Methods for overcoming barriers to using adaptive designs in Low and Middle-Income Countries

Samuel Kwakye Sarkodie

Thesis submitted for the degree of
Doctor of Philosophy

Population Health Sciences Institute
Faculty of Medical Sciences
Newcastle University
Newcastle upon Tyne
United Kingdom

January 2024

I dedicate this thesis to my parents, Mr. James Sarkodie and Mrs Georgina Otchere.

Acknowledgements

The successful completion of this phase in my academic career would have been unattainable without the divine strength and grace of the Almighty God, for which I am immensely grateful.

My appreciation extends to the Newcastle University Overseas Research Scholarship (NUORS) for the funding opportunity to undertake my PhD studies. This new academic milestone was made possible by their generous assistance.

To my supervisors, Prof James MS Wason and Dr Michael J Grayling, you have my utmost gratitude for all efforts channelled towards the outcome of my good. Your vast expertise, intelligent remarks, and constructive criticisms have significantly influenced my academic growth and elevated this thesis to its current state of excellence. Working with you these past years have been such a wonderful experience with me developing skills and knowledge which would impact my career. I also appreciate the appraisal and recommendations from my assessors, Prof John Matthews and Dr Michele Castelli for their advice, unwavering support, and guidance at every stage of my PhD study.

My appreciation also goes out to my parents, my siblings, and friends for their relentless support, prayers, love and counsel. Their consistent belief in me served as the compass of my academic pursuit. All these are invaluable assets that I cannot repay. I am also indebted to Dr & Mrs Duah Danso, Mr. Armando Martie, Mr. Edmund Asamoah, Dr Apreh-Siaw, Dr Gabriel Aboyadana, Dr Yaw Boakye-Ansah, Dr Shadrach Dare, Dr Kevin Wilson, Mr. & Mrs Adu-Amankwa, as well as Mr. & Mrs Osei for their support. Your encouragement has not only given me solace but also served as the impetus for my tenacity all through my PhD.

This acknowledgement would not be complete without honouring Matilda, Kimberly, Angela, and Dorathy for their diverse contributions to the success of this thesis. To the wonderful individuals I met during my studies – Ruqayya, Luke, Faye, Aritra, Pela, Laura, Lou, and the members of the BRG – your contributions have been invaluable in various aspects of my academic journey, and I express eternal gratitude for your support.

Abstract

Research indicates that populations in Low and Middle-Income Countries (LMICs) could benefit from clinical trials due to their high disease burden. Yet, they are underrepresented in clinical research. This is because the advancement of clinical research in LMICs is hindered by several obstacles, among which insufficient funding has been identified as a key barrier. One notable innovation, that holds promise for increasing the number of trials conducted in LMICs, is the utilisation of adaptive designs as they can increase the value derived from limited resources. Considering that adaptive designs are not always beneficial, the methods presented in this thesis determine instances where adaptations could provide greater utility within the context of individually (IRTs) and cluster randomised trials (CRTs) with continuous outcomes. The thesis begins by proposing a seamless MAMS framework where a cheaper (intermediate) endpoint is used at an interim analysis. With adaptive designs heavily reliant on the quality and quantity of available information at interim analyses for decision-making, this study proposes an optimal timing for conducting the interim analysis to avert wrongly dropping a relevant arm at the interim analysis. To offer a contextual foundation for the subsequent chapters on CRT methodologies, the thesis next presents the results of a comprehensive review undertaken to explore various strategies for specifying the intra-cluster correlation coefficient (ICC) and other essential sample size parameters. The review highlighted that many HTA trials neither reported the uncertainty around the assumed ICC nor justified their selected values, which could affect the trial's validity. Based on the review, I then evaluated how uncertainty in the ICC impacts whether a parallel-group or stepped-wedge CRT design is more efficient in terms of the required sample size. Here, the uncertainty was captured by placing independent priors on key parameters and averaging over the possible range of values. The results indicated that in many cases, when there is uncertainty regarding the ICC, a stepped-wedge CRT design tends to be more efficient compared to a parallel-group CRT design. A limitation of the above approach is that its utility can be highly dependent on the choice of prior. Thus, I next introduce an adaptive design where pre-trial knowledge about the ICC is captured by placing a prior on it, which is then updated at an interim analysis using the study data to form a posterior that allows reestimation of the sample size. It was clearly demonstrated in the results that when there is low-quality evidence available to guide the choice of prior, this approach to sample size reestimation provides greater utility than previously proposed frequentist methods. Results show the proposed methodologies are both robust and cost-effective. Therefore, I conclude by discussing how the adoption of these methods could enhance LMICs' capacities to conduct high-quality research.

Contents

List of Figures

List of Tables

Publications related to this thesis

Sarkodie SK, Wason JMS, Grayling MJ. A hybrid approach to comparing parallel-group and stepped-wedge cluster-randomized trials with a continuous primary outcome when there is uncertainty in the intra-cluster correlation. *Clinical Trials* 2022; 20(1): 59–70. doi: 10.1177/17407745221123507

Chapter 1

Introduction

This chapter introduces the fundamental concepts that form the focus of this thesis. It begins by discussing the concepts of individual and cluster-level randomisation in clinical trials, as well as adaptive designs (ADs) and their rationale. The associated methodological implications of these concepts are also explored. Additionally, the chapter provides an overview of clinical trials conducted in low and middle-income countries (LMICs), highlighting the opportunities and challenges they present. This background sets the stage for the innovative methods proposed in subsequent chapters, which aim to address particular barriers to executing (adaptive) clinical trials in LMICs.

1.1 Clinical trials

Clinical trials are research studies that evaluate the safety and efficacy of treatments and interventions (Friedman *et al.*, 2015; Crowley & Hoering, 2012). They play a vital role in advancing medical knowledge and improving healthcare outcomes. Since the introduction of randomised controlled trials (RCTs) in 1946 (Bhatt, 2010), many medical intervention programs, diagnostic tools, and drugs have been evaluated using RCTs. They continue to be the "gold" standard for evaluating current innovations in medical practice. Simultaneously, they also spur questions for new research that lead to further discoveries over time, helping us to, e.g., better understand diseases (Pocock, 2013; Friedman *et al.*, 2015; Haeussler & Assmus, 2021). Hence, their significant contribution to the health and well-being of the general population cannot be overemphasised.

Clinical trials typically belong to several distinct phases, each serving a specific purpose in the development and evaluation of a new intervention. These phases progressively move from assessing safety and dosage to evaluating efficacy and confirming previous findings, ultimately leading to regulatory approval and post-marketing monitoring. The commonly recognised phases as summarised by the American Cancer Society (2020) are:

- **Phase 0 (Exploratory)**: Phase 0 trials involve a small number of participants and are designed to gather preliminary data on how a new drug or treatment interacts with the body. These trials are exploratory and may involve micro-dosing or exposing participants to very low doses of the intervention.
- **Phase 1 (Safety)**: Phase 1 trials assess the safety and dosage range of the intervention. These trials usually involve a small number of healthy volunteers or individuals with the target condition. The focus is primarily on determining the treatment's safety, identifying potential side effects, and establishing appropriate dosage levels.
- **Phase 2 (Efficacy)**: Phase 2 trials aim to evaluate the efficacy of the intervention in treating the target condition. These trials involve a larger number of participants and are typically randomised controlled studies. The focus is on gathering preliminary evidence of the treatment's effectiveness, optimal dosages, and potential adverse effects.
- **Phase 3 (Confirmation)**: Phase 3 trials involve a larger number of participants and are designed to confirm the effectiveness of the intervention demonstrated in Phase 2. These trials often include a control group and employ rigorous randomisation and blinding techniques. Phase 3 trials provide more comprehensive data on the benefits, risks, and potential side effects of the treatment.
- **Phase 4 (Post-Marketing Surveillance)**: Phase 4 trials are conducted after the regulatory approval of the intervention and its release to the market. These trials aim to gather additional information about the long-term safety, efficacy, and optimal use of the treatment in larger and more diverse patient populations. Phase 4 trials can identify rare side effects, assess the intervention's performance in real-world settings, and compare it to other treatment options.

A fundamental principle in the majority of clinical trials is randomisation. Through the random allocation of patients to treatment groups, any patient-specific factors, often referred to as confounding variables, that could have influenced the treatment outcome are expected to be evenly distributed between the arms. In turn, any observed difference in outcomes between the arms can be attributed to the effects of the treatment under investigation. Thus, randomisation supports causal inference, reduces bias, enhances statistical validity, and ensures ethical treatment allocation leading to more reliable conclusions about the effectiveness and safety of treatments (Lim & In, 2019).

Evans *et al.* (2011) catalogues several instances where treatment effects derived from observational data were initially considered beneficial, but when randomised studies were conducted, they were found to be potentially harmful. One notable case is the Women's

Health Initiative Trial (Prentice *et al.*, 1998), in which a hormone replacement therapy initially seemed to lower the risk of heart disease in women. However, upon closer evaluation through randomisation, it was revealed to slightly elevate the risk. This empirical evidence highlights the importance of randomisation in determining the true effects of treatments. I elaborate on individual and cluster-level randomisation in Section 1.1.1, as the methods proposed in this thesis will consider both types of trial design, but refer the reader to Chow & Liu (2014) for other characteristics of clinical trials such as blinding.

The benefits of evaluation through clinical trials notwithstanding, there are unfortunately several drawbacks to clinical trials that must be recognised. Firstly, lengthy phases of clinical development are needed for evaluating a novel treatment (Friedman *et al.*, 2015). The duration required for testing and approving a drug is not standardised, but it can typically range from 10 to 15 years to complete all phases of clinical research before reaching the licensing stage (Cancer Research UK, 2022). As a result, evaluating long-term effects may also require a significant amount of time, delaying the availability of results and potential therapeutic benefits. In some conditions, e.g., in chronic diseases, keeping participants on a placebo arm for a very long duration may be especially difficult. Therefore, many trials only assess the short-term benefit of an intervention (Siderowf, 2004).

Another major issue with clinical trials is that they are very expensive to conduct (Martin *et al.*, 2017). Studies have shown that it costs an average of $2.6 billion to successfully develop a new drug (Mullard, 2014; DiMasi *et al.*, 2016), of which a high proportion is due to the clinical trials required for evaluating efficacy and safety. Specifically, a report from 2014 indicated that the average cost of a phase 3 trial was approximately $20 million (ASPE, 2014). Critically, despite these substantial financial investments, the clinical research process carries considerable risks of financial losses as 70% of Phase 2 trials and 50% of Phase 3 trials do not succeed due to various factors (Fogel, 2018). Here, a trial is defined as unsuccessful if it fails to demonstrate efficacy or safety for reasons, including, but not limited to, under-recruitment, flawed study design, and inappropriate endpoints. Nonetheless, it is important to acknowledge that even trials that do not demonstrate efficacy contribute valuable insights to the body of knowledge. Amongst the reasons for the failure of phase 3 trials, Hwang *et al.* (2016) posit that 22% is due to lack of funding. Thus, the high cost of running the trial and the notably low success rate make the trial process considerably expensive.

A big challenge to many clinical trials is choosing a suitable sample size. Indeed, the ability of the trial to detect a true effect if it exists, often referred to as statistical power, is a critical consideration when designing a trial (Rothwell *et al.*, 2018). Given that the power is a function of the sample size, it is imperative to carefully and adequately justify the sample size selection. Although best practices require that the sample size is calculated before commencing the trial, key parameters for estimating the sample size may be subject

to substantial uncertainty at the design stage of the trial (Bauer & Kohne, 1994). Even in situations where estimates of these parameters may be available from a pilot or previous relevant study, their inherent uncertainty is often not quantified and incorporated in the sample size calculation. Thus, the likelihood of erroneously specifying the required sample size becomes high, impacting the internal validity of the trial (Teare *et al.*, 2014).

For instance, under-estimated nuisance parameters (such as the variance for a normally distributed outcome) can lead to small sample sizes and can result in a trial lacking sufficient power to draw clear conclusions, even if a true effect is present (Altman, 1980; Grayling *et al.*, 2018). Conversely, over-estimated nuisance parameters may lead to an excessively large sample size that increases the trial cost, raises ethical concerns such as needlessly exposing individuals to potentially inferior or harmful treatments, and also delays the study's outcome (Altman, 1980; Grayling *et al.*, 2018). It is worth noting that studies that have investigated the factors associated with the high cost of trials have cited sample size (Martin *et al.*, 2017; Bentley *et al.*, 2019; Nevens *et al.*, 2019), consistently showing that increasing the sample size leads to higher trial costs (Faber & Fonseca, 2014; Bacchetti *et al.*, 2008; Schie & Moerbeek, 2014; Baio *et al.*, 2015; Rutterford *et al.*, 2015).

Finally, the lack of generalisability, also known as external validity, is another drawback often associated with clinical trials (Siderowf, 2004; Chow & Liu, 2014). Generalisability refers to the extent to which trial results can be correctly applied to other situations or populations. Research carried out in the field of Parkinson's disease, for example, revealed that the cumulative incidence of the condition was greater among black individuals when compared to whites and Hispanics (Mayeux *et al.*, 1995). Nonetheless, trials focusing on Parkinson's disease generally include a very small proportion of black participants, often less than 5% (Di Luca *et al.*, 2023; Parkinson Study Group, 2000; Adler *et al.*, 1998). This pattern of under-representation is similarly observed in trials related to multiple sclerosis (Hogancamp *et al.*, 1997). Such limitations affect the external validity of the trial and underscore the necessity for a framework that incorporates under-served populations in clinical trials, thereby enhancing the clinical relevance and overall utility of the trial results (Witham *et al.*, 2020).

Addressing the above challenges requires careful planning, collaboration amongst stakeholders, as well as developing novel methods that strike the right balance for conducting an effective and ethically sound trial. Before discussing potentially useful novel methods, I proceed by presenting an overview of the two types of trial design that will be focused upon in this thesis.

1.1.1 Individually randomised vs. cluster randomised trials

The two types of trial designs that are widely used in practice are individually randomised trials (IRTs) and cluster randomised trials (CRTs). The primary distinction between

them lies in the unit of randomisation. Therefore, many trial designs encountered in the IRT domain (crossover, two-arm superiority, equivalence and non-inferiority trials, etc.) have a CRT equivalent. In IRTs, individual participants are randomly assigned to different treatment groups (Elley *et al.*, 2004) and outcomes are measured at the individual level. Consequently, this type of trial design is commonly used in clinical research to evaluate the efficacy and safety of interventions on individual patients (Torgerson, 2001). An underpinning assumption that participant outcomes are independent leads to simple statistical analyses being employed to determine whether the intervention was effective. The advantages of IRT designs include simplicity and the provision for precise estimates of treatment effects at the individual level. The design can also allow for better control of confounding variables.

However, when investigating the impact of an intervention delivered in group settings, it can pose greater challenges to randomise at the individual level. A CRT, therefore, is essentially a trial which randomises groups (clusters) of individuals rather than the individuals themselves (Eldridge & Kerry, 2012). In practice, these pre-existing groups or randomisation units may comprise for example participants in GP practices (Coventry *et al.*, 2015), health-facilities (Coronado *et al.*, 2014; Cundill *et al.*, 2015; Menya *et al.*, 2015), nursing homes (Gravenstein *et al.*, 2016; Mor *et al.*, 2017), schools (He *et al.*, 2015; Hankonen *et al.*, 2016; Jukes *et al.*, 2017), or communities (Roca *et al.*, 2011; Prudhomme O'Meara *et al.*, 2018). Although randomisation takes place at the cluster level, outcomes are typically measured on individuals within each cluster. The CRT design has a wide utility in health services research, especially where the intervention seeks to influence behavioural and organisational change (Rahman *et al.*, 2008; Kumakech *et al.*, 2009; Butler *et al.*, 2013; Harris *et al.*, 2018), evaluate an infectious disease prevention program (Khan *et al.*, 2012; Perriat *et al.*, 2018), or implement some low-risk intervention within whole communities (O'Brien *et al.*, 2018).

CRTs date back to the 1940s when classroom units were utilised to evaluate an intervention in the educational field (Lindquist, 1940). The methodological difficulties encountered by researchers conducting cluster (community) intervention trials were first published by Cornfield in 1978 (Cornfield, 1978). Since then, the CRT design has become widely accepted and reported in medical research (Moberg & Kramer, 2015), necessitating continued development of its design (Murray *et al.*, 2004; Campbell *et al.*, 2007), analysis (Kerry & Bland, 1998; Hussey & Hughes, 2007), and reporting methodology (Campbell *et al.*, 2012).

An important factor to note about CRTs is that outcomes from individuals within a cluster are expected to be correlated. The degree of correlation is typically quantified by the intra-cluster correlation (ICC) coefficient, which becomes an essential parameter during sample size determination. An extended review of the ICC and its impact on CRTs

is presented in Chapter 3. Given the non-independence of clustered data, the sample size of a CRT must be inflated. That is why, when compared to IRTs, CRTs necessitate a larger sample size to achieve the same level of power (Hemming *et al.*, 2015; Lewis & Julious, 2021). Also, due to the correlated outcomes in CRTs, conventional statistical analyses that assume independence become inappropriate. Specifically, using such methods of analysis may result in the inflation of the type I error rate, incorrect confidence interval coverage, and spurious p-values. CRTs also introduce several other challenges, e.g., they are often unblinded, which introduces a chance for recruitment bias. Together, these issues add an additional layer of complexity to CRTs, meaning careful consideration must be given before opting for a CRT over an IRT.

