	AESOPS	Average drinks per day (ADD) 12m	GP Practice	0.0570	NR	NR	Primary care
3	Body Therapy for Schizophrenia	Calgary 10w	Physio	0.0010	NR	NR	Mental health
3	Body Therapy for Schizophrenia	MANSA 10w	Physio	0.0010	NR	NR	Mental health
3	Body Therapy for Schizophrenia	PANSS Positive 6m	Physio	0.0010	NR	NR	Mental health
3	Body Therapy for Schizophrenia	PANSS Positive 10w	Physio	0.0010	NR	NR	Mental health
3	Body Therapy for Schizophrenia	SAS 10w	Physio	0.0010	NR	NR	Mental health
3	Body Therapy for Schizophrenia	SIX 6m	Physio	0.0010	NR	NR	Mental health
3	Body Therapy for Schizophrenia	SIX 10w	Physio	0.0010	NR	NR	Mental health
3	Body Therapy for Schizophrenia	SAS 6m	Physio	0.0070	NR	NR	Mental health
3	Body Therapy for Schizophrenia	CAINS Expression 10w	Physio	0.0220	NR	NR	Mental health
3	Body Therapy for Schizophrenia	CAINS Expression 6m	Physio	0.0230	NR	NR	Mental health
3	Body Therapy for Schizophrenia	CAINS Experience 10w	Physio	0.0370	NR	NR	Mental health
3	Body Therapy for Schizophrenia	CAINS Experience 6m	Physio	0.0410	NR	NR	Mental health
3	Body Therapy for Schizophrenia	MANSA 6m	Physio	0.0500	NR	NR	Mental health
3	Body Therapy for Schizophrenia	PANSS Marder 6m	Physio	0.0750	NR	NR	Mental health
3	Body Therapy for Schizophrenia	Calgary 6m	Physio	0.0860	NR	NR	Mental health
3	Body Therapy for Schizophrenia	PANSS General 10w	Physio	0.0960	NR	NR	Mental health
3	Body Therapy for Schizophrenia	PANSS Negative 10w	Physio	0.0990	NR	NR	Mental health
3	Body Therapy for Schizophrenia	PANSS Negative 6m	Physio	0.1370	NR	NR	Mental health
3	Body Therapy for Schizophrenia	PANSS General 6m	Physio	0.2050	NR	NR	Mental health
3	Body Therapy for Schizophrenia	PANSS Marder 10w	Physio	0.6780	NR	NR	Mental health
6	CASPER	PHQ-9 4 m	Case manager	0.0069	0.0000	0.0644	Mental health
6	CASPER	PHQ-9 12m	Case manager	0.0072	0.0000	0.0676	Mental health
7	CASPER Plus	PHQ-9 4 m	Case manager	0.0001	0.0000	0.0644	Mental health
10	COBRA	PHQ-9 12m	NR	0.0400	NR	NR	Mental health
12	Families for Health	BMI z-score 12m	Family-level	0.4713	0.3170	0.0396	Obesity
12	Families for Health	BMI z-score 12m	Families for Health delivery group	0.0000	NR	NR	Obesity
13	Getting out of the house	NEADL 12m	Occupational therapist/Physio	0.0012	NR	NR	Stroke
13	Getting out of the house	SF36-v2 12m	Occupational therapist/Physio	0.0025	NR	NR	Stroke
13	Getting out of the house	GHQ-12 (participant) 6m	Occupational therapist/Physio	0.0028	NR	NR	Stroke
13	Getting out of the house	GHQ-12 (participant) 6m	NHS stroke services	0.0035	NR	NR	Stroke
13	Getting out of the house	GHQ-12 (participant) 12m	NHS stroke services	0.0043	NR	NR	Stroke
13	Getting out of the house	SF36-v2 6m	Occupational therapist/Physio	0.0051	NR	NR	Stroke
13	Getting out of the house	NEADL 6m	Occupational therapist/Physio	0.0052	NR	NR	Stroke
13	Getting out of the house	RMI 12m	NHS stroke services	0.0057	NR	NR	Stroke
13	Getting out of the house	RMI 6m	NHS stroke services	0.0060	NR	NR	Stroke
13	Getting out of the house	GHQ-12 (carer) 12m	NHS stroke services	0.0066	NR	NR	Stroke
13	Getting out of the house	GHQ-12 (participant) 12m	Occupational therapist/Physio	0.0075	NR	NR	Stroke
13	Getting out of the house	SWOM 12m	NHS stroke services	0.0093	NR	NR	Stroke
13	Getting out of the house	SF36-v2 6m	NHS stroke services	0.0099	NR	NR	Stroke