Notwithstanding the complexities associated with cluster-level randomisation, there are several pragmatic reasons why a CRT may be preferred over a standard IRT. Eldridge & Kerry (2012), as well as Campbell & Walters (2014), provide some advantages and justification for adopting a CRT. A commonly cited reason for the use of CRTs is to prevent contamination between the intervention and control groups. For example, it would be unlikely that residents in a community assigned to the control arm would not notice an education program broadcast on television that was targeted solely at the intervention arm. Thus, individual randomisation may not be appropriate for such an intervention, as the control group would have access to information designated for the intervention group (Hauck *et al.*, 1991; Eldridge *et al.*, 2004). It is important to add that contamination can also occur among those delivering the intervention. For instance, a surgeon who has been trained on a new surgical procedure may find it extremely difficult to alternate between the new and old surgical procedures on different patients. In such a situation, employing randomisation at the level of the individuals responsible for delivering the intervention can also help prevent this possible contamination.

Another reason to use a CRT design is that it may be cost-effective and more convenient to administer the intervention to a group, e.g., if the intervention involves an expensive piece of equipment with high-level training. In some cases, the intervention may also specifically be targeted at clusters with the aim of influencing the knowledge, attitude, or practices of the whole group. Sometimes, ethical reasons may require the use of a CRT since it may be unethical to withhold certain types of intervention from people within the same cluster. For example, it would be unethical to withhold a vaccine intervention from some members within the cluster knowing its life-saving benefits. Instances where this concept is relevant could include large trials conducted in LMICs, where it becomes more feasible for fieldworkers to sequentially roll-out the same intervention among villages.

In sum, the choice between an IRT and a CRT depends on various factors, such as the nature of the intervention, the research question, available resources, logistical considerations, and the level at which the intervention can be applied. Careful consideration

of these factors is necessary to select the most appropriate trial design for specific study objectives.

In what follows, I present a comprehensive overview of both fixed and ADs to establish a contextual foundation for the application of ADs in LMICs.

1.1.2 Fixed designs vs. adaptive designs

Fixed trial designs have long been the common approach for evaluating the safety and efficacy of new treatments since the inception of clinical research. Fixed designs follow a structured process that includes designing the trial, conducting it as per the predetermined protocol, and analysing the data according to the pre-specified plan. They are still prevalent and relevant in clinical research, despite the growing interest in innovative trial methodologies. Though they are simpler to design and conduct, fixed designs offer little flexibility to make potentially desirable or necessary changes to the trial's design features. Hence, some ethical issues can arise in traditional fixed trials due to the unresponsive nature of their design (Bothwell & Kesselheim, 2017).

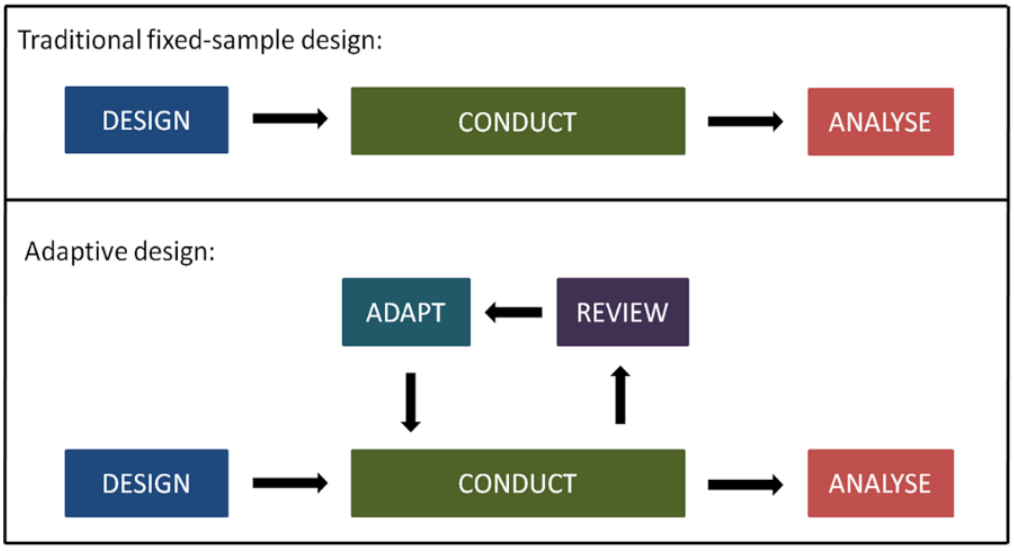

Figure 1.1: Schematic of a traditional clinical trial design with fixed sample size, and an AD with pre-specified review(s) and adaptation(s). Figure adapted from Pallmann *et al.* (2018).

While fixed designs have been the backbone of clinical trials for decades and have contributed significantly to medical advancements, there is growing recognition of the potential benefits of utilising ADs and novel statistical methodologies to enhance trial efficiency and address complex research questions. ADs are a broad class of approaches that aim to provide a more cost-effective, efficient, and ethical trial compared to classical fixed alternatives (Pallmann *et al.*, 2018). They allow the investigator to make pre-specified

changes during the trial based on one or more interim analyses of the accumulating data while maintaining the integrity and validity of the trial. Design features that could be modified include the sample size (via sample size re-estimation), the allocation ratios to the trial arms (via outcome adaptive randomisation), which treatments are present (via dropping of arms), and the considered dosages (via dose-finding/ranging).

By making adjustments in response to emerging data, ADs can improve the probability of a successful trial. These adaptations make ADs more efficient compared to traditional fixed designs. For example, by dropping ineffective treatment arms early in the trial, ADs have the potential to optimise resource utilisation by reducing the number of participants needed. This efficiency can lead to cost savings, making better use of limited resources. Moreover, the ability to make interim decisions in ADs accelerates the drug development process. Thus, researchers can identify successful treatments more quickly, potentially speeding up the time it takes for new therapies to reach patients. Furthermore, early stopping rules in ADs can be incorporated to protect patient safety. For instance, if there is clear evidence of harm or benefit, the trial can be stopped early to prevent unnecessary exposure to ineffective or unsafe treatments. ADs have other advantages, particularly in situations where there is uncertainty about the most effective treatment or intervention (Sampson & Sill, 2005; Chow & Chang, 2008). They are also applicable to all phases of clinical research and are becoming more widespread in their use (Pocock, 2013; Friedman *et al.*, 2015; Pallmann *et al.*, 2018; Haeussler & Assmus, 2021).

While ADs offer several advantages, they also come with certain disadvantages. They are inherently more complex and may require more sophisticated statistical methods to handle the dynamic nature of the trial than traditional fixed designs. The need for pre-specified adaptation rules and the potential for multiple interim analyses can make the trial design, conduct, and analysis more intricate, requiring specialised expertise. Besides, the flexibility in adapting the trial may increase the probability of making a Type I error (incorrectly concluding that there is a treatment effect when there isn't). Therefore, appropriate adjustments and control mechanisms are necessary to mitigate this risk. Given the complexities associated with ADs, Wason *et al.* (2019) assert that they are not always beneficial and provide some considerations around their utility. These include:

- **Number of interim analyses:** According to the authors, having more than two interim analyses during a trial is rarely justified, as the additional benefit gained is not significant enough to outweigh the burden involved, unless the trial spans an extended period or involves the addition of new treatment arms as it progresses.
- **Limitation due to early stopping:** Although early stopping is a benefit of ADs (allowing for quick decisions on effective treatments and saving patients from ineffective ones), it also has its drawbacks. Ending the trial quickly also means missing

out on valuable information about secondary outcomes, safety assessments, and efficacy within subgroups. Besides, there is a concern that the estimated treatment effect might not be as accurate with a smaller sample size due to the early stopping decision. This poses a challenge when attempting to implement changes in clinical practice based on such limited evidence.

- **Long term outcomes:** The efficiency of an AD is derived from the valuable information available at interim analysis, which allows for more accurate predictions of the trial's eventual outcomes if it were to continue to completion. However, realising these benefits heavily relies on making reliable decisions during the interim analysis. Therefore, the outcome information used for adapting the trial must be observed rapidly enough compared to the planned recruitment period of the trial. If this condition is not met, the trial might conclude recruitment before enough outcomes are observed to be able to reliably modify the trial's course.
- **Limitation due to additional administrative and logistical complexity:** A well-established infrastructure and significant resource investment is needed for prompt and accurate interim analysis. Therefore, the lack of logistics to facilitate timely data return, rigorous data cleaning, and effective communication can diminish the efficiency benefits of ADs.

Additional non-methodological barriers to ADs include the absence of readily available, well-documented, and user-friendly software, insufficient funding structures, a shortage of experts in ADs, the considerable time needed for trial design with ADs, and concerns about potential operational biases arising from their implementation (Chow & Corey, 2011; Kairalla *et al.*, 2012; Dimairo *et al.*, 2015; Grayling & Wheeler, 2020). Despite these challenges, ADs remain a valuable tool in clinical research when appropriately planned and executed. Given the benefits and challenges associated with both fixed and ADs, their appropriateness depends on the specific research question and context. It is therefore essential to strike a balance between the advantages of well-established traditional designs and the opportunities offered by more flexible and adaptive approaches to ensure the advancement of medical knowledge and the improvement of patient outcomes.

In this thesis, particular attention is given to the *drop-the-loser* design discussed in Section 1.1.3 and the seamless phase II/III design, covered in Chapter 2. I refer the reader to Pallmann *et al.* (2018) for an extensive overview of the different types of ADs in clinical trials.

1.1.3 Drop the loser adaptive design

In the drug development process, it can be of great importance to evaluate multiple treatments or doses alongside a control group. This evaluation allows for the option to eliminate

less effective treatments or advance more promising ones at specified stages of the trial. Such adaptive trial design is broadly referred to as the multi-arm multi-stage (MAMS) design. When the primary interest of the trial lies in dropping less effective treatment arms at the end of the first stage, based on some pre-specified ranking criteria, this variant of the MAMS design is referred to as the *drop-the-loser* (or sometimes referred to as *pick-the-winner*) design.

A drop-the-losers design can be particularly beneficial during phase II or early clinical development, especially in situations where uncertainties exist concerning varied treatments or dose levels (Bauer & Kieser, 1999; Posch *et al.*, 2005). This is because the design allows for the selection of one superior treatment (or dose) in the first stage and a confirmation of the selected treatment's efficacy in the second stage under a single trial protocol (Sampson & Sill, 2005). An important characteristic of the design is a shared control arm and the use of data from both stages of the trial for the final analysis.

In practice, such trials are typically designed to achieve a specified power at the completion of the second stage. Thus, there might be no statistical power constraint for the analysis at the end of the initial stage, for the purpose of eliminating ineffective treatments (Chow & Chang, 2008). In such instances, a common approach is to eliminate less successful treatments based on the estimated treatment means. Specifically, if the primary endpoint represents a measure where higher values signify a more favourable treatment response, the treatment exhibiting the highest average response during the first stage will be chosen for further exploration.

The characteristics of the design, such as the avenue for multiple treatment evaluation, the shared control, the simplicity of the selection criteria, and the combination of two trial phases into a single trial protocol offer competitive advantages over traditional designs. The main benefits include a reduced need for sample size by employing a shared control group instead of individual control groups for each treatment with the added benefit of direct comparison between the experimental treatments (Wason *et al.*, 2017). Also, the use of a single protocol can greatly shorten the drug development timeline, possibly eliminating completely the delays between the phases.

Another primary advantage of the drop-the-loser design is the potential for resource savings by concentrating resources on treatments that show the most potential and are more likely to produce meaningful results. It is well established that the clinical trial paradigm can be resource-intensive and time-consuming. Therefore, by discontinuing arms that show little promise, researchers can direct their resources more efficiently toward arms that demonstrate better outcomes. This not only leads to cost savings but also expedites the trial process, ultimately benefiting patients and improving the speed at which promising treatments can reach the market.

In addition, the drop-the-loser design addresses some ethical considerations. In fixed

designs, for example, maintaining ineffective arms throughout the study raises ethical concerns, as it involves exposing more patients to treatment arms with limited potential benefits. By promptly discontinuing these arms, the drop-the-loser design prioritises patient welfare and ensures that participants are not unnecessarily subjected to ineffective treatments.

Despite the advantages of the drop-the-loser design, it is important to note that it is not suitable for all clinical trials. As indicated earlier, it is most commonly applied in early-phase trials, where a range of treatment arms are initially explored to identify the most promising options. As a result, this methodology may not be as appropriate for later-phase trials, where a treatment's efficacy and safety profile are better established. Additionally, conventional approaches employed for making inferences using data from both stages may result in tests with significance levels higher than desired due to issues with multiple testing. Some approaches to correct these issues are discussed in detail in Chapter 2. Furthermore, by empirically selecting the "optimal" treatment, it is important to acknowledge that this choice might not represent the best treatment from a population perspective (Sampson & Sill, 2005). This is due to the fact that dose or treatment groups that are excluded may hold valuable information concerning the dose-response of the treatment being investigated (Chow & Chang, 2008).

It is worth noting that some fixed designs also allow for multiple treatment arms to be included with an equal allocation of resources and participants to each arm in most cases. In such settings, the allocations remain unchanged throughout the trial's duration, regardless of emerging data indicating the performance of individual arms. In contrast, the drop-the-loser design allows for real-time data analysis, and as results accumulate, ineffective or less promising arms are discontinued. By doing so, this design significantly improves the efficiency of the trial relative to the multiple comparisons in fixed designs. While the AD drop-the-loser design is not a one-size-fits-all solution, it addresses some of the key limitations of fixed trial designs. It also holds great promise for improving the efficiency of clinical trials, making it a valuable tool in the quest for new and effective medical treatments.

Having established the usefulness and complexities of ADs, the next section discusses trials in LMICs, highlighting the challenges associated with clinical research in those settings and exploring how ADs may offer potential solutions to overcome these barriers.

1.2 Trials in low and middle-income countries

According to the World Bank, LMICs are economies with a gross national income per capita between $1,136 and $4,465 (The World Bank, 2022). Therefore, these countries are typically characterised by their relatively lower economic development and income

levels compared to high-income countries (HICs). LMICs often face significant challenges in terms of infrastructure, healthcare systems, education, and access to resources. It is important to note that the classification of countries into income groups is not static and can change over time as countries undergo economic development and experience shifts in their socio-economic status. Therefore, at the time of submitting this thesis, there are 134 LMICs, which together represent the majority of the global population (The World Bank, 2022; UNCTAD, 2022).

Collectively, LMICs carry approximately 90% of the global disease burden, primarily consisting of preventable infectious diseases (Global Forum for Health, 2002). Moreover, there is a growing prevalence of non-communicable diseases in these nations (Marshall, 2004; Alemayehu *et al.*, 2018), a transition which places an additional burden on their already strained resources (Alwan, 2011). Although studies suggest that populations in LMICs stand to benefit from clinical trials because of the greater health burden in such settings (Mbuagbaw *et al.*, 2011; Grover *et al.*, 2017; Rosala-Hallas *et al.*, 2018), they are under-represented in health research (Mbuagbaw *et al.*, 2011). A study conducted in 2018 revealed that there was a significant disparity in the distribution of clinical trial sites, with approximately 83% of trials taking place in 25 HICs, while less than 5% were conducted across 91 LMICs (Drain *et al.*, 2018). A similar phenomenon was found in pediatric health, where only 25% of pediatric trials were conducted in LMICs although 98% of the global disease burden is carried by children who reside in those regions (Nor Aripin *et al.*, 2010).

While certain clinical research priorities might appear relevant to both LMICs and HICs, variations in the specific needs of these two regions still exist. In the UK, for example, a strong emphasis is placed on recruitment and retention which is in contrast to the top 10 most critically important priorities for LMIC researchers (Rosala-Hallas *et al.*, 2018). Therefore, addressing questions that are specific to the particular context is essential for designing interventions that effectively enhance equitable health outcomes (Barreto, 2009). It is against this background that the World Health Organisation encourages countries to become independent producers and consumers of health research due to the distinctive health needs and unique challenges in individual countries (WHO, 2013).

Despite the WHO's advocacy for bespoke trials, substantial progress within LMICs has not been achieved. This is because the advancement of clinical research in LMICs is hindered by several significant obstacles including limited financial resources, unnecessarily lengthy delays in ethical approval procedures, and regulatory and administrative challenges (Joseph *et al.*, 2016; Alemayehu *et al.*, 2018). Among these challenges, Alemayehu and colleagues identified insufficient funding as a key barrier, highlighting that the majority of clinical trial funding originates from HICs or pharmaceutical companies based in those regions. In terms of global health, only 10% of global health research funds are allocated towards addressing medical conditions in LMICs, even though they have the

highest disease burden (Global Forum for Health, 2002). Also, developing nations typically allocate minimal funds for research and overall healthcare, thereby exacerbating the funding constraints faced in clinical research. This evidence suggests that despite the high disease burden in LMICs, most of these countries lack the necessary resources to close the under-representation gap (Khoja *et al.*, 2019).

Given the scarce resources in LMICs, ADs may be advantageous in bridging the under-representation gap as they make better use of resources. The flexibility of ADs could enable researchers to allocate resources where they are most likely to yield meaningful results. For example, by reallocating resources to promising treatment arms through adaptive randomisation, AD could increase the likelihood that limited resources are used effectively. Likewise, if the observed effect size is larger than anticipated based on interim data analysis, the trial might be stopped earlier, conserving resources while still achieving meaningful results. Owing to this, over 95% of researchers and methodologists working on trials in LMICs regarded research topics on ADs as important, while 67% (210/314) saw it to be critically important (Rosala-Hallas *et al.*, 2018). More so, the use of ADs in global health is highly encouraged (Lang, 2011; Rosala-Hallas *et al.*, 2018), and the advent of the COVID-19 pandemic has enhanced such calls (Stallard *et al.*, 2020).

Bridging the under-representation gap in LMICs through cost-effective clinical trials presents several opportunities and potential benefits. Firstly, including LMICs in clinical research allows for a more diverse study population, ensuring the generalisability of research findings and the applicability of interventions across different ethnicities and geographical regions. This diversity contributes to a more comprehensive understanding of treatments and interventions' efficacy and safety profiles. Moreover, conducting trials in LMICs can enhance access to innovative therapies and interventions for local populations. In some cases, trial participants may gain early access to potentially life-saving treatments that are not yet available in their countries through regular healthcare channels. This can bring significant benefits to patients who may otherwise face limited treatment options due to economic constraints or underdeveloped healthcare systems. Additionally, clinical trials in LMICs contribute to capacity building and research infrastructure development.

It is worth noting that, both IRTs and CRTs designs are commonly used in developing countries. However, CRTs are particularly appealing for evaluating interventions in LMICs, especially in regions such as sub-Saharan Africa (Isaakidis & Ioannidis, 2003), and the usage of CRTs in this context dates back to the early 1970s (Eldridge & Kerry, 2012). Despite the widespread use of CRTs in LMICs, though, empirical evidence suggests that there is still a lack of extensive recognition regarding the methodological challenges associated with CRTs in research conducted within the region (Isaakidis & Ioannidis, 2003). The authors further noted that there was a lack of consideration and proper reporting of the prerequisite design and analysis aspects of cluster randomisation in the majority of the

trials they reviewed. Specifically, the ICC and design effect which are essential to sample size calculation were rarely reported. These methodological issues have implications for the validity of trials and highlight the urgent requirement for conducting research aimed at identifying robust, cost-effective, and efficient methodologies for trialists operating in resource-constrained settings (Franzen *et al.*, 2017; Grover *et al.*, 2017).

Considering the wide context in which ADs in LMICs are situated, I now move forward by establishing the scope of this study. This scope is mainly determined by factors such as relevance and the practicality of achieving results within a reasonable academic time frame. These predefined boundaries and constraints of the research are anticipated to guide the central focus of the thesis, provide a premise for the research objectives, and provide a context for the novelty within the thesis.

1.3 Scope of the thesis

A recent study investigating the participation of LMICs in randomised clinical trials conducted by HICs discovered that the majority of such trials were predominantly conducted in India and Ukraine. In fact, a substantial 96% of the total trials conducted in LMICs were in these two countries (Rubagumya *et al.*, 2022). This finding suggests that African populations are underrepresented in these trials, raising concerns about the generalisability and applicability of the research findings to diverse global populations, particularly those from African regions. In 2005, the WHO highlighted the global importance of creating African-owned research centres with the capability to conduct independent clinical trials (Matsoso *et al.*, 2005). However, the advancement of this initiative has been constrained(Cardoso *et al.*, 2015; Zegers-Hochschild, 2011). It is also important to acknowledge that, although LMICs typically have common resource limitations, specific nations may have distinct issues (Rosala-Hallas *et al.*, 2018). For example, LMICs such as Brazil, Ukraine, and China are more resourced than the LMICs in Africa.

Therefore, while the methods espoused in this thesis are envisaged to have a wider utility, even among HICs, the focus of the thesis is to empower LMICs in Africa with cost-effective methodologies that can improve their participation in clinical research and bridge the existing under-representation gap in clinical trials. This approach seeks to promote equitable access to innovative trial designs and ensure that research findings are more representative and applicable to diverse global populations, ultimately advancing healthcare outcomes for all. Thus, reference to LMICs in the thesis shall connote LMICs in Africa.

Although a plethora of issues have been highlighted as barriers to using ADs in LMICs, this thesis will focus on proposing methodologies that are both robust and cost-effective. Specifically, the efficiency espoused in this thesis shall be related to sample size for the

following key reasons:

1. **Improve internal validity of trials:** If LMICs are to adhere to the call by the WHO urging for self-reliance in generating and utilising health research, it is essential that trials conducted in these countries possess internal validity. As previously noted, trials might lack internal validity due to the dependence of sample size on parameters that are difficult to accurately determine at the commencement of the trial; whereas inaccurately specifying these parameters can undermine the internal validity of the trial. In essence, an accurate sample size determination is a cornerstone of a well-designed clinical trial, ensuring that the study produces valid, reliable, and generalisable results that can inform medical practice and decision-making.

2. **Resource optimisation:** Determining the optimal sample size helps avoid unnecessary resource wastage, as recruiting and managing more participants than needed can be costly and time-consuming. Due to the relationship between sample size and cost, developing cost-effective and efficient sample size methodologies maximises the limited resources in LMICs and enhances their capacity to conduct high-quality research.

It is imperative to add that sample size determination was among the top 10 most critical items listed in both the priority rankings for LMICs and HICs (Rosala-Hallas *et al.*, 2018). While the primary focus or scope remains on these methodological issues, I comment on non-methodological issues in the discussion presented in Chapter 6.

1.4 Aims and objectives

The research aims to develop novel methodologies that address barriers to conducting clinical trials in LMICs, with a particular focus on ADs. To achieve this aim, the research was guided by the following specific objectives:

- To develop a general framework for the optimal *drop-the-loser* trials when an intermediate endpoint is used for interim selection.

- To develop a hybrid approach to designing parallel-group and stepped-wedge cluster randomised trial designs when there is uncertainty in the intra-cluster correlation.

- To develop methods that facilitate sample size re-estimation in cluster randomised trials.

1.5 Novel contributions within this thesis

According to Munos (2009), the escalating costs associated with drug development have led to a greater demand for faster and more cost-effective trial results. Consequently, there has been a conscious effort to optimise methodological research into the design, conduct, analysis and reporting of trials (Rosala-Hallas *et al.*, 2018). Nonetheless, there is still considerable scope for further improvement, particularly in LMICs. Therefore, the novel contribution in this thesis constitutes an advancement in the methodological literature by 1) addressing gaps in existing proposals and, 2) introducing innovative perspectives and methodologies that extend the current boundaries of knowledge and pave the way for future research endeavours.

Since LMICs are less likely to undertake expensive data collection methods due to resource constraints, studies could benefit from identifying which treatments might work before they commit to using expensive data collection methods. Therefore, the first objective proposes a framework extending the drop-the-losers design to be more amenable to a seamless phase II/III setting. Specifically, this design allows for the simultaneous comparison of multiple treatments with a shared control in the phase II stage to identify the most promising treatment. Subsequently, a selected treatment is then seamlessly advanced to the phase III stage for confirmatory testing. Owing to the limited resources within LMICs, an intermediate endpoint is assumed in phase II to identify the most promising treatment. Therefore, I stipulated that early patients in the trial will only need their intermediate outcomes to have been measured for the interim analysis, facilitating a cheaper and more efficient method of data collection and lower trial cost. This concept revolves around the notion that the adaptation occurs before most patients experience a definitive outcome. I.e. there is no need to measure the definitive outcome for the arm that is removed from the trial. Thus, the interim analysis informs which arm to drop and which one to collect more expensive data on for the remainder of the trial.

The optimal timing of interim analysis has not been extensively considered for such designs, and will be given specific attention. This is significant because the adaptations and decisions within ADs heavily rely on the quality and amount of information available at the interim analysis. Previous studies that have investigated the optimal timing for interim analysis have predominantly focused on discrete intervals (Wason *et al.*, 2017) or examined the relationship between the optimal timing and the expected sample size relative to its standard deviation (Grayling & Mander, 2021). In this study, I examine the optimal timing for interim analysis as a continuous variable and determine the percentage reduction in sample size compared to alternative timings. Subsequently, I investigate the combinations of values for both the intermediate and definitive outcomes, along with their correlation, that maximises the family-wise error rate (FWER).

Having established from the literature that CRTs are widely employed in LMICs, the next objective provides a possible solution to the issues around specifying the ICC (or other design parameters), known as the 'hybrid' (sometimes called 'Bayesian-frequentist') approach. By considering the impact of the ICC prior, the aim is to gain a deeper understanding of which CRT design is more efficient under varying ICC scenarios, providing valuable insights for the selection of an appropriate CRT design. While hybrid methods have been extensively studied in IRTs, they have received limited attention in the context of CRTs (see, e.g., Kunzmann *et al.*, 2021). The most pertinent study related to CRTs designed within the hybrid framework is that of Lewis & Julious (2021). They utilised confidence intervals to define a plausible range for the ICC and incorporated this uncertainty into the sample size calculation. However, they did not associate a specific prior density with each potential ICC value. In this study, I assign parametric prior distributions to the ICC, total variance, as well as the target effect at which the trial is powered. Considering that the success of a clinical trial relies heavily on accurately determining the required sample size, which also impacts the trial's cost, the methods advocated in this research aim to improve the likelihood of a successful CRT. This likelihood is quantified using the 'probability of success', which typically takes the place of the frequentist power in the hybrid literature.

Building on the previous work, the final objective considers an AD in the form of sample size reestimation (SSRE) in CRTs within a hybrid framework. This allows for adjustments to the final sample size, taking into account the information gained from the interim data. By leveraging an adaptive approach, this objective aims to enhance the accuracy of the final sample size, thereby improving the validity and efficiency of the overall trial design. To the best of my knowledge, previous work on SSRE in CRTs has been conducted in purely frequentist (see, e.g., Lake *et al.*, 2002; Schie & Moerbeek, 2014; Grayling *et al.*, 2018) and Bayesian frameworks (see, e.g., Wang, 2007; Brakenhoff *et al.*, 2019; Zhong *et al.*, 2013), leaving avenue for research within the hybrid framework. Specifically, the prior distribution for an ICC is updated based on interim data, utilising the posterior to determine the final sample size. The results from the hybrid SSRE procedure are compared to frequentist SSRE, using frequentist operating characteristics such as the type I error rate. I then ascertain instances where the hybrid SSRE framework offers a distinct advantage over the frequentist approach, in particular assessing when the hybrid SSRE approach overcomes known challenges of the large variation in the final sample size that is associated with the frequentist approach. It is worth noting that this is the first study to examine the use of Mean Squared Error (MSE) as an evaluation metric for a CRT SSRE approach, and this approach is attractive compared to the typical practice of separately comparing power and sample size summaries, which is the conventional method.

1.6 Organisation of the thesis

Chapter 2 provides an overview of seamless phase II/III trial designs, discussing their advantages and exploring statistical issues in recent studies. It identifies a gap in the literature concerning the use of intermediate endpoints within seamless designs. It then proceeds to explore the statistical methods utilised in such trial designs, focusing on the drop-the-loser design. Particular attention is given to developing a methodology for determining the optimal time for interim analysis. The subsequent section of the chapter presents results obtained from the selected motivating example, namely the TAILoR trial. Finally, the chapter concludes with a comprehensive discussion of the results and their implications.

Chapter 3 reviews various approaches used in practice for specifying the ICC and other CRT design parameters. This review serves as motivation for the development of the proposed methods in Chapter 4 and Chapter 5. The methodology section of the chapter captures the data sources used and the review strategy employed, the inclusion and exclusion criteria, as well as the data extraction and synthesis. The evidence from the review is then summarised using descriptive statistics, frequency distribution tables, and graphs. A discussion of the results also notes potentially useful methods for specifying the ICC that appear to have been sparsely utilised to date.

In Chapter 4, a sample size estimation approach is introduced, which addresses the uncertainty in sample size parameters by employing parametric distributions for the key sample size parameters. The chapter begins by defining the parallel group (PG) and stepped wedge (SW) CRT designs considered within the study along with their statistical principles and assumptions underpinning both designs. The introduction section further highlights some possible approaches to account for uncertainty in the sample size parameters and argues for the preferred hybrid approach. Subsequently, these hybrid quantities, the analysis models of the CRT designs, motivating examples for the study, and the choice of priors are extensively discussed with notations in the methodology sections. Thereafter, a comprehensive assessment of how incorporating a prior on the ICC affects the efficiency comparison between a PG and a SW CRT design is conducted in the results section. The chapter concludes with a discussion of the results, the limitations of the study, and the implications of the findings on CRT designs.

Chapter 5 begins by identifying gaps in the hybrid literature and proposes how incorporating an AD element could address the limitations of the proposed method presented in Chapter 4. In the introductory section, the SSRE concept, along with its advantages and drawbacks, is elucidated in detail. The methodology section provides a high-level overview of how SSRE is performed both in the frequentist and hybrid frameworks. The section further captures the simulation study, the Bias and MSE formulae, as well as the

plot of the selected priors. In the results section, I evaluate, among other factors, whether the hybrid SSRE technique effectively addresses the well-known challenges associated with frequentist SSRE in CRTs. I also assess the bias, MSE, and power trade-off based on the informativeness of the prior. The final section of the chapter discusses the results, acknowledges the limitations of the study, and makes practical recommendations for the use of this method.

In each chapter, wherever possible, real examples from clinical trials are incorporated to illustrate the practical application of clinical and statistical concepts. These motivating examples serve to elucidate the relationships and interactions between these concepts, providing a tangible context for readers.

In Chapter 6, the thesis concludes with a discussion of its primary contributions. The main findings, innovative approaches, and valuable insights identified in the previous chapters are summarised, highlighting their significance in the context of LMICs. Additionally, it discusses the potential benefits and challenges specific to these settings and how ADs can be tailored to address the unique healthcare landscape and resource constraints in LMICs. Likewise, the chapter explores directions for further work in the area of ADs in LMICs. It presents potential research opportunities and avenues for advancing the implementation and understanding of adaptive methodologies in these regions.

Chapter 2

Optimal drop-the-loser trials when an intermediate endpoint is used for interim selection

In this chapter, a general framework for the design of multi-arm two-stage drop-the-loser trials is proposed for when an intermediate outcome variable is used at the interim analyses. The optimal timing of this interim analysis, to minimise the required study sample size, is also described. Finally, an evaluation is conducted regarding the performance of the proposed methodology in comparison to more traditional design approaches.

2.1 Introduction

As highlighted in the preceding chapter, in clinical development multiple experimental treatments are often under consideration for evaluation in phase II, with one promising experimental treatment then selected for a confirmatory phase III trial. Quan *et al.* (2020) suggest if everything runs smoothly then it typically takes around 9 months (the "white gap") between the analysis of such phase II data and the recruitment of patients for the phase III trial, owing to the development of hypotheses and receiving of study protocol approval from regulatory boards for the phase III trial. This is one example, amongst many, of the ways in which significant challenges are faced in drug development because of the extensive time and resources required to discover, develop, and demonstrate the advantages of a new drug (Maca, 2006).

To overcome this challenge, considerable effort has been directed towards finding ways to expedite and enhance the efficiency of drug development, while maintaining the integrity and validity of the process. One such approach is to combine within a single trial the objectives that have conventionally been handled in separate phase II and phase III trials,

using a seamless phase II/III design (Jenkins *et al.*, 2011; Stallard & Todd, 2011; Hampson & Jennison, 2015; Friede *et al.*, 2020). Seamless phase II/III designs remove the white gap that would have existed between conducting the phase II and phase III trials separately, and could also potentially offer additional efficiencies in terms of the total required sample size (Bretz *et al.*, 2006; Wason *et al.*, 2017). An illustrative schematic of a seamless design, where a decision is taken at an interim analysis regarding the experimental treatment arm that will proceed to the confirmatory stage is presented in Figure 2.1.

Principally, two categories of seamless phase II/III design exist. The *inferentially* seamless design, which shall be explored methodologically in this chapter, incorporates data from patients enrolled both before and after the adaptation process (i.e., the phase II and the phase III data) in the final analysis. Whereas, the *operationally* seamless design only includes data from the phase III portion of the trial in the final analysis (Bretz *et al.*, 2006; Maca, 2006; Quan *et al.*, 2020). For the inferentially seamless approach, a major statistical challenge arises from combining data from the two stages: control of the family-wise error rate (FWER) (Friede *et al.*, 2020).

Fortunately, some methods now exist to combine data from the two stages of a seamless phase II/III design while appropriately controlling the FWER. The majority of such proposed methods can be categorised into two groups. The first group builds on the group-sequential methodology developed by Thall *et al.* (1988, 1989), and requires that there is only a single treatment and control that continue beyond the first stage (Stallard & Todd, 2003), or that the number of treatments at each stage is predetermined (Stallard & Friede, 2008). The combination test approach developed by Bauer & Kohne (1994) serves as the foundation for the second group of methodologies. While these methods offer more adaptability (see, for example, Bauer & Kieser, 1999; Posch *et al.*, 2005; Bretz *et al.*, 2006), they might be less powerful in certain scenarios (Friede *et al.*, 2011).

Magirr *et al.* (2012) introduced a group-sequential method that permits entirely flexible treatment selection. I.e., the number of treatment arms within any stage need not be prespecified. However, this flexibility might come at the expense of being conservative and consequently experiencing a decrease in power. The Dunnett test (Dunnett, 1955) has also been extended by Koenig *et al.* (2008) to a two-stage design with adjustable treatment selection using the conditional error principle of Müller & Schäfer (2001). It has been demonstrated that this strategy outperforms alternative approaches in terms of power (Stallard & Friede, 2008). Stallard & Todd (2011) have since effectively highlighted the similarities and differences amongst available data combination methods. However, it is still not clear how to select an optimal data combination method for all trial settings (Hampson & Jennison, 2015). In essence, the optimal method for combining data may vary depending on the specific characteristics of the clinical trial design and questions of interest.