13	Getting out of the house	RMI 6m	Occupational therapist/Physio	0.0111	NR	NR	Stroke
13	Getting out of the house	SWOM 6m	NHS stroke services	0.0123	NR	NR	Stroke
13	Getting out of the house	RMI 12m	Occupational therapist/Physio	0.0134	NR	NR	Stroke
13	Getting out of the house	GHQ-12 (carer) 12m	Occupational therapist/Physio	0.0136	NR	NR	Stroke
13	Getting out of the house	NEADL 6m	NHS stroke services	0.0155	NR	NR	Stroke
13	Getting out of the house	NEADL 12m	NHS stroke services	0.0171	NR	NR	Stroke
13	Getting out of the house	GHQ-12 (carer) 6m	Occupational therapist/Physio	0.0176	NR	NR	Stroke
13	Getting out of the house	GHQ-12 (carer) 6m	NHS stroke services	0.0234	NR	NR	Stroke
13	Getting out of the house	SWOM 6m	Occupational therapist/Physio	0.0289	NR	NR	Stroke
13	Getting out of the house	SWOM 12m	Occupational therapist/Physio	0.0315	NR	NR	Stroke
13	Getting out of the house	SF36-v2 12m	NHS stroke services	0.0597	NR	NR	Stroke
14	IMPACT	MFQ centred time variable in model	Orthopaedic surgeon	0.0001	NR	NR	Mental health
15	KATMETAL	EQ-5D 12m	Orthopaedic surgeon	0.0000	0.0000	0.0250	O/R/MSK
15	KATMETAL	EQ-5D 12m	NHS Hospital	0.0060	0.0000	0.0700	O/R/MSK
15	KATMETAL	EQ-5D 3m	NHS Hospital	0.0000	0.0000	0.000*	O/R/MSK
15	KATMETAL	EQ-5D 3m	Orthopaedic surgeon	0.0140	0.0000	0.1040	O/R/MSK
15	KATMETAL	EQ-5D 60m	Orthopaedic surgeon	0.0000	0.0000	0.000*	O/R/MSK
15	KATMETAL	EQ-5D 60m	NHS Hospital	0.0000	0.0000	0.000*	O/R/MSK
15	KATMETAL	Operating time (min)	NHS Hospital	0.4490	0.1720	0.7260	O/R/MSK
15	KATMETAL	Operating time (min)	Orthopaedic surgeon	0.5140	0.2780	0.7270	O/R/MSK
15	KATMETAL	Oxford knee score 12m	NHS Hospital	0.0210	0.0000	0.0420	O/R/MSK
15	KATMETAL	Oxford knee score 12m	Orthopaedic surgeon	0.0560	0.0000	0.1300	O/R/MSK
15	KATMETAL	Oxford knee score 3m	NHS Hospital	0.0000	0.0000	0.000*	O/R/MSK
15	KATMETAL	Oxford knee score 3m	Orthopaedic surgeon	0.0070	0.0000	0.0570	O/R/MSK
15	KATMETAL	Oxford knee score 60m	NHS Hospital	0.0000	0.0000	0.000*	O/R/MSK
15	KATMETAL	Oxford knee score 60m	Orthopaedic surgeon	0.0020	0.0000	0.0210	O/R/MSK

ID	Trial	Outcome with follow-up	Cluster	ICC	95% CI		Clinical area
15	KATMOBILE	EQ-5D 60m	NHS Hospital	0.0150	0.0000	0.0800	O/R/MSK
15	KATMOBILE	EQ-5D 12m	Orthopaedic surgeon	0.0240	0.0000	0.0800	O/R/MSK
15	KATMOBILE	EQ-5D 3m	Orthopaedic surgeon	0.0710	0.0210	0.1490	O/R/MSK
15	KATMOBILE	EQ-5D 12m	NHS Hospital	0.0400	0.0040	0.1030	O/R/MSK
15	KATMOBILE	Oxford knee score 60m	NHS Hospital	0.0440	0.0000	0.1230	O/R/MSK
15	KATMOBILE	EQ-5D 60m	Orthopaedic surgeon	0.0190	0.0000	0.0650	O/R/MSK
15	KATMOBILE	Operating time (min)	Orthopaedic surgeon	0.1990	0.0730	0.3900	O/R/MSK
15	KATMOBILE	EQ-5D 3m	NHS Hospital	0.0600	0.0050	0.1330	O/R/MSK
15	KATMOBILE	Oxford knee score 12m	NHS Hospital	0.0630	0.0230	0.1340	O/R/MSK
15	KATMOBILE	Oxford knee score 12m	Orthopaedic surgeon	0.0590	0.0120	0.1340	O/R/MSK
15	KATMOBILE	Oxford knee score 3m	Orthopaedic surgeon	0.0680	0.0180	0.1300	O/R/MSK
15	KATMOBILE	Oxford knee score 3m	NHS Hospital	0.0730	0.0160	0.1580	O/R/MSK
15	KATMOBILE	Operating time (min)	NHS Hospital	0.1670	0.0920	0.2940	O/R/MSK
15	KATMOBILE	Oxford knee score 60m	Orthopaedic surgeon	0.0510	0.0000	0.1210	O/R/MSK
15	KATPATELLA	EQ-5D 60m	NHS Hospital	0.0020	0.0000	0.0210	O/R/MSK