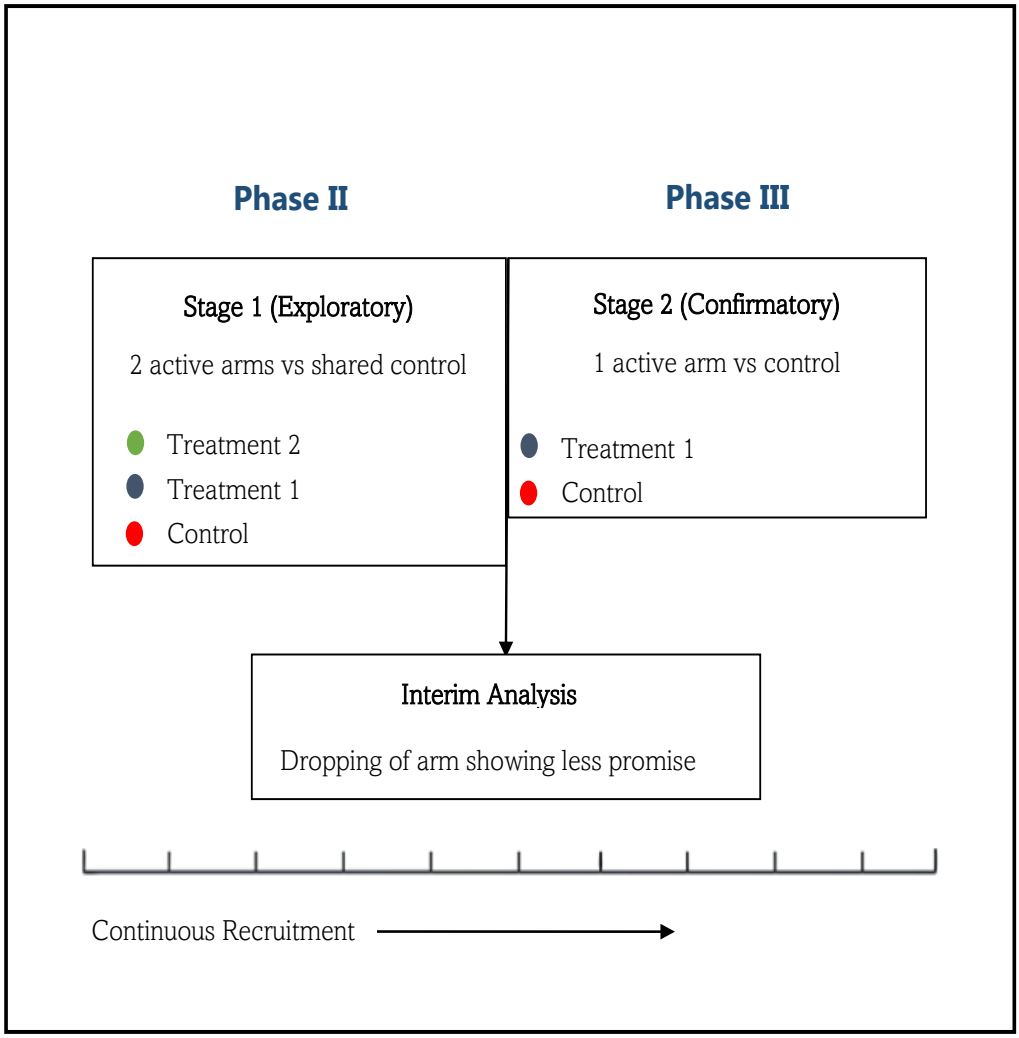

Figure 2.1: Schematic of a seamless phase II/III design where a treatment arm is dropped at an interim analysis and the remaining promising treatment arm proceeds to a confirmatory stage.

Given these complexities, the FDA has indicated that seamless designs should be used with caution, as the design is not well-understood and the statistical methods underpinning the design are not well-established (FDA, 2010; Gallo *et al.*, 2010). The use of extensive simulations to evaluate the operating characteristics of these complex designs at the trial design stage has therefore been encouraged (Friede *et al.*, 2010; Benda *et al.*, 2010).

It is important to note that not all clinical development programs may be suitable candidates for a seamless phase II/III design approach (Maca, 2006). For example, an adaptive seamless design would probably not be suitable for a phase II program in a novel disease area if the objective is to establish the primary endpoint to be used in phase III. Maca (2006) discusses other criteria that determine the feasibility and merits of a

seamless approach. In all, while seamless designs offer the potential for improved efficiency and faster development timelines, they also carry certain limitations and challenges. In turn, significant effort is therefore needed to carefully plan their execution and ensure the validity of the proposed analysis (Hampson & Jennison, 2015).

In this chapter, the focus will be on facilitating such careful design of a specific type of seamless design: one in which an intermediate endpoint is used to select which of two experimental treatments to bring forward to the confirmatory stage. Therefore, I next discuss the various types of endpoints typically employed in clinical trials.

2.1.1 Clinical endpoints

Selecting appropriate and relevant endpoints is critical in the design of a clinical trial as it is ultimately the results for these endpoints that determine trial success and the ability to have established meaningful results that support regulatory approval and clinical decision-making. In general, an endpoint is an event or outcome that can be objectively or subjectively measured to determine the benefits of a studied intervention (Chow & Liu, 2014; Friedman *et al.*, 2015). There are principally two types of endpoint, namely, primary and secondary endpoints. The primary endpoint is the main outcome that the study is designed to assess. Almost always it is used to determine the treatment's efficacy and/or safety features. The success or failure of the intervention is often very clearly determined by whether it achieves a predefined level of significance for the primary endpoint. Therefore, the primary endpoint plays a crucial role in providing the basis for regulatory actions (Food and Drug Administration, 2017). Conversely, secondary endpoints provide supplementary information about the intervention's effects and can help researchers gain a broader understanding of its impact on patients (FDA-NIH Biomarker Working Group, 2021). Unlike the primary endpoint, trials are not always powered to detect differences in secondary endpoints.

Endpoints can take many measurement forms, depending on the nature of the trial and the medical condition being studied. We broadly categorise these various types of measures as either definitive or intermediate.

Definitive endpoints are direct measures of patient health or disease status, such as survival rates, symptom improvement, disease progression, or relapse rates. While direct objective measures of how a patient feels, performs, or survives have impacted clinical practice and are often considered optimal, they typically take a long time to observe and require large sample sizes with increased cost (Wittes *et al.*, 1989). In cancer trials, for example, overall survival is a common endpoint but can take an extended period to assess (Huang *et al.*, 2009; Bratton *et al.*, 2016). This issue is exacerbated in settings where a study of the definitive endpoint may be impossible due to low prevalence, or where it may be less frequently possible to use expensive direct measures throughout the trial due to

low resources.

Intermediate endpoints, on the other hand, are indirect measures that are used as a substitute for a definitive endpoint when the definitive endpoint is difficult or time-consuming to measure (Wittes *et al.*, 1989; Dafni & Tsiatis, 1998). Through validation, intermediate outcomes could become surrogate outcomes. However, the process of establishing an intermediate outcome as a surrogate endpoint is complex and requires a multidisciplinary approach involving clinical, statistical, and regulatory expertise. According to Lagakos & Hoth (1992), the AIDS epidemic stimulated a growing interest in utilising intermediate endpoints as a foundation for assessing treatment efficacy. Thereafter, the use of intermediate outcomes has gained increasing popularity and has established a wide utility in practice. Examples of intermediate outcomes include biomarkers, physiological measures, imaging results, laboratory test results, or specific biological indicators that can be measured in the body and may provide insights into the intervention's mechanism of action or treatment response (DeMets *et al.*, 2020).

intermediate endpoints hold immense potential in drug development, offering the opportunity to significantly enhance efficiency and cost-effectiveness (Dafni & Tsiatis, 1998). Specifically, using an intermediate endpoint which is usually cheaper and measured early can offer a reduced required sample size compared to a trial involving a definitive endpoint (Wittes *et al.*, 1989; Friedman *et al.*, 2015; Benjamin *et al.*, 2016). Moreover, the quick observation of intermediate endpoints plays a crucial role in accelerating the trial process, ensuring swift access to a successful drug for patients who depend on it. This expedited approach is not only efficient but also holds ethical significance in promptly addressing the needs of patients awaiting effective treatments. It is important to note that several drugs have now received FDA approval using intermediate endpoints (FDA, 2018).

Despite their advantages, intermediate endpoints are not always a simple or reliable tool for evaluating the benefit-to-risk ratio of interventions, as the relationship between them, definitive endpoints, and the specific intervention under evaluation might be complex (Temple, 1999; Frangakis & Rubin, 2002; Fleming & Powers, 2012). For example, the correlation between intermediate and definitive endpoints may not always be perfect and changes in intermediate endpoints may not reliably predict changes in definitive endpoints. In particular, intermediate endpoints may show short-term benefits, but the long-term impact on definitive endpoints may differ. Moreover, interventions may have unintended effects on intermediate endpoints that do not necessarily translate into meaningful clinical benefits. Furthermore, the relationship between intermediate and definitive endpoints may vary among different subgroups in diverse patient populations.

As a result of this complex relationship between intermediate and definitive endpoints, there are several instances where inappropriate intermediate endpoints have provided misleading information (Echt *et al.*, 1991; International Chronic Granulomatous Disease Co-

operative Study Group, 1991). An instance of this is the Normal Hematocrit trial, which involved patients with end-stage renal disease who also have cardiac conditions. These patients often experience anemia (low hematocrit) due to erythropoietin deficiency. In this trial, the intermediate intervention of normalising the hematocrit levels using erythropoietin stimulating agent adversely impacted the overall survival and, in some instances, increased the risk of myocardial infarction. (Besarab *et al.*, 1998). A similar phenomenon was observed in a type 2 diabetes mellitus trial where an intermediate endpoint of a therapeutic strategy providing an absolute 1% reduction in H_bA_{1c} led to elevated mortality rate along with an increased risk of hypoglycemia (The Action to Control Cardiovascular Risk in Diabetes Study Group, 2017). Thus, incorrectly designating an intermediate endpoint as a reliable intermediate could result in endorsing an ineffective drug, steering patients away from potentially more beneficial alternatives. Therefore, accurately distinguishing an intermediate endpoint as a valid or invalid intermediate for definitive outcomes is essential.

A good intermediate endpoint must possess certain qualities that make it a reliable and valid substitute for a definitive endpoint in a clinical trial. Validity, in this context, means that the intermediate should accurately capture the treatment's impact on the definitive endpoint. It should be a valid representation of the underlying disease or condition being studied (Ciani *et al.*, 2021). In fact, changes in the intermediate endpoint should precede or coincide with changes in the definitive endpoint. This temporal relationship ensures that the intermediate provides early indications of treatment efficacy. Key attributes that have been proposed for a good intermediate endpoint are:

- **Strong correlation:** A high correlation between the intermediate endpoint and the definitive endpoint is important. The intermediate should reliably predict the treatment effect on the definitive outcome. This is a commonly held belief among some medical researchers, including Prentice (1989). According to Kunz *et al.* (2017) and Liu *et al.* (2019), it is more beneficial to use short-term intermediate information at an interim analysis if the correlation between the intermediate and definitive endpoint is strong.

- **Biological plausibility:** Chataway *et al.* (2011) defined the concept of a "biologically plausible" outcome as an indicator that provides insight into whether the mechanism of action of a test treatment is operating as expected. In effect, there should be a clear and understood biological rationale explaining the relationship between the intermediate and definitive endpoints. This ensures that changes in the intermediate endpoint are likely to reflect changes in the definitive outcome. Hence, the intermediate endpoint should possess the capability to function as a causal pathway to the definitive endpoint (Bratton *et al.*, 2016; Kunz *et al.*, 2017).

- **Clinical relevance:** An intermediate endpoint does not necessarily have to be clin-

ically relevant itself, but it should be closely associated with a clinically relevant outcome. However, in certain instances where the intermediate endpoint becomes a valid surrogate and has a direct impact on the patient's condition, it becomes clinically relevant (Buyse & Molenberghs, 1998). Therefore, the important question that arises is whether we can ascertain that any small observed effect on the intermediate endpoint is meaningful and likely to translate into clinically relevant outcomes (Moll *et al.*, 2006).

- **Measurability:** According to Zhuang & Chen (2020), a quantifiable and well-defined relationship should exist between the intermediate and definitive outcomes. Thus, the intermediate endpoint should be easily measurable and quantifiable. Reliable and consistent measurement methods are essential for the intermediate to be practical in a clinical trial setting.

- **Generalisability:** The relationship between the intermediate and definitive endpoints should be consistent across different patient populations, varying disease stages, and treatment modalities. This means that, following adjustment for the treatment's impact on the intermediate endpoint, the relationship between a time-varying covariate (such as medication dosage which could be adjusted at specific time points based on the participant's health status, response to treatment, or other clinical considerations) and the definitive outcome should remain consistent across different treatment groups (Dafni & Tsiatis, 1998). This makes the intermediate more widely applicable.

- **Regulatory acceptance:** Regulatory agencies often play a role in determining the acceptability of an intermediate endpoint. Endpoints that are recognised and accepted by regulatory authorities enhance the likelihood of successful trial outcomes. In recent times, regulatory bodies such as the European Medicines Agency (EMA) and the US Food and Drug Administration (FDA) have increasingly granted approvals based on intermediate endpoints (Darrow *et al.*, 2020; FDA, 2018). Therefore, the selection of an intermediate should ideally be conducted in a manner that ensures its validity and enhances the likelihood of regulatory approval. This consideration holds significance primarily within regulated settings; many trials conducted in LMICs, for example, may not necessarily seek regulatory approval for their interventions or outcomes.

- **Ethical considerations:** The use of intermediate endpoints should align with ethical standards, ensuring that the substitution does not compromise patient safety or well-being.

When selecting treatments at interim analysis, the definitive outcome could serve as a

guide if it can be measured quickly. Alternatively, if this is not feasible, an intermediate outcome could be considered. This intermediate outcome doesn't need to meet all the criteria of surrogacy; less stringent conditions might suffice (Burzykowski *et al.*, 2005). Particularly, the aim is to prevent instances in which the intermediate outcome demonstrates a negative effect while the definitive outcome would have indicated a positive effect. Such occurrences could result in the erroneous exclusion of a relevant arm from consideration. In terms of the operational characteristics of the seamless phase II/III design, both the correlation between the intermediate and definitive outcomes at an individual patient level and the treatment effects at a population level are important factors to consider (Friede *et al.*, 2020).

2.1.2 Proposed design

In Chapter 1, I described how the drop-the-losers design can be a very sample-size-efficient approach to evaluating multiple experimental interventions. Wason *et al.* (2017) provide very helpful results for designing such trials, including a simple method for strongly controlling the type I FWER. This method is specific though to the case where a single endpoint is used throughout the trial. This limits the potential application of this approach in a seamless phase II/III setting, where it would be more likely an intermediate outcome is to be used for faster treatment selection.

Therefore, in this chapter, I extend the drop-the-losers design described by Wason *et al.* (2017) to make it more suitable to a seamless setting by allowing for the use of an intermediate endpoint for treatment selection. More specifically, I consider a trial design that integrates Phase II and phase III into a single, uninterrupted study with an interim analysis. In the phase II component, patients will be randomised to two treatment arms as well as a control arm. At the interim analysis, one of these treatment arms will then be selected to continue to the phase III component, for formal powered comparison to the control arm. The choice between the treatment arms will be based on evidence regarding efficacy using an intermediate outcome. The final phase III analysis will then be based on comparative efficacy established using a definitive outcome.

Expanding the methodology in Wason *et al.* (2017), my proposals accommodate varying treatment effects across the intermediate and definitive endpoints and account for the correlation between them. Sub-cases will address situations in which this correlation is treated as known or unknown. Furthermore, given that the decision to drop an arm is heavily dependent on the quality and quantity of available information at the interim analysis, I explore the optimal timing for conducting the interim analysis to minimise the total required sample size.

The remainder of the chapter is organised as follows. In the next section, I introduce the methodology underpinning the drop-the-loser design. Section 2.2.2 then covers the FWER

and power of the considered design. In Section 2.2.3, I provide information relating to a motivating example. The results and subsequent discussion are then detailed in Section 2.3 and Section 2.4 respectively.

2.2 Methodology

This section presents the methodology underpinning the drop-the-loser design considered in this chapter. It explains the setting, high-level design assumptions, and the criteria for decision-making. Additionally, it provides detailed information about the distributional assumptions regarding the test statistics, the approach to FWER control, and the example used in the Results.

2.2.1 Notation, hypotheses, and test statistics

Consider a trial involving a control and two experimental treatments, with two stages. I explore the scenario where an intermediate outcome may be utilised in the first stage and a definitive outcome in the second stage of the trial. For simplicity, both endpoints are assumed to be normally distributed, though extension to alternative distributional forms follows easily using asymptotic results (Jaki & Magirr, 2013). The design incorporates a fixed sample size in each stage, while also assuming equal allocation between the present arms in both stages. It is further assumed that subjects from both phases are used in the final decision-making process.

To facilitate these assumptions, assume that at the end of stage j there are N_j patients, indexed $i \in \{1, \ldots, N_j\}$, with data on treatment arm k. Let $k = 0$ denote the control arm and $k \in \{1, 2\}$ denote the experimental arms. We assume patients $i \in \{1, \ldots, N_1\}$, from stage $j = 1$, provide outcomes Y_{ikI} and Y_{ikD} for the intermediate and definitive outcomes respectively. By contrast, patients $i \in \{N_1 + 1, \ldots, N_2\}$, from stage $j = 2$, need only provide outcome Y_{ikD} as the intermediate outcome is not used in the final analysis. Additionally, we assume $Y_{iko} \sim N(\mu_{ko}, \sigma_o^2)$ for $o \in \{I, D\}$, with σ_I and σ_D treated as known. To account for the correlation between outcomes from the same patient on the two outcomes in what follows, we set $\text{Corr}(Y_{ikI}, Y_{ikD}) = \rho$.

The null hypotheses are then defined as $H_k : \delta_{kD} = \mu_{kD} - \mu_{0D} \leq 0$, for $k \in \{1, 2\}$. Subsequently, I will refer to the situation where the set of k treatment effects $\delta_{1D} = \delta_{2D} = \delta_{1I} = \delta_{2I} = 0$ as the global null hypothesis, H_G, for the FWER control. One logical approach to weak FWER control is when you are at the boundary of all of the null hypotheses, which is 0. Therefore, it stands to reason that as you approach 0, the FWER should increase or be maximised. To select a treatment arm at the interim analysis, and to test the null hypotheses, we employ Wald test statistics:

$$Z_{jko} = \frac{\hat{\delta}_{jko}}{\sqrt{\text{Var}(\hat{\delta}_{jko})}} = \frac{\frac{1}{N_j}\sum_{i=1}^{N_j} Y_{iko} - \frac{1}{N_j}\sum_{i=1}^{N_j} Y_{i0o}}{\sqrt{\frac{2\sigma_o^2}{N_j}}}.$$

Extending Wason *et al.* (2017), evaluating the operating characteristics of the proposed design then depends on knowing the joint distribution of $Z = (Z_{11I}, Z_{12I}, Z_{21D}, Z_{22D})^\top$. As demonstrated by Law *et al.* (2020), this is

$$\begin{pmatrix} Z_{11I} \\ Z_{12I} \\ Z_{21D} \\ Z_{22D} \end{pmatrix} \sim MVN \left\{ \begin{pmatrix} \delta_{1I}\sqrt{\frac{N_1}{2\sigma_I^2}} \\ \delta_{2I}\sqrt{\frac{N_1}{2\sigma_I^2}} \\ \delta_{1D}\sqrt{\frac{N_2}{2\sigma_D^2}} \\ \delta_{2D}\sqrt{\frac{N_2}{2\sigma_D^2}} \end{pmatrix}, \begin{pmatrix} 1 & \frac{1}{2} & \rho\sqrt{\frac{N_1}{N_2}} & \frac{\rho}{2}\sqrt{\frac{N_1}{N_2}} \\ \frac{1}{2} & 1 & \frac{\rho}{2}\sqrt{\frac{N_1}{N_2}} & \rho\sqrt{\frac{N_1}{N_2}} \\ \rho\sqrt{\frac{N_1}{N_2}} & \frac{\rho}{2}\sqrt{\frac{N_1}{N_2}} & 1 & \frac{1}{2} \\ \frac{\rho}{2}\sqrt{\frac{N_1}{N_2}} & \rho\sqrt{\frac{N_1}{N_2}} & \frac{1}{2} & 1 \end{pmatrix} \right\}.$$

2.2.2 Family-wise error rate and power

When performing multiple statistical tests (e.g., comparing multiple treatment arms), the probability of making at least one type I error, referred to as the FWER, will increase unless a statistical adjustment is applied. In a seamless design setting, it will often be the case that the FWER must be controlled to some nominal α level, to limit the chance of spuriously recommending an ineffective experimental treatment. Before describing the approach that was therefore taken in this chapter to control the FWER, I first describe some alternative multiple correction methods that were considered for this study.

The first, and most common, approach to FWER control is *Bonferroni's* correction, where the significance threshold for each individual test is adjusted to be more stringent by dividing the desired overall significance level (e.g., $\alpha = 0.05$) by the number of tests being conducted ($\alpha^* = \alpha/K$). Bonferroni's correction is simply to apply but is known to in many instances be conservative. To reduce this conservatism, alternative methods do exist that require limited, if any, additional assumptions. These include the *Bonferroni-Holm* and *Šidák's* corrections. None of these methods though leverage known correlations between test statistics; in the case of the considered drop-the-loser design, the correlation between the test statistics has a known distributional form. To attempt to leverage this, I considered FWER control using a generalised version of *Dunnett's* correction.

To understand the treatment effect scenarios for which FWER control is considered in this work, we first define a general power function for the probability that H_1 or H_2 is rejected. To specify this power function, we must nominate our rules for treatment selection and hypothesis rejection. We assume that treatment arm $k \in \{1, 2\}$ is selected for continuation to stage 2 of the trial if $Z_{1kI} > Z_{1k'I}$ for $k' \neq k$. That is, in this study, we assume the experimental arm with the highest test statistic for the intermediate outcome

is selected at the interim analysis. Then, H_k is subsequently rejected if $Z_{2kD} > c$, for critical rejection threshold c. The probability of H_1 being rejected is thus given by

$$P_1(\delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c) = \mathbb{P}(Z_{11I} > Z_{12I}, Z_{21D} > c \mid \delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c).$$

Note that here, for brevity, we have written this probability conditional only on the parameters that will be used in describing how the FWER is controlled; the probability is also conditional on σ_I and σ_D.

We can compute this probability using an affine transformation approach, as described by Wason *et al.* (2017). Specifically, let

$$A_1 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Then, $Y = (Y_1, Y_2)^\top = AZ \sim MVN\{A_1\mathbb{E}(Z), A_1\text{Cov}(Z)A_1^\top\}$, where Y_1 is the difference between the interim test statistics and Y_2 is the test statistic in the second stage and

$$\begin{aligned}\mathbb{P}(Z_{11I} > Z_{12I}, Z_{21D} > c \mid \delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c) \\ = \mathbb{P}(Y_1 > 0, Y_2 > c \mid \delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c).\end{aligned}$$

Thus, we have

$$\begin{aligned}P_1(\delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c) &= \mathbb{P}(Y_1 > 0, Y_2 > c \mid \delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c), \\ &= \int_0^\infty \int_c^\infty \phi\{(y_1, y_2)^\top, A_1\mathbb{E}(Z), A_1\text{Cov}(Z)A_1^\top\}\, \text{d}y_2\text{d}y_1.\end{aligned}$$

Here, $\phi(y, \mu, \Sigma)$ is the probability density function of an $MVN(\mu, \Sigma)$ distribution evalauted at y. Defining

$$A_2 = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

we similarly have

$$\begin{aligned}P_2(\delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c) &= \mathbb{P}(\text{Reject } H_2 \mid \delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c), \\ &= \int_0^\infty \int_c^\infty \phi\{(y_1, y_2)^\top, A_2\mathbb{E}(Z), A_2\text{Cov}(Z)A_2^\top\}\, \text{d}y_2\text{d}y_1.\end{aligned}$$

Note that the case where the definitive outcome is used at the interim analysis, as in Wason *et al.* (2017), can be recovered in the above by simply setting $I = D$ and $\rho = 1$.

Family-wise error rate control

Note that using the above, and the fact that only one null hypothesis is ever tested in stage 2, we can compute the family-wise error rate for any set of treatment effects as

$$FWER(\delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c) = \mathbb{I}(\delta_{1D} \leq 0) P_1(\delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c) + \mathbb{I}(\delta_{2D} \leq 0) P_2(\delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c).$$

In general, for particular choices of N_1 and N_2, one then might wish to choose the rejection parameter c to achieve *strong* control of the FWER, i.e., control of the FWER to some level α regardless of the values of the parameters δ_{1I}, δ_{2I}, δ_{1D}, δ_{2D}, and ρ. Or, in equation form, choosing c such that

$$\max_{\delta_{1I} \in \mathbb{R},\ \delta_{2I} \in \mathbb{R},\ \delta_{1D} \in \mathbb{R},\ \delta_{2D} \in \mathbb{R},\ \rho \in [0,1]} FWER(\delta_{1I}, \delta_{2I}, \delta_{1D}, \delta_{2D}, \rho, N_1, N_2, c) \leq \alpha.$$

For the case where the definitive outcome is used for treatment selection (i.e., where $I = D$ and $\rho = 1$), Wason *et al.* (2017) demonstrate this can be achieved by ensuring

$$FWER(0, 0, 0, 0, 1, N_1, N_2, c) \leq \alpha.$$

In the case where an intermediate outcome is used for treatment selection, we have not been able to formally prove which combinations of values of δ_{1I}, δ_{2I}, δ_{1D}, δ_{2D}, and ρ maximise the FWER. However, our observations (see, e.g., Bratton *et al.*, 2016) indicate it may be when $\delta_{1D} = \delta_{2D} = 0$ and $\rho = 1$, with the values of δ_{1I} and δ_{2I} immaterial as both treatments are then ineffective for the definitive outcome as so it does not matter which is brought forward. We explore this hypothesis later.

Alternatively, strong control of the FWER might be considered to be an unnecessarily strict requirement. This may be particularly true for non-regulated trials, or when the relationship between the intermediate and definitive outcomes is better understood. Following this logic, we consider design under two weaker sets of scenarios for FWER control: each assumes the global null hypothesis H_G is true. They then vary in how they treat ρ

- ***ρ* treated as known:** If the relationship between the endpoints is well-understood, one might opt to choose c as the minimal value such that

  $$FWER(0, 0, 0, 0, \rho, N_1, N_2, c) \leq \alpha,$$

  for a particular assumed value of ρ.

- ***ρ* treated as unknown:** To ensure FWER control across possible values of ρ

under H_G, we can alternatively choose c such that

$$\max_{\rho\in[0,1]} FWER(0,0,0,0,\rho,N_1,N_2,c) \leq \alpha.$$

In our explorations (see, e.g., Jennison & Turnbull, 2006; Magirr *et al.*, 2012; Jaki & Magirr, 2013; Wason *et al.*, 2017), this has been observed to be found equivalent to requiring

$$FWER(0,0,0,0,1,N_1,N_2,c) \leq \alpha,$$

which relates to our hypothesis above that $\rho = 1$ for the scenario in which the FWER is maximised.

Power

A primary objective of this chapter is to identify the optimal timing of the interim analysis to minimise the required sample size and thus reduce resource burden while controlling power to a desired level.

To achieve this, we assume that $N_1 = N\theta/3$ and $N_2 = N_1 + N(1-\theta)/2$, such that there are $N\theta$ patients in stage 1 and then $N(1-\theta)$ patients in stage 2, giving N as the total sample size for both stages. Thus, the timing of the interim is reflected in the parameter θ. Note that $\theta = 0.5$ would reflect equal sample sizes in the two stages, while $\theta = 0.6$ reflects equal sample sizes per (present) arm in each stage.

Observe that under the global null hypothesis H_G, the distribution of the test statistics depends on N_1 and N_2 only through their ratio, N_1/N_2. Based on the above restrictions, this ratio depends only on θ and not on N. Thus, for any given θ, we may first find the value $c = c(\theta)$ that controls the FWER according to the particular requirements stipulated in the previous subsection. We then choose N as the minimal value such that

$$P_1(\tau_{I1},\tau_{I0},\tau_{D1},\tau_{D0},\rho,N\theta/3,N\theta/3+N(1-\theta)/2,c) \geq 1-\beta,$$

for some nominated value of ρ, and where $\tau_{s1} > \tau_{s0}$ for $s \in \{I, D\}$. Here, τ_{s1} and τ_{s0} represent the interesting and the uninteresting assumed treatment effects respectively. That is, we choose N to control the probability of rejecting H_1, assuming treatment 1 has a stronger effect on both the intermediate and definitive outcomes. We do not require power control over all ρ, as we assume this would be considered too conservative in practice. Thus a fixed ρ will always be assumed for the power calculation in designing the trials below, but this fixed ρ may or may not be leveraged in the FWER control requirement.

2.2.3 Motivating example

The TAILoR trial is utilised as a practical example to illustrate the research findings. However, certain adjustments were made to align it with the framework of a seamless phase II/III design. The trial originally aimed to assess the effects of four different doses of Telmisartan, which is believed to mitigate insulin resistance in individuals with HIV in combination with antiretroviral therapy. The primary goal of the trial was to measure the reduction in insulin resistance in the groups receiving Telmisartan compared to the control group, at the 24-week mark. The design strategy controlled the FWER to the $\alpha = 0.05$ level, while ensuring 90% power ($\beta = 0.1$) for $\tau_{D1} = 0.545$, $\tau_{D0} = 0.178$, and $\sigma_D = 1$. These parameter values are assumed for computing all designs below.

No assumptions were made by TAILoR regarding an intermediate selection outcome, as an intermediate outcome was not used in this way. Nonetheless, we assume fasting glucose levels were employed as the intermediate outcome since elevated levels within the normal range can be associated with insulin resistance. For simplicity, we focus on the case where $\delta_{I1} = \delta_{D1}$ and $\delta_{I0} = \delta_{D0}$, though we explore the impact of the values of the means for the intermediate outcomes more generally.

2.3 Results

The results are presented in four sections, in the following order: the impact of the correlation ρ on the required sample size, the optimal timing of the interim analysis, an evaluation of the influence of the treatment effects on the FWER, and the efficiency gain of the seamless design in comparison to a simpler non-adaptive approach to trial design in this three-arm setting.

2.3.1 Impact of ρ on required sample size

A key question in practice when considering the use of the explored design might be the impact of the value of ρ assumed in the power (and also potentially the FWER) calculation. Sensitivity to this ρ may cause concern, as it may imply a heightened risk of an over- or under-powered trial. Therefore, in this section, I evaluate the impact of the assumed ρ on the required sample size. This is done for the two considered requirements for FWER control (treating ρ as known or unknown), in the situation where $\theta = 0.6$. The results are presented in Figure 2.2.

The results show an inverse relationship between the level of correlation and the required sample size when ρ is treated as unknown in the FWER control requirement. Specifically, as the level of correlation between the intermediate and the definitive increases, the sample size required decreases. This finding is consistent with previous literature that has

asserted that when there is a perfect correlation between these two endpoints, it often leads to a reduction in the required sample size.

On the contrary, if the correlation is assumed to be known in the FWER control requirement, the critical threshold c increases in this assumed value of ρ. This results in a larger required sample size for increased ρ. I.e., the power gain (for any fixed rejection rule) from larger ρ is outweighed by the power loss caused by the increase in c for larger ρ. Observe though that assuming a known ρ in the FWER control requirement always results in a lower required sample size than assuming ρ is unknown in the FWER control requirement, for any given value of ρ assumed in the power calculation.

Note also that across both lines in Figure 2.2, we can see that the variation in the required sample size is relatively low, even when ρ is altered from 0 to 1. I return to this point in the Discussion.

2.3.2 Optimal timing of the interim analysis

Given that the reliability of the decision to discontinue a treatment arm is strongly influenced by the quality and quantity of information at the interim analysis, I now proceed to vary the timing of the interim analysis and assess the impact on the required sample size. In particular, this evaluation aimed to identify the optimal timing that yields the smallest required sample size for a specified power. To do this, I explored a wide range of interim analysis timings, defined by $\theta \in [0.25, 1)$, considering four cases where the correlation between the intermediate and definitive endpoints was given by $\rho \in \{0.25, 0.5, 0.75, 1\}$. Once again, I considered cases where this ρ is treated as known or unknown for the purposes of controlling the FWER. The results are presented in Figure 2.3.

In general, the plots exhibit approximately U-shaped curves, suggesting that the lowest sample size is observed somewhere in the middle of the considered range of θ. This should not be surprising, as with a very small sample before the interim analysis, the interim may not reliably capture information for an effective selection, leading to a loss in efficiency. Contrarily, with a very long interval before the selection, the trial may lose out on the opportunity to effectively direct resources towards a much better-performing arm, again affecting efficiency. Thus, the approximately U-shaped curve indicates that there is a spot (optimal time for the interim analysis) where the trial benefits most from reliably adapting the design.

When the correlation between the two endpoints ρ is treated as unknown in the FWER control requirement, the optimal time for the interim analysis appears to be when approximately 65% ($\theta = 0.65$) of patients have been allocated. This is true across the considered values of ρ. By contrast, when the correlation ρ is assumed known in the FWER control requirement, the optimal timing of the interim analysis depends on the specific assumed value of ρ. Specifically, the larger the assumed value of ρ the earlier the optimal timing of

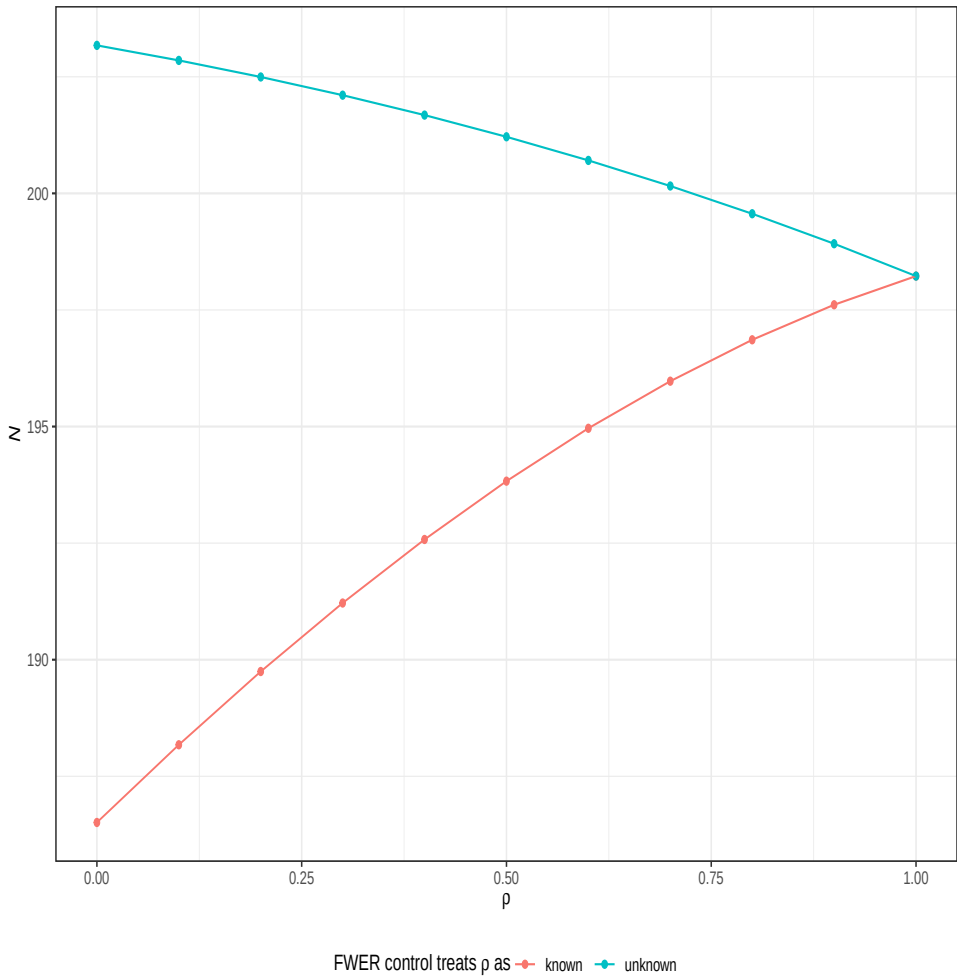

Figure 2.2: Shows the required sample size as a function of the assumed value of ρ in the power calculation. The blue line is the case where ρ is treated as unknown in the FWER requirement and the red line is the case where it is, like in the power requirement, treated as known in the FWER requirement. Equal allocation of sample size to each arm in both stages is assumed ($\theta = 0.6$).

the interim analysis (i.e., smaller θ becomes optimal). Observe from Figure 2.3 that when $\rho = 0.25$, the optimal timing of the interim analysis occurs at $\theta = 0.75$, and for $\rho = 1$, the optimal timing of the interim analysis occurs at $\theta = 0.65$.