15	KATPATELLA	EQ-5D 12m	Orthopaedic surgeon	0.0400	0.0110	0.0760	O/R/MSK
15	KATPATELLA	EQ-5D 3m	NHS Hospital	0.0080	0.0000	0.0410	O/R/MSK
15	KATPATELLA	EQ-5D 3m	Orthopaedic surgeon	0.0050	0.0000	0.0320	O/R/MSK
15	KATPATELLA	EQ-5D 12m	NHS Hospital	0.0170	0.0000	0.0520	O/R/MSK
15	KATPATELLA	Oxford knee score 12m	NHS Hospital	0.0270	0.0040	0.0710	O/R/MSK
15	KATPATELLA	EQ-5D 60m	Orthopaedic surgeon	0.0090	0.0000	0.0470	O/R/MSK
15	KATPATELLA	Operating time (min)	Orthopaedic surgeon	0.4450	0.3600	0.5240	O/R/MSK
15	KATPATELLA	Oxford knee score 3m	NHS Hospital	0.0410	0.0140	0.0870	O/R/MSK
15	KATPATELLA	Oxford knee score 60m	NHS Hospital	0.0450	0.0160	0.0860	O/R/MSK
15	KATPATELLA	Oxford knee score 12m	Orthopaedic surgeon	0.0470	0.0150	0.0940	O/R/MSK
15	KATPATELLA	Oxford knee score 3m	Orthopaedic surgeon	0.0500	0.0210	0.0870	O/R/MSK
15	KATPATELLA	Operating time (min)	NHS Hospital	0.3700	0.2540	0.4700	O/R/MSK
15	KATPATELLA	Oxford knee score 60m	Orthopaedic surgeon	0.0370	0.0030	0.0740	O/R/MSK
16	OCTET Guided self-help	Y-BOCS-OR 12m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	Y-BOCS-OR 3m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	Y-BOCS-OR 6m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	CORE-OM 12m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	CORE-OM 6m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	CSQ-8 3m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	GAD-7 6m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	PHQ-9 3m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	PHQ-9 6m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	SF-36 v2 MCS 3m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	WSAS 3m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	WSAS 6m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Guided self-help	SF-36 v2 PCS 12m	Psychological wellbeing practitioner	0.0020	NR	NR	Mental health
16	OCTET Guided self-help	CORE-OM 3m	Psychological wellbeing practitioner	0.0030	NR	NR	Mental health
16	OCTET Guided self-help	Y-BOCS-SR 12m	Psychological wellbeing practitioner	0.0040	NR	NR	Mental health
16	OCTET Guided self-help	SF-36 v2 PCS 6m	Psychological wellbeing practitioner	0.0040	NR	NR	Mental health
16	OCTET Guided self-help	SF-36 v2 MCS 12m	Psychological wellbeing practitioner	0.0070	NR	NR	Mental health
16	OCTET Guided self-help	Y-BOCS-SR 3m	Psychological wellbeing practitioner	0.0090	NR	NR	Mental health
16	OCTET Guided self-help	PHQ-9 12m	Psychological wellbeing practitioner	0.0110	NR	NR	Mental health
16	OCTET Guided self-help	WSAS 12m	Psychological wellbeing practitioner	0.0120	NR	NR	Mental health
16	OCTET Guided self-help	GAD-7 3m	Psychological wellbeing practitioner	0.0300	NR	NR	Mental health
16	OCTET Guided self-help	SF-36 v2 MCS 6m	Psychological wellbeing practitioner	0.0480	NR	NR	Mental health
16	OCTET Guided self-help	Y-BOCS-SR 6m	Psychological wellbeing practitioner	0.0780	NR	NR	Mental health
16	OCTET Guided self-help	GAD-7 12m	Psychological wellbeing practitioner	0.0820	NR	NR	Mental health
16	OCTET Guided self-help	CSQ-8 6m	Psychological wellbeing practitioner	0.0970	NR	NR	Mental health
16	OCTET Guided self-help	SF-36 v2 PCS 3m	Psychological wellbeing practitioner	0.1780	NR	NR	Mental health