2.3.3 Family-wise error rate control

In this section, I explore the effect of the values of δ_{1I}, δ_{2I}, δ_{1D}, δ_{2D}, and ρ on the FWER. It is logical based on much previous research to expect that the FWER will be maximised when $\delta_{1D} = \delta_{2D} = 0$. I therefore focus on varying the values of δ_{1I}, δ_{2I}, and ρ. Understanding the influence of ρ on the FWER may be anticipated to be critical based on the common belief that the correlation between the intermediate and definitive

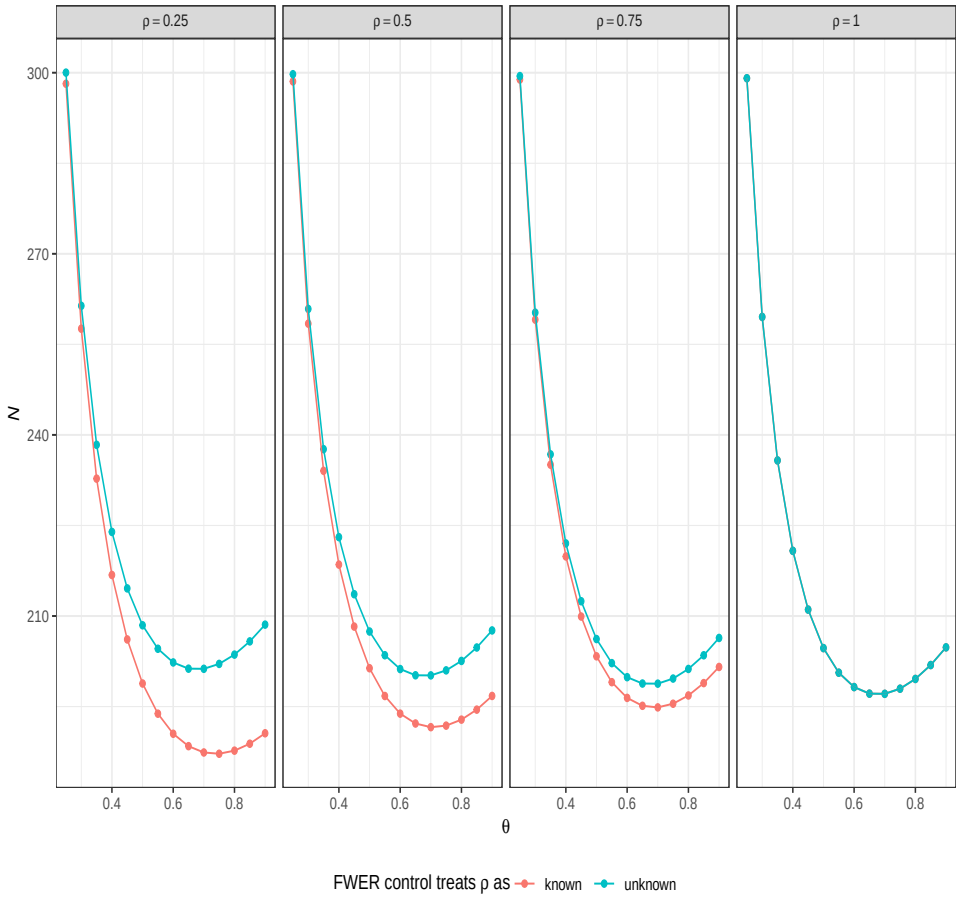

Figure 2.3: Impact of the timing of the interim analysis (as defined by θ) on the required sample size, for varying correlations ρ between the intermediate and definitive outcomes for cases where ρ is treated as known and unknown in the FWER control requirement.

outcomes plays a crucial role in the context of a seamless design; values of ρ were selected to represent cases where the correlation between the intermediate and definitive endpoint was weak through perfect, namely $\rho \in \{0.25, 0.5, 0.75, 1\}$. I then explore the grid with $(\delta_{1I}, \delta_{2I} \in [-2\tau_{D1}, 2\tau_{D1}] = [-2(0.545), 2(0.545)]$. As the principal interest lies in whether modification of the values of δ_{1I}, δ_{2I}, and ρ causes an increase or decrease to the FWER, I simply fix $c = \Phi^{-1}(0.975)$, $\theta = 0.6$, and $N = 200$, with N chosen to reflect the values observed in Figure 2.2. The results are presented in Figure 2.4.

Generally, the results show that for given values of δ_{1I} and δ_{2I}, as the correlation between the intermediate and definitive endpoints strengthens, the FWER increases. This is because a weak correlation between the two endpoints means that they are less aligned, and thus, the interim analysis may yield results that do not strongly predict the results of the final analysis. More formally, using properties of the multivariate normal distribution, conditional on the observed value of Z_{1kI}, the expectation of Z_{2kD} increases as a function

of ρ. Thus, larger ρ is associated with an increased likelihood of committing a family-wise error.

Note that for a given value of ρ the FWER is largest when $\delta_{1I} = \delta_{2I}$. This is unsurprising, because of the assumption that $\delta_{1D} = \delta_{2D} = 0$: as both definitive endpoint means are assumed to be zero, the situation in which $\delta_{1I} = \delta_{2I}$ maximises the ability to select whichever of the two experimental treatments is experiencing a random high, to the knock-on effect of an increased FWER.

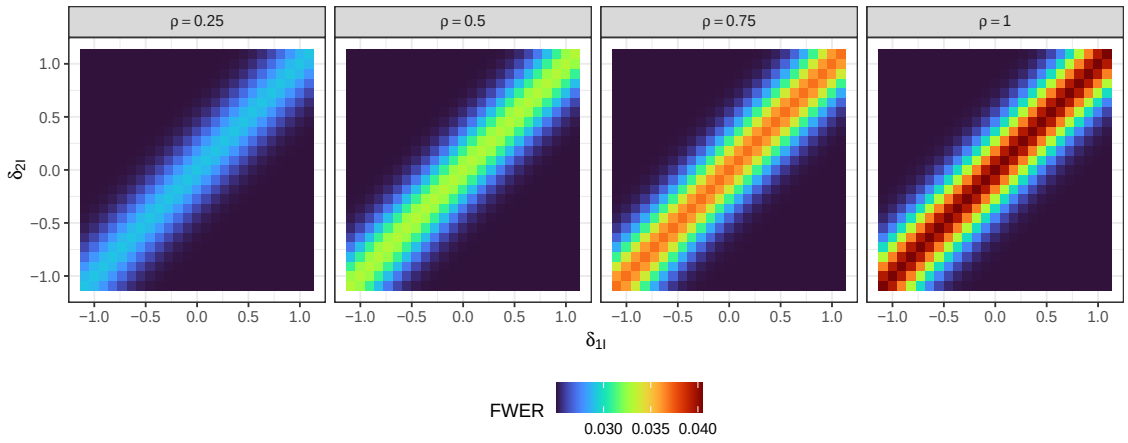

Figure 2.4: Shows the FWER as a function of the assumed value of $\rho \in \{0.25, 0.5, 0.75, 1\}$ and $\delta_{1I}, \delta_{2I} \in [-2\tau_{D1}, 2\tau_{D1}] = [-2(0.545), 2(0.545)]$
. Here, $N = 200$ and equal allocation of sample size to each arm in both stages is assumed ($\theta = 0.6$).

2.3.4 Comparison of adaptive and non-adaptive sample sizes

It has been established in the literature that a primary advantage of the seamless phase II/III design is an efficiency gain through a reduction in the required sample size. In this section, I quantify one such type of efficiency gain by evaluating the percentage decrease in the required sample size using the seamless phase II/III design when compared against

a standard three-arm trial without an interim analysis. This comparison is an important one, as this non-adaptive approach could be considered a viable alternative in practice if the seamless design does not provide a strong efficiency gain, owing to its increased simplicity. The definitive outcome assumptions, $\tau_{D1} = 0.545$, $\tau_{D0} = 0.178$, and $\sigma_D = 1$, from the motivating example, were utilised in the multiarm developed by Grayling & Wason (2020) to compute the sample size required by the non-adaptive three-arm design. A significance level of $\alpha = 0.05$ was used, in combination with Dunnett's correction, assuming 1:1:1 randomisation. The identified design was found to require a sample size of 207.

With regards to the seamless design, $\theta = 0.75$ was assumed as an efficient option when ρ is treated as known, and $\theta = 0.65$ was assumed as an efficient option when ρ is treated as unknown in the FWER control requirement, as discussed in section 2.3.2. Furthermore, $\rho = 0.5$ was assumed as a suitably conservative value that may reflect realistic trial design scenarios.

The results in Table 2.1 show a 7.8% and 3.5% reduction in the required sample size for the seamless phase II/III designs compared to the non-adaptive approach when the correlation between the intermediate and the definitive endpoints is treated as known and unknown respectively in the FWER control requirement. Generally, the seamless design results in a relatively small reduction in the required sample size, but they significantly decrease the number of patients for whom the definitive outcome needs to be measured. Depending on the nature of the trial, these reductions may reflect a sizeable benefit in terms of reduced required resources, particularly because patients in the dropped arm do not require the measurement of the definitive outcome. To quantify this, a separate calculation is needed to determine the precise extent of the reduction.

Note though that part of this observed efficiency gain arises from the inherent nature of the drop-the-loser design, which is limited in its ability to identify only one effective arm, irrespective of the effectiveness of both arms. By contrast, the non-adaptive three-arm design can reject both null hypotheses.

Non-adaptive sample size	**Seamless design when ρ is known**		**Seamless design when ρ is unknown**	
	Sample Size	Percentage reduction (%)	Sample Size	Percentage reduction (%)
207	192	7.8	200	3.5

Table 2.1: Comparison of the seamless design and the single-stage multi-arm trial for $\alpha = 0.05$, $\beta = 0.1$, $\tau_{D1} = 0.545$, $\tau_{D0} = 0.178$, $\tau_{I1} = 0.545$, $\tau_{I0} = 0.178$, $\rho = 0.5$, and $\sigma_I = \sigma_D = 1$.

2.4 Discussion

In this study, a framework for the design of a seamless phase II/III drop-the-loser design was proposed. This framework may help expedite the drug development process by eliminating the time-lag that would exist if separate phase II and phase III trials were conducted, with it focused on the situation where one of two treatment arms is to be selected for further testing at an interim analysis based on an intermediate outcome measure. Another notable advantage of this design, as evidenced in both the results and existing literature, is the reduced fixed required sample size compared to a traditional multi-arm trial. The fixed sample size inherent in this design also addresses the uncertainties associated with the variable sample size required by other designs, such as group sequential multi-arm multi-stage designs. This reduction in uncertainty contributes to easing challenges related to securing funding and managing the logistical aspects of the trial.

Despite the advantages of the proposed design, it has some practical issues. First, adhering to the requirement of dropping a treatment at the end of the first stage may be challenging in practice. For instance, if both experimental treatments demonstrate strong performance compared to the control, it might be considered inappropriate to drop an arm.

Second, the drop-the-loser design, like other adaptive designs, faces the limitation of outcome delay. Since adaptation requires a measurable parameter, delayed responses result in fewer observations at interim analyses, lowering the probability of identifying optimal treatments. This limitation has the potential to compromise the trial's efficiency by recruiting patients into arms that are later dropped before assessing their responses. According to Wason *et al.* (2017), the extent of efficiency loss depends on the trial's recruitment rate and the timing of the endpoint measurement(s). The incorporation of an intermediate outcome into this proposed design, which inherently assumes the presence of a quickly measurable factor for adaptation, addresses the limitation of outcome delay to a large extent. This assertion is backed up by Hampson & Jennison (2013), who argued that the efficiency loss due to outcome delays can be mitigated by leveraging data on correlated short-term endpoints. I have not here though evaluated how fast the intermediate outcome needs to be measured for the proposed design to be useful.

It was observed from the results that the change in required sample size from no correlation ($\rho = 0$) to perfect correlation ($\rho = 1$) between the intermediate and the definitive endpoint was small. This underscores a crucial point that precise knowledge of the correlation is not essential. Further, it highlights that a large value of ρ does not imply that an effect on the intermediate endpoint will be accompanied by an effect on the definitive endpoint. This assertion is emphasised by Baker & Kramer (2003) who demonstrated that

having a perfect correlation between a potential intermediate and an unobserved definitive outcome does not guarantee accurate inference based on this potential intermediate endpoint. This is because ρ does not indicate the relationship between the means of the two endpoints. Hence, from a design perspective, the quality of available data to inform assumptions about ρ may not significantly impact the design.

While we acknowledge that achieving a perfect correlation between the intermediate and definitive treatment effects is rare due to various influencing factors, I present a hypothetical scenario where such a correlation may exist. Consider a clinical trial evaluating a new medication's efficacy in reducing blood pressure. The intermediate endpoint may be the reduction in systolic blood pressure (SBP) after one month of treatment, while the definitive endpoint could be the SBP reduction after six months. If the new medication consistently and perfectly lowers SBP after one month, it is probable that it will also lower SBP after six months, resulting in a perfect correlation between the reductions at these time points.

The results also demonstrated that conducting an interim analysis when 65% ($\theta = 0.65$) of patients have been allocated for cases where ρ is treated as unknown in the FWER control requirement is highly efficient, often even optimal. The closely related work regarding the optimal timing of the interim analysis is that of Walter *et al.* (2020), whose systematic review revealed that the majority of trials scheduled their interim analysis around the midpoint of the data collection. Although the trials analysed by Walter and colleagues did not specifically target seamless or drop-the-loser designs, their findings still hold statistical and non-statistical significance in this context. This is because such strategic practice of conducting an interim analysis around the midpoint of a trial offers other advantages in terms of relatively early identification of issues with the study. By contrast, conducting an interim analysis too early or too late in a clinical trial can have various disadvantages. According to Thorlund *et al.* (2018), conducting an interim analysis too early can be problematic because limited data are prone to increased type I error rates, unreliable results, ineffective adaptations, and ethical concerns. This can consequently result in premature termination of an arm without sufficient evidence. On the other hand, delaying the interim analyses diminishes the trial's efficiency (Huang, 2016) due to missed opportunities for adaptation, prolonged trial timelines, increased costs, and potential patient exposure to ineffective treatments. Thus, conducting an interim analysis around the midpoint of a trial may in general be recommended as an effective approach on statistical and non-statistical grounds.

We observed that, at least for the considered design scenario, the FWER appeared to be maximised when $\delta_{1I} = \delta_{2I}$ and $\rho = 1$. This is a useful finding, as it means the rejection threshold computed for the case where the definitive outcome is used for treatment selection, provided by Wason *et al.* (2017), may still provide strong control of the FWER

in the case where an intermediate outcome is used. However, we note we have not formally proved this to be the case, and if guaranteed strong FWER control is required, this would currently necessitate shifting towards an approach involving p-value combination or a similar methodology.

Finally, note that the methodology detailed in this chapter relied on a normal approximation. Future research could focus on exploring joint distributions for different types of outcome data. For instance, the intermediate outcome could take a binary form, indicating response or non-response, with then fully specified distributions given for the definitive outcome, conditional on the value of the intermediate outcome. While a normal approximation may work well in general, a higher level of precision of the true operating characteristics may well be achievable through the adoption of a more sophisticated modelling framework.

Chapter 3

A review of approaches to specifying the intra-cluster correlation and other design parameters

In the clinical trial process, sample size plays a critical role in determining how many participants are required to adequately power the trial. Although it is regarded as best practice to calculate the sample size at the beginning of the trial, this estimation requires knowledge of certain key parameters that are unknown at the design stage. These parameters include the SD, the effect size, and the ICC when estimating the sample size for a CRT. In this chapter, we review the approaches to specifying these parameters used in CRT reports in the HTA journal. We further highlight the justifications underscored by the researchers for their adopted approaches.

3.1 Introduction

As indicated in Chapter 1, CRTs are frequently employed in primary care research, and outcomes from a cluster tend to be more similar (Killip *et al.*, 2004; Campbell & Walters, 2014). To illustrate the factors that account for the similarities between outcomes from individuals within a cluster, let us consider a cluster such as a GP practice. Firstly, people living in the geographical location of the GP practice may have similar socioeconomic characteristics. Therefore, a socioeconomic factor such as deprivation, for example, which is known to impact health outcomes (Theocharidou & Mulvey, 2018; Hawkins *et al.*, 2012) may contribute to the similar disease conditions presented at the GP practice. Moreover, since these patients are likely to be treated by the same health professionals within the

practice, their method of treating a particular disease condition may also result in similar recovery rates or outcomes among patients. As a result of these factors, some homogeneity in health outcomes within the cluster is expected.

The two levels of clustering that are commonly encountered in CRTs are one-level clustering and nested clustering. In one-level clustering, the data is grouped into clusters at a single level. Each observation belongs to one and only one cluster, and there is no hierarchical structure beyond this single level. A study where individuals within different households are the units of analysis, and the households are considered as clusters is a case of one-level clustering. Here, each individual belongs to one household, and there is only one level of clustering. Nested clustering, on the other hand, involves multiple levels of grouping or clustering. Observations are nested within higher-level clusters, and these higher-level clusters, in turn, may be nested within even higher-level clusters, forming a hierarchical structure. Consider a study examining students within classrooms within schools. Students are nested within classrooms, and classrooms are nested within schools. This hierarchical structure constitutes nested clustering. Each student is in one classroom, and each classroom is in one school. For brevity, I shall restrict my attention to one-level clustering throughout this thesis.

The intra-cluster correlation coefficient (ICC) is the primary measure used for the homogeneity between the outcomes from individuals in a cluster. While the definition and interpretation of the ICC may be expressed in a different number of ways depending on the outcome measure, modelling approach, or when dealing with covariates, it is usually the same quantity measured in different contexts (Eldridge & Kerry, 2012; Eldridge *et al.*, 2009). When determining the sample size for a CRT, the ICC becomes an important parameter to prevent erroneous conclusions (Killip *et al.*, 2004), and also becomes useful in the interpretation of the primary outcome (Eldridge & Kerry, 2012). Although other measures of cluster homogeneity exist (e.g., coefficient of variation), the ICC is widely utilised in health services research.

Let us consider a two-arm parallel-group CRT with a continuous outcome, denoted as Y_{ij}, for each participant i in cluster j. A commonly assumed linear mixed-effect model for the outcomes is

$$Y_{ij} = \theta + X_j\mu + c_j + e_{ij}, \tag{3.1}$$

where

- θ is an intercept term,
- $X_j = 1$ if cluster j is allocated to the experimental arm and $X_j = 0$ otherwise,
- $c_j \sim N(0, \sigma_c^2)$ is a random effect for cluster j, and
- $e_{ij} \sim N(0, \sigma_e^2)$ is the individual-level error.

Here, the ICC (ρ) is defined as the ratio of the variation due to the clusters (σ_c^2) and the total (between and within cluster) variation (σ^2); mathematically expressed as

$$\rho = \frac{\sigma_c^2}{(\sigma_c^2 + \sigma_e^2)} = \frac{\sigma_c^2}{\sigma^2}.$$

Considering the ICC formula above, in a study where the within-cluster variance is zero ($\sigma_e^2 = 0$), then individual outcomes within the cluster are identical ($\rho = 1$) while no similarities exist between the outcomes when σ_c^2 or $\rho = 0$. Thus, the value of the ICC typically ranges between 0 and 1, where values closer to unity indicate a strong correlation between the outcomes within a cluster. Researchers have extensively outlined other methods of calculating the ICC for binary and other outcome types (Ridout *et al.*, 1999; Campbell & Walters, 2014; Eldridge *et al.*, 2009); these are considered beyond the scope of this thesis.

While it is rare to observe negative ICCs, it sometimes occurs in practice (see, e.g., Heller *et al.*, 2014). There are some theoretical assumptions or reasons associated with observing a negative ICC. In the study by Heller and colleagues, which sought to improve type 1 diabetes management using dose adjustment for normal eating, the negative ICC could have arisen through uneven dose adjustment among patients. This is a widely held theory among researchers, who believe that negative ICCs can be observed when resources or logistics are limited and unevenly distributed among the clusters (Campbell & Walters, 2014). Ukoumunne (2002) also attributes the plausibility of a negative ICC to sampling error, while Eldridge & Kerry (2012) associates such phenomenon to chance.

Analytically, there are various methods of estimating the ICC. Some of these methods include the one-way ANOVA, Generalised Estimating Equations (GEE), and the linear mixed effect model (Equation 3.1, sometimes referred to as hierarchical models or random effect models). However, these methods of estimating the ICC require observed data, which are typically unavailable at the design stage of the trial for sample size determination. As a result of this difficulty, trialists have adopted some approaches to specifying values for the ICC to assume at the beginning of the trial. In what follows, we discuss three of the most common approaches

- **Published trial reports:** Published trial reports are a valuable source of documentation to obtain ICC values and serve as a useful guide for researchers who might conduct similar trials in the future. According to several studies (Rutterford *et al.*, 2015; Ring *et al.*, 2018), this is a widely adopted method of obtaining ICCs.

- **Pilot and feasibility studies:** This method provides an estimate for the ICC using a small amount of trial data in the setting under consideration. This small study may be internal or external to the main trial. The advantage of this method is that if the pilot trial is conducted in an identical setting to the main trial, it is reasonable

to assume the ICC estimate from it will be better than that from any other available source. However, given that a small amount of data is used to estimate the ICC, the variability around this estimate is often large (Eldridge *et al.*, 2016). Additionally, such pilot evaluations need to be conducted in an often prohibitively timely fashion to allow the result to feed into the main trial.

- **ICC databases or lists:** Some databases exist that have collated ICCs from publication lists of varied interventions and outcomes. Amongst the list include ICCs calculated for schools, hospitals, GP practices, and other organisations. This wealth of ICCs obtained from relevant trials within these databases is sometimes combined through the application of the Bayesian hierarchical model to generate an ICC coefficient estimate for a planned CRT (see, e.g., Tishkovskaya *et al.*, 2023). One such online database is hosted by the University of Aberdeen (`https://www.abdn.ac.uk/hsru/what-we-do/tools/index.php#panel177`.

Other important parameters in sample size calculation are the effect size (or target difference) and the SD. Although the terms target difference and effect size are sometimes used interchangeably, a subtle difference exists between them. The effect size quantifies the magnitude of the difference observed between the treatment groups by providing a standardised measure of the treatment's impact on the outcome of interest. It is typically calculated by taking the difference in means (or other appropriate statistics) between the groups and dividing it by a measure of the variability, such as the SD. Therefore, the effect size becomes useful when interpreting the study results and helps determine the clinical significance of the intervention.

By contrast, the target difference is the predefined difference in outcomes between the treatment groups that researchers aim to detect. It is defined based on the primary objective of the clinical trial and helps guide the sample size calculation to ensure the study has adequate power to detect a meaningful difference if it truly exists. The success of a clinical trial is often assessed based on whether the observed results meet or exceed the predefined target effect and warrant the adoption of a new treatment or intervention over the standard of care. Throughout this thesis, we shall use the effect size.

Similarly to the ICC, obtaining parameters for the effect size and SD can sometimes be problematic. As a result, Cook *et al.* (2018) proposed seven guidance points in choosing the effect size for sample size calculations in RCTs. Their proposed approaches, which included evidence from previous trials and pilot studies, were similar to the approaches for specifying an ICC discussed above. Based on Cook's work, Rothwell *et al.* (2018) conducted a review of trials reported in the HTA journal and found that the most common method of effect size specification was the review of evidence and the use of previous similar studies. This finding has been observed to hold for the SD at which trials are powered as

well (Julious & Owen, 2011).

Previous authors have performed similar evaluations regarding how the effect size was set in great depth (Cook *et al.*, 2018; Rothwell *et al.*, 2018). However, their reviews were within the IRT context. Thus, we are unable to ascertain if those findings hold in the context of CRTs. Within the CRT context, reviews have focused on either the statistical analysis of CRTs (Offorha *et al.*, 2022) or the adherence to the CONSORT extension for CRTs (Han *et al.*, 2019). We are unaware of any study that has extended its scope to review the justifications for the assumed ICC value. Therefore, the role of this review is to effectively demonstrate a specific point: whether high-quality publications in the HTA journal often omit justification for their ICC and other sample size parameters.

Our selection of the HTA journal was premised on the fact that it is the largest programme within the National Institute of Health Research (NIHR), which also happens to be the major funder of health research in the UK (National Institute for Health and Care Reseearch, 2023). Moreover, the HTA bases funding for commissioned and health-related research on a wide range of considerations, but one essential requirement is that researchers provide a detailed report that is published in the HTA journal regardless of the statistical significance of the findings. Therefore, the lack of publication bias (both statistically significant and non-significant findings are published) in the HTA journal makes it ideal for reviews (Rothwell *et al.*, 2018).

In Section 3.2, I describe the methodology for the review which includes the data sources and strategy, inclusion and exclusion criteria, as well as data extraction and synthesis. Descriptive statistics outlining the characteristics of the trials that met the eligibility criteria, and plots showing the impact of the CONSORT statement on ICC review is presented in Section 3.3. In Section 3.4, I discuss the findings of the review, highlighting how the current findings relate to existing knowledge in the field and offer thoughtful reflections on the broader implications of the findings of the review.

3.2 Methodology

This section outlines the search strategy for identifying publications relating to CRTs, including the inclusion and exclusion criteria for the review.

3.2.1 Data sources and review strategy

The following search was run on PubMed on 08/01/21: (((“Health technology assessment” [Journal]) OR (“Health technology assessment reports” [Journal])) OR (“Health technology assessment (Winchester, England)” [Journal])) AND (“cluster”).

Fifty-four returned articles were equally allocated between three reviewers (myself, Michael Grayling and James Wason) to determine inclusion in the data extraction. Specif-

ically, articles were included if they related to the report of a completed CRT. Here, we defined a completed CRT as a trial (i.e., not an observational study) with an appropriate power calculation (i.e., one that should indicate our target parameters had been set values). For included articles, information was then extracted on the values and justification given for the assumed ICC, observed ICC, effect size, and the SD. In cases where an article was unclear in regards to the value/justification for one or more of these parameters, a discussion between the three reviewers was undertaken to resolve the extraction.

Following the completion of data extraction by the three reviewers, I synthesised the extracted data into a unified dataset. I then performed descriptive analyses of the values typically assumed for the ICC in the trial reports, as well as other relevant characteristics which are detailed in the results section of this chapter.

3.2.2 Inclusion and exclusion criteria

The search was principally interested in CRT reports published in English with no restriction on the trial publication date. The publication date of the trial was disregarded for the following reasons: (i) to enable extraction from a larger number of studies for the review, since only one Journal was searched, and (ii) to assess adherence to the CONSORT extension for CRTs pre and post the publication of this reporting guideline. As indicated above, only studies that met the above definition of a completed CRT were included in the review.

Studies that did not meet the eligibility criteria were mainly IRTs that recommended the use of CRTs in future trials. For example, the study by Bonell *et al.* (2015) was a pilot IRT that advocated for the adoption of a CRT design in the definite phase III trial to examine the effectiveness and cost-effectiveness of the intervention. In some cases, the studies retrieved were small feasibility studies (e.g., Koffman *et al.* (2019)), systematic reviews (e.g., Liu *et al.* (2006)), or did not include a formal sample size calculation (eg., Wright *et al.* (2016)). These studies were part of the total returned articles because the search keyword "cluster" was present. In total, 20 (37%) articles were excluded from the review.

3.2.3 Data extraction and synthesis

The required information was extracted using a predefined extraction form after the cluster trial reports from the HTA journal had been selected for inclusion. Of the 34 (63%) articles that met the eligibility criteria, data for further analysis was extracted and stored using an Excel template which all three reviewers could securely access from a shared location. In cases where the desired information could not be located in the report, it was noted with *"Not given"* to indicate that the author(s) did not explicitly state the information of

interest or that it could not be inferred.

The information extracted was captured under three key thematic areas: (i) the characteristics of the trial, (ii) the values of the assumed parameters that were utilised for the sample size calculation, and (iii) the justification provided for the selected assumed parameters. To answer a specific question regarding the difference between the assumed and observed ICC, information on the observed ICC was also extracted, while an additional column was added to note interesting findings, such as consideration of sample size reestimation or making allowance for uncertainty in the assumed values.

3.3 Results

This section presents the findings and analysis of the review guided by the key thematic areas outlined above in the form of frequency distribution tables, descriptive statistics, and graphs. The section concludes with an assessment of the impact of the CONSORT statement on ICC reporting.

3.3.1 Characteristics of the trials

Fifty-four reports were extracted from the HTA journal on 08/01/21, of which 34 met the inclusion criteria. These 34 trials represent a wide range of study settings and are reflective of true practice. Therefore, as a first step in the analysis, I present a general overview of the reviewed trials in the form of frequencies and percentages for all categorical levels of the trial characteristics. Categorical levels with counts of one were recorded under "Others" for brevity. It is worth noting that all of the reviewed trials had standard two-arm parallel group CRT designs. None of the trials were, e.g., stepped-wedge, factorial, or cross-over CRTs. A PRISMA diagram detailing the number of eligible and ineligible studies is presented in Figure 3.1.

A summary of the trial characteristics in Table 3.1 shows that the majority of the trials had continuous primary outcomes (25/34, 74%) or binary outcomes (7/34, 20%), while two trials had other primary outcomes (e.g., time-to-event or composite outcomes). With regards to the trial setting, GP practices constituted the majority (10/34, 29%), followed by schools (3/34, 9%), while nursing homes, communities, and families accounted for 6% of studies each. Moreover, 15 trials, representing 44% of the sample, were conducted in settings other than the aforementioned. In terms of the clinical area, it is unsurprising that primary care recorded the same frequency as GP settings (10/34, 29%), owing to the fact that primary care is generally administered in GP practices. Other clinical areas included mental health (4/34, 12%), public health (3/34, 9%), sexual health (2/34, 6%), and emergency care (2/34, 6%). Concerning the type of intervention, 10 interventions representing 29% of the sample were medical training programs, 2 (6%) were non-medical

training programs, 2 (6%) assessed the effectiveness of health technology, while 20 (59%) interventions were made up of other therapies such as drugs, etc. On the issue of blinding, the majority of the trials were unblinded (25/34, 73%). Regarding the number of clusters, 2 (6%) of the trials had less than ten clusters, 7 (21%) had 10-30 clusters, 4 (12%) had 31-60 clusters, 2 (6%) had 61-90 clusters, 4 (12%) had more than 90 clusters, while 15 (43%) did not state the number of clusters utilised in the trial. With respect to cluster sizes, 3 (9%) of the trials had an average of less than ten participants per cluster, 9 (26%) had an average of 10-30 samples, 2 (6%) trials had an average of 31-60, similarly, 2 (6%) trials had an average of 61-90, and 2 (6%) trials had an average of more than 90 cluster sizes each, while 16 (47%) did not state the average number of cluster size utilised in the trial.

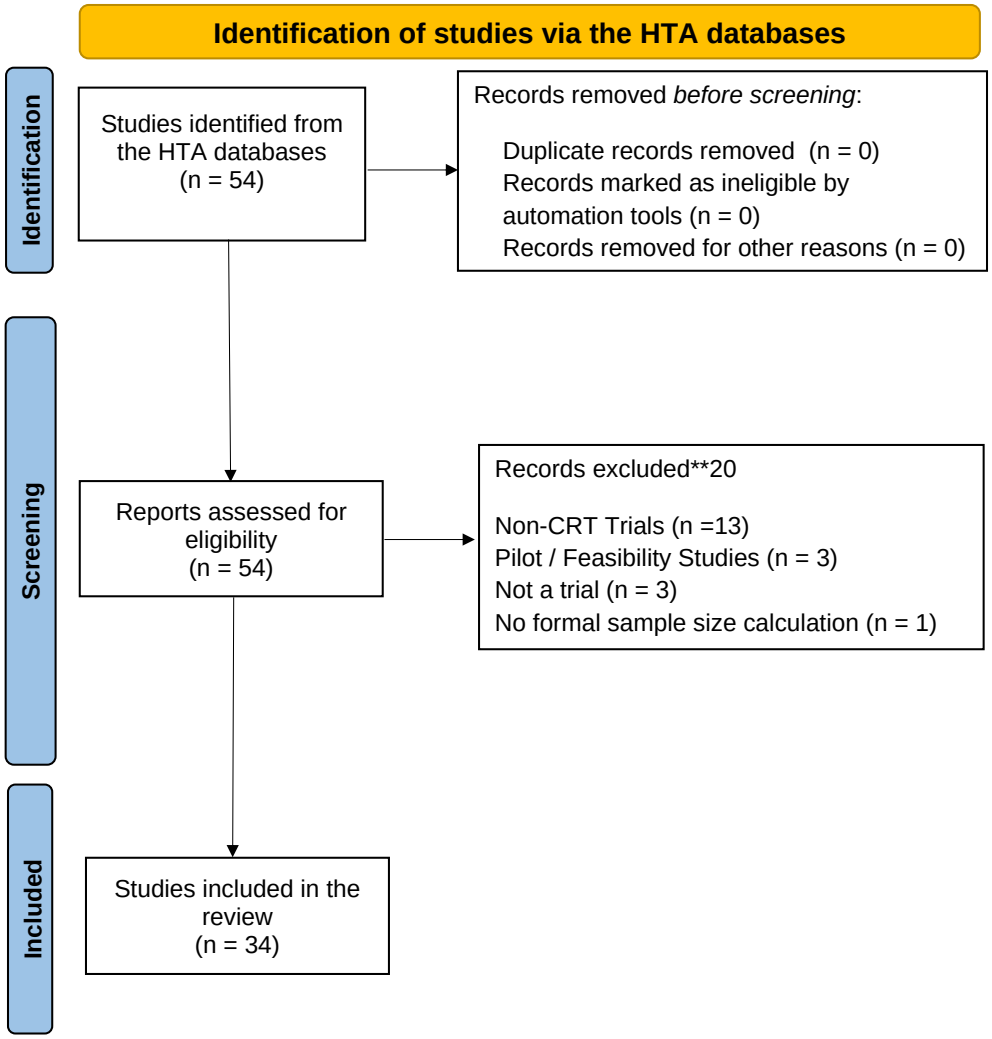

Figure 3.1: PRISMA flow-diagram of articles selected and included in the review.