16	OCTET Supported cCBT	Y-BOCS-OR 6m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Supported cCBT	Y-BOCS-SR 6m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Supported cCBT	CORE-OM 12m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Supported cCBT	CORE-OM 6m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Supported cCBT	CSQ-8 3m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Supported cCBT	GAD-7 6m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Supported cCBT	PHQ-9 3m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health

ID	Trial	Outcome with follow-up	Cluster	ICC	95% CI		Clinical area
16	OCTET Supported cCBT	WSAS 3m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Supported cCBT	WSAS 6m	Psychological wellbeing practitioner	0.0010	NR	NR	Mental health
16	OCTET Supported cCBT	Y-BOCS-SR 3m	Psychological wellbeing practitioner	0.0030	NR	NR	Mental health
16	OCTET Supported cCBT	SF-36 v2 PCS 12m	Psychological wellbeing practitioner	0.0110	NR	NR	Mental health
16	OCTET Supported cCBT	SF-36 v2 MCS 6m	Psychological wellbeing practitioner	0.0180	NR	NR	Mental health
16	OCTET Supported cCBT	PHQ-9 6m	Psychological wellbeing practitioner	0.0190	NR	NR	Mental health
16	OCTET Supported cCBT	CORE-OM 3m	Psychological wellbeing practitioner	0.0260	NR	NR	Mental health
16	OCTET Supported cCBT	Y-BOCS-OR 3m	Psychological wellbeing practitioner	0.0270	NR	NR	Mental health
16	OCTET Supported cCBT	GAD-7 3m	Psychological wellbeing practitioner	0.0280	NR	NR	Mental health
16	OCTET Supported cCBT	CSQ-8 6m	Psychological wellbeing practitioner	0.0410	NR	NR	Mental health
16	OCTET Supported cCBT	SF-36 v2 MCS 12m	Psychological wellbeing practitioner	0.0900	NR	NR	Mental health
16	OCTET Supported cCBT	Y-BOCS-SR 12m	Psychological wellbeing practitioner	0.0980	NR	NR	Mental health
16	OCTET Supported cCBT	Y-BOCS-OR 12m	Psychological wellbeing practitioner	0.1090	NR	NR	Mental health
16	OCTET Supported cCBT	GAD-7 12m	Psychological wellbeing practitioner	0.1510	NR	NR	Mental health
16	OCTET Supported cCBT	WSAS 12m	Psychological wellbeing practitioner	0.1580	NR	NR	Mental health
16	OCTET Supported cCBT	PHQ-9 12m	Psychological wellbeing practitioner	0.1600	NR	NR	Mental health
16	OCTET Supported cCBT	SF-36 v2 PCS 6m	Psychological wellbeing practitioner	0.1640	NR	NR	Mental health
16	OCTET Supported cCBT	SF-36 v2 PCS 3m	Psychological wellbeing practitioner	0.1820	NR	NR	Mental health
16	OCTET Supported cCBT	SF-36 v2 MCS 3m	Psychological wellbeing practitioner	0.2250	NR	NR	Mental health
18	PEPS	SPSI-R 72 w	Problem-solving therapy groupgroup	0.0100	0.0100	0.1700	Mental health
18	PEPS	Three main problems client's assessment of severity 72 w	Problem-solving therapy groupgroup	0.0100	0.0100	0.1500	Mental health
18	PEPS	SFQ 72 w	Problem-solving therapy groupgroup	0.0700	0.0000	0.2900	Mental health
18	PEPS	HADS-T 72 w	Problem-solving therapy groupgroup	0.1100	0.0000	0.2900	Mental health
19	PhysioDirect	SF-36 v2 PCS 6m	NHS Primary Care Trust	0.0050	NR	NR	O/R/MSK
19	PhysioDirect	SF-36 v2 PCS 6m	GP Practice	0.0050	NR	NR	O/R/MSK
20	POWER+	EQ5-D 12m	GP Practice	0.0140	0.0000	0.3600	Obesity
20	POWER+	Mean weight reduction 12m	GP Practice	0.0150	0.0000	0.0900	Obesity
20	POWER+	Clinically important weight reduction 12m	GP Practice	0.0180	0.0000	0.1600	Obesity
22	SARAH	MHQ overall hand function 12m	Hand therapist	0.0000	NR	NR	O/R/MSK
23	SHEAR	Form 90: Alcohol 6mconsumption in last 90 days	Clinician	0.0120	0.0001	0.6486	Mental health
24	START	HADS-T 12m	Psychology graduate	0.0000	0.0000	0.0700	Mental health
24	START	HADS-T 18 m	Psychology graduate	0.0000	0.0000	0.0700	Mental health
24	START	HADS-T 8 m	Psychology graduate	0.0000	0.0000	0.0800	Mental health
24	START	HADS-T 4 m	Psychology graduate	0.0200	0.0000	0.0900	Mental health
26	SWAP Group (intervention)	Weight 6m	Weight management group	0.0130	0.0000	0.1050	Obesity
26	SWAP Group (intervention)	Weight 12m	Weight management group	0.0010	0.0000	0.0760	Obesity
26	SWAP Group (intervention)	Waist 6m	Weight management group	0.1170	0.0000	0.2650	Obesity
26	SWAP Group (intervention)	Waist 12m	Weight management group	0.0800	0.0000	0.2030	Obesity
26	SWAP Group (intervention)	SBP 6m	Weight management group	0.0320	0.0000	0.1350	Obesity
26	SWAP Group (intervention)	SBP 12m	Weight management group	0.0200	0.0000	0.1080	Obesity
26	SWAP Group (intervention)	DBP 6m	Weight management group	0.1560	0.0000	0.3220	Obesity
26	SWAP Group (intervention)	DBP 12m	Weight management group	0.0570	0.0000	0.1690	Obesity
26	SWAP Group (intervention)	BMI 6m	Weight management group	0.0280	0.0000	0.1290	Obesity
26	SWAP Group (intervention)	BMI 12m	Weight management group	0.0010	0.0000	0.0760	Obesity
26	SWAP Group (intervention)	Lost 5% of body weight 6m	Weight management group	0.0010	0.0000	0.0840	Obesity
26	SWAP Group (intervention)	Lost 5% of body weight 12m	Weight management group	0.0010	0.0000	0.0760	Obesity
26	SWAP Group (intervention)	Lost 10% of body weight 6m	Weight management group	0.0010	0.0000	0.0840	Obesity
26	SWAP Group (intervention)	Lost 10% of body weight12m	Weight management group	0.0010	0.0000	0.0760	Obesity
26	SWAP Group (intervention)	Food knowledge assessment 6m	Weight management group	0.0010	0.0000	0.0840	Obesity
26	SWAP Group (intervention)	Food knowledge assessment 12m	Weight management group	0.0460	0.0000	0.1510	Obesity
26	SWAP Group (intervention)	Food craving index – frequency domain 6m	Weight management group	0.0390	0.0000	0.1460	Obesity
26	SWAP Group (intervention)	Food craving index – frequency domain 12m	Weight management group	0.0010	0.0000	0.0770	Obesity
26	SWAP Group (intervention)	Food craving index – strength domain 6m	Weight management group	0.0190	0.0000	0.1150	Obesity

ID	Trial	Outcome with follow-up	Cluster	ICC	95% CI		Clinical area
26	SWAP Group (intervention)	Food craving index – strength domain 12m	Weight management group	0.0010	0.0000	0.0780	Obesity
26	SWAP Group (intervention)	Three factor eating – cognitive restraint domain 6m	Weight management group	0.0010	0.0000	0.0850	Obesity
26	SWAP Group (intervention)	Three factor eating – cognitive restraint domain 12m	Weight management group	0.0010	0.0000	0.0780	Obesity
26	SWAP Group (intervention)	Three factor eating – uncontrolled eating domain 6m	Weight management group	0.0810	0.0000	0.2130	Obesity
26	SWAP Group (intervention)	Three factor eating – uncontrolled eating domain 12m	Weight management group	0.0350	0.0000	0.1350	Obesity
26	SWAP Group (intervention)	Three factor eating – emotional eating domain 6m	Weight management group	0.0650	0.0000	0.1870	Obesity
26	SWAP Group (intervention)	Three factor eating – emotional eating domain 12m	Weight management group	0.0630	0.0000	0.1790	Obesity
26	SWAP Group (intervention)	IPAQ – METS minutes/week domain 6m	Weight management group	0.0010	0.0000	0.1010	Obesity
26	SWAP Group (intervention)	IPAQ – METS minutes/week domain 12m	Weight management group	0.0500	0.0000	0.1730	Obesity
26	SWAP Group (intervention)	IPAQ – sitting domain domain 6m	Weight management group	0.0280	0.0000	0.1470	Obesity
26	SWAP Group (intervention)	IPAQ – sitting domain domain 12m	Weight management group	0.0010	0.0000	0.1020	Obesity
26	SWAP Nurse (control)	Weight 6m	Nurse	0.0010	0.0000	0.1010	Obesity
26	SWAP Nurse (control)	Weight 12m	Nurse	0.0010	0.0000	0.0860	Obesity
26	SWAP Nurse (control)	Waist 6m	Nurse	0.0010	0.0000	0.1010	Obesity
26	SWAP Nurse (control)	Waist 12m	Nurse	0.0010	0.0000	0.0860	Obesity
26	SWAP Nurse (control)	SBP 6m	Nurse	0.0010	0.0000	0.1010	Obesity
26	SWAP Nurse (control)	SBP 12m	Nurse	0.0550	0.0000	0.2220	Obesity
26	SWAP Nurse (control)	DBP 6m	Nurse	0.0010	0.0000	0.1010	Obesity
26	SWAP Nurse (control)	DBP 12m	Nurse	0.0010	0.0000	0.0860	Obesity
26	SWAP Nurse (control)	BMI 6m	Nurse	0.0010	0.0000	0.1010	Obesity
26	SWAP Nurse (control)	BMI 12m	Nurse	0.0010	0.0000	0.0860	Obesity
26	SWAP Nurse (control)	Lost 5% of body weight 6m	Nurse	0.0010	0.0000	0.1010	Obesity
26	SWAP Nurse (control)	Lost 5% of body weight 12m	Nurse	0.0140	0.0000	0.1210	Obesity
26	SWAP Nurse (control)	Lost 10% of body weight 6m	Nurse	0.0010	0.0000	0.1020	Obesity
26	SWAP Nurse (control)	Lost 10% of body weight12m	Nurse	0.0010	0.0000	0.0860	Obesity
26	SWAP Nurse (control)	Food knowledge assessment 6m	Nurse	0.0010	0.0000	0.1010	Obesity
26	SWAP Nurse (control)	Food knowledge assessment 12m	Nurse	0.0010	0.0000	0.0860	Obesity
26	SWAP Nurse (control)	Food craving index – frequency domain 6m	Nurse	0.0500	0.0000	0.2220	Obesity
26	SWAP Nurse (control)	Food craving index – frequency domain 12m	Nurse	0.0800	0.0000	0.2800	Obesity
26	SWAP Nurse (control)	Food craving index – strength domain 6m	Nurse	0.0930	0.0000	0.3190	Obesity
26	SWAP Nurse (control)	Food craving index – strength domain 12m	Nurse	0.0170	0.0000	0.1300	Obesity
26	SWAP Nurse (control)	Three factor eating – cognitive restraint domain 6m	Nurse	0.0160	0.0000	0.1430	Obesity
26	SWAP Nurse (control)	Three factor eating – cognitive restraint domain 12m	Nurse	0.0970	0.0000	0.3190	Obesity
26	SWAP Nurse (control)	Three factor eating – uncontrolled eating domain 6m	Nurse	0.2370	0.0000	0.6070	Obesity
26	SWAP Nurse (control)	Three factor eating – uncontrolled eating domain 12m	Nurse	0.1070	0.0000	0.3460	Obesity
26	SWAP Nurse (control)	Three factor eating – emotional eating domain 6m	Nurse	0.1150	0.0000	0.3680	Obesity
26	SWAP Nurse (control)	Three factor eating – emotional eating domain 12m	Nurse	0.1830	0.0000	0.5010	Obesity
26	SWAP Nurse (control)	IPAQ – METS minutes/week domain 6m	Nurse	0.0010	0.0000	0.1240	Obesity
26	SWAP Nurse (control)	IPAQ – METS minutes/week domain 12m	Nurse	0.0010	0.0000	0.1000	Obesity
26	SWAP Nurse (control)	IPAQ – sitting domain domain 6m	Nurse	0.0010	0.0000	0.1210	Obesity

*NR:Not reported; O/R/MSK: Orthopedics/ Rheumatology/ Musculoskeletal (including back pain)