3.3.2 Descriptive statistics and distribution of the target parameters

Summary statistics on the target parameters are presented in this section. Data is described using measures of central tendency and spread. An emphasis is placed on the difference between the assumed and observed ICC, due to the sensitivity of power to the ICC and the implications of discrepancies on the trial. In addition to summary statistics in Table 3.2, the distribution of the assumed effect size is also displayed in this section through a histogram in Figure 3.2. This was particularly emphasised as the effect size is the sole standardised parameter, while the remaining target parameters were assessed on distinct scales across each trial.

According to Table 3.2, all 34 (100%) eligible trials reported an assumed ICC with a mean and variance of 0.0729 and 0.0111 respectively. The assumed values for the ICC were positively skewed on the range [0.002, 0.5] with a median of 0.05. In terms of the observed ICC for the analysed primary outcome, 26 trials (77%) reported this, with a mean of 0.0445 and a variance of 0.0027. Although the observed ICC was also positively skewed, it was skewed on [0.001, 0.24], with a median of 0.0285. The higher mean and median values of the assumed ICC in comparison to the observed ICC suggest that many of the trials were potentially overpowered. I highlight the implications of an overpowered trial in the discussion section.

Eighteen studies (53%) clearly reported the effect size for which the trial was powered. The mean effect size was 0.33, with a variance of 0.0289. It also had a range of 0.565 [0.005, 0.57], with a median of 0.32. The 16 trials (47%) that reported an SD recorded a mean and variance of 2.30 and 6.46 respectively. The SD was also positively skewed on [0.01, 7.5] with a median of 1.25.

Even trials that utilised pilot studies to calculate the ICC for the main trial acknowledged the estimated ICC from the pilot will be very imprecise (see, e.g., Froggatt *et al.*, 2020). Yet, none of the reviewed trials incorporated uncertainty around the value of the ICC by assuming a prior distribution for this parameter. In addition, few stated the power for a selection of point ICC values (see, e.g., (Campbell *et al.*, 2015)). This is particularly surprising given the frequency with which the ICC was not well understood at the design stage.

3.3.3 Justification for the target parameters

Having presented an overview of the eligible trials with some summary statistics of the target parameters, I now explore the rationale for the selected values of the target parameters in sample size calculations in this section.

Results presented in Table 3.3 reveal that 12 trials (35%) selected the value for the assumed ICC based on previous similar studies. In the absence of a previous study, a value

Characteristics	**Frequency (n)**	**Percentage**
Primary Outcome		
Continuous	25	74
Binary	7	20
Others	2	6
Setting		
GP practice	10	29
Nursing or care homes	2	6
Schools	3	9
Community	2	6
Families	2	6
Others	15	44
Clinical area		
Primary care	10	29
Mental health	4	12
Sexual health	2	6
Emergency care	2	6
Public health	3	9
Others	12	35
Type of intervention		
Medical training program	10	29
Non-medical training program	2	6
Health technology	2	6
Others	20	59
Blinding		
Blinded	9	27
Unblinded	25	73
Number of clusters		
< 10	2	6
10 - 30	7	21
31 - 60	4	12
61 - 90	2	6
> 90	4	12
Not given	15	43
Average cluster size		
< 10	3	9
10 - 30	9	26
31 - 60	2	6
61 - 90	2	6
> 90	2	6
Not given	16	47

Table 3.1: Characteristics of the 34 CRTs under review.

Variable	N	Mean	Variance	Median	Min	Max
Assumed ICC	34	0.0729	0.0111	0.0500	0.0020	0.5000
Observed ICC	26	0.0445	0.0027	0.0285	0.0001	0.2400
Effect size	18	0.3300	0.0289	0.3200	0.0050	0.5700
SD	16	2.3000	6.4600	1.2500	0.0100	7.5000

Table 3.2: Descriptive statistics of the variables of interest and effect size.

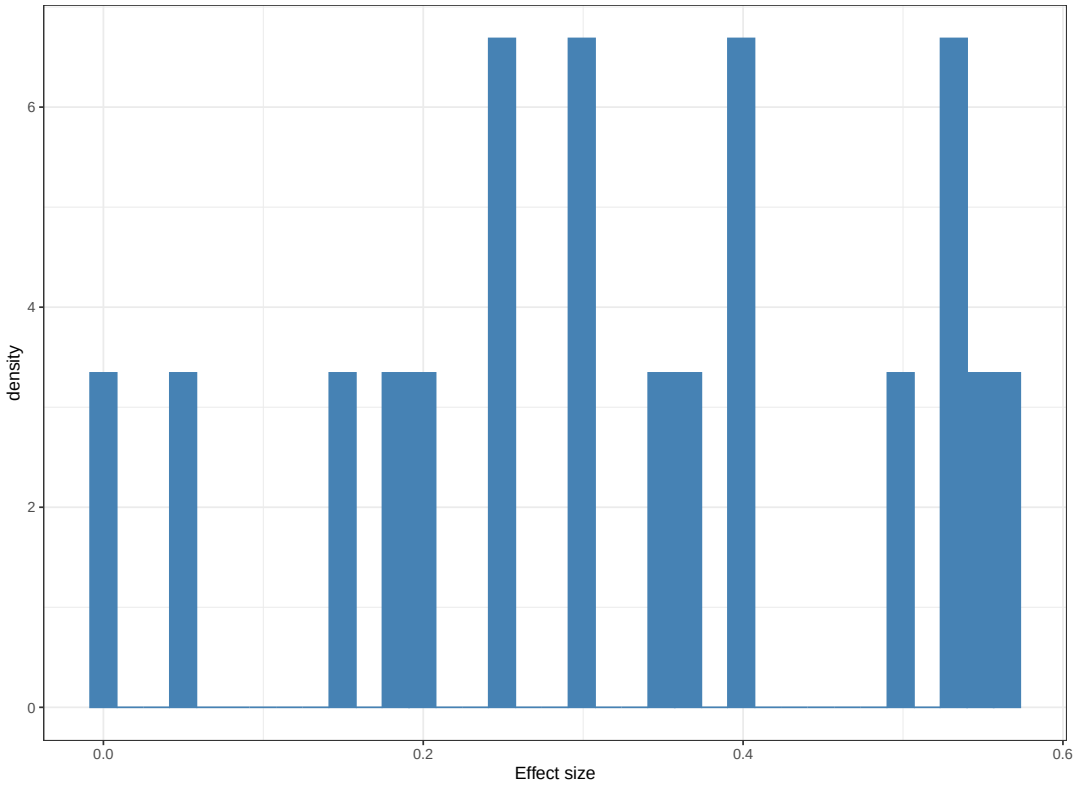

Figure 3.2: A histogram showing the distribution of the effect size.

assumed to be sufficiently conservative to power the trial was most often selected (7/34, 20%), while some studies obtained their ICC values from pilot studies (3/34, 9%). Other sources (4/34, 12%) of obtaining the ICC included ICC databases, audit, or survey data. Eight studies (24%) provided no justification for the selected value of their ICC. This is an indication of the difficulty many trials face in estimating or assuming a suitable value for the ICC during sample size calculation. A case in point is the AMBER care feasibility CRT where the researchers were ultimately unable to estimate the ICC due to a small number of clusters (Koffman *et al.*, 2019).

Similar to the ICC, the most common reason stated by trialists for the selected value of the effect size was that it was based on a review of evidence from previous studies (10/18, 56%), followed by a pilot study (3/18, 16%). Five studies (28%) offered no rationale for

their assumed effect size.

Concerning the assumed SD, 7 studies (44%) did not justify the selected SD. However, some trials obtained their selected values from a previous study (6/16, 38%), while the remaining studies extracted the SD from a pilot study (3/16, 18%).

Parameter	Justification	n	Percentage
Assumed ICC (N = 34)			
	Previous study	12	35
	Conservative value	7	20
	Pilot Study	3	9
	Others	4	12
	Not given	8	24
Assumed effect size (N = 18)			
	Previous study	10	56
	Pilot study	3	16
	Not given	5	28
Assumed SD (N = 16)			
	Previous Study	6	38
	Pilot Study	3	18
	Not given	7	44

Table 3.3: Justifications for the assumed ICC, effect size and the SD.

3.3.4 Impact of the CONSORT guideline on ICC reporting

The consequences of failing to report the ICC are serious, as the ICC is a key parameter in designing and planning future CRTs. It is also important in the interpretation of trial results. As a result, a CONSORT extension was developed and published in 2012 to guide authors on CRT reporting. In this section, I assess the impact of the CONSORT extension to CRTs on ICC reporting; particularly interest is given to whether the CONSORT reporting guideline is followed. The results are presented in Figure 3.3.

Surprisingly, there seems to be low adherence to the CONSORT statement following its publication in relation to both the assumed and observed ICC being reported. In the case of the assumed ICC, all the trials that predated the publication of the CONSORT statement reported their assumed ICC. Conversely, some trials post-publication of the guideline did not publish their assumed ICC. In terms of the observed ICC, all the trials except one conducted pre-publication of the CONSORT statement reported an observed ICC. However, only 2014 and 2015 were the years in which a 100% adherence rate was achieved for reporting an observed ICC post-publication of the CONSORT extension.

The results from Figure 3.3 suggest that reporting of the assumed and observed ICC has become less frequent in practice. This highlights one of the major difficulties in

specifying the ICC at the design stage of a trial, given that previous studies are the most common sources of obtaining suitable values. It is, however, worth noting that there are considerably more trials included in the review from post the CONSORT extension's publication, and this increase could potentially contribute to difficulty in achieving a 100% reporting rate for the ICCs. Nonetheless, the ICC should always be reported owing to its relevance to CRTs, and we would have especially expected its presence in reports within the HTA Journal.

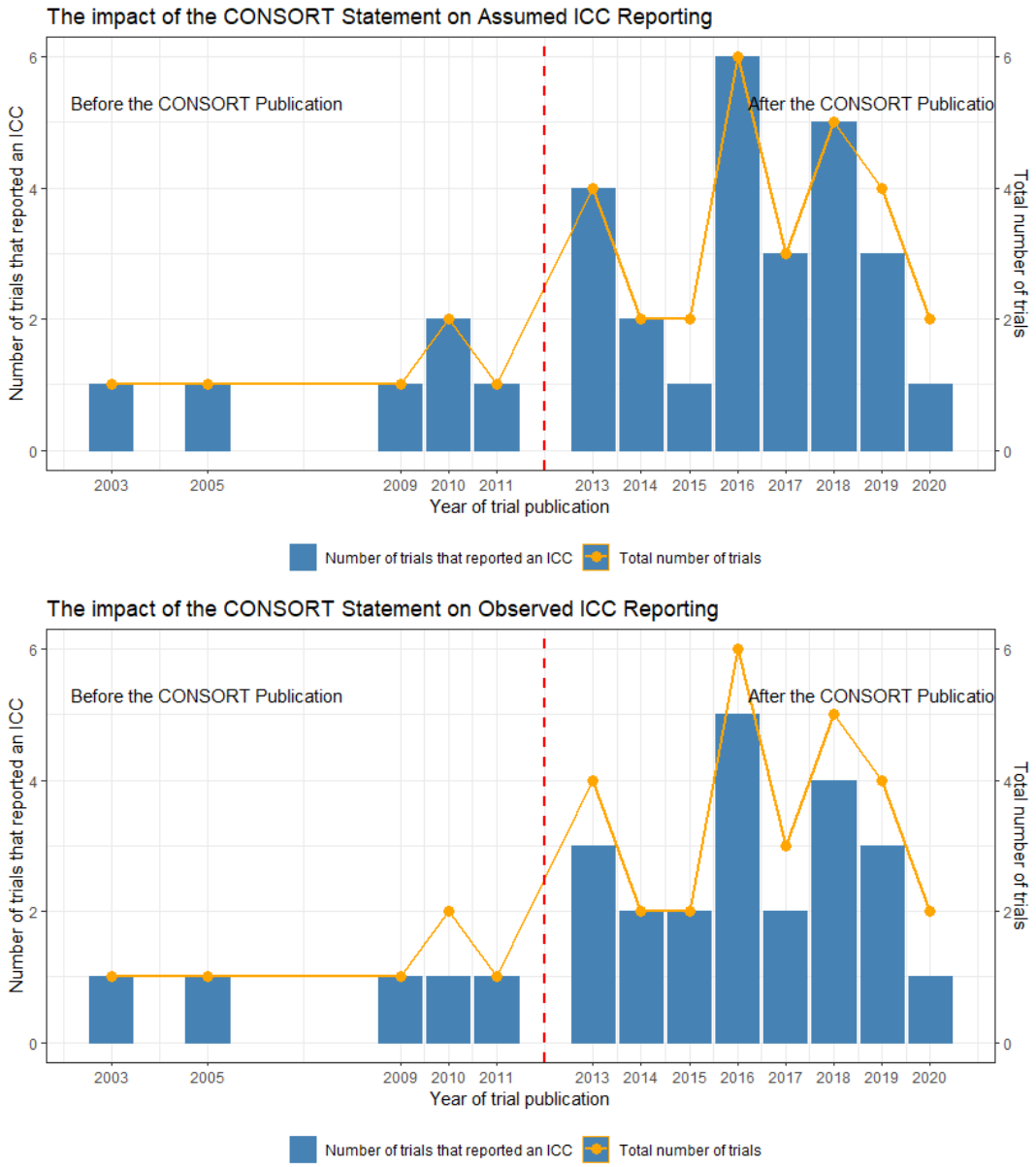

Figure 3.3: Plot comparing trends in ICC reporting pre and post the publication of the CONSORT extension for CRTs.

3.4 Discussion

For any CRT design, an issue in practice is obtaining relevant and accurate estimates for the ICC to assume at the trial design stage for sample size estimation. This issue is exacerbated by the fact that estimates of the ICC from pilot studies are often inaccurate due to the smaller sample sizes employed in such studies and how precision of estimation of ICCs scales in the number of clusters and sample size per cluster (Eldridge *et al.*, 2016). A similar statement also holds, though arguably to a smaller degree, regarding the SD and effect size at which the trial is powered.

Owing to this, we have conducted a review of the values typically utilised during sample size calculations and the justifications associated with the selected values based on studies reported in the HTA journal. Our selection of the HTA journal (reports within which are typically more extensive and longer than a clinical journal article) was premised on the fact that it generally serves as a basis for policy implications/recommendations, evidence reviews, and technology acquisition, and should arguably represent the upper-end of quality of trial reports (Carlos & Goeree, 2009).

Therefore, the poor reporting of ICCs and other key design parameters was a disappointing finding. This finding is consistent with a recent review of publicly funded trials in the UK which found that 42% of ICCs for the analysed primary outcomes were not reported, while 12% of the studies did not report an ICC at all (Offorha *et al.*, 2022). Similarly, a systematic review of ICC reporting by Han *et al.* (2019) corroborated this finding, with only 26% of the 281 CRTs analysed reporting estimated values of the ICC. Although the CONSORT extension for CRTs encourages the reporting of ICC values, the low adherence to this guideline as revealed in our findings contributes to the lack of relevant available estimates of ICCs for future trials. In the LIFELAX trial, for example, the authors stated that *"no data currently exist from which a relevant intracluster correlation coefficient (ICC) for this trial can be calculated"* (Speed *et al.*, 2010). Similarly, Pallan *et al.* (2019) bemoaned the lack of ICC for their study and had to rely on patterns of ICCs from other sources to inform their calculation. This phenomenon may lead to problems in practice, as many CRT trials rely on previously reported estimates for their sample size determination.

Even when there are relevant historical trials with ICC values available, how to effectively combine them to reflect their varying degrees of relevance to the planned trial is often an issue. Although Turner *et al.* (2004) proposed a method for assigning weights to each of the ICCs from relevant historical trials, their approach was purely subjective within a Bayesian design. Moreover, an overview of the reviewed trials indicated that many criteria, such as the number of clusters, average cluster size, subjects, setting, stratification, and outcomes that need to be considered to establish if the ICC is useful for

the study being designed were often not reported, as was also found in Chakraborty *et al.* (2009).

Given the above, the ICC is likely to be imprecisely specified at the design stage (Lewis & Julious, 2021). However, of the trials that reported ICC values, the imprecision inherent in the ICC was often not quantified and reported for use in future trials. The dominant use of point estimates from the reviewed trials neglects any consideration of uncertainty around the assumed parameter value. Misspecification of the nuisance parameters as a result of failing to account for uncertainty around the estimates may impact the trial's validity and power (Grayling *et al.*, 2018).

The results of this study revealed a considerable difference between the assumed and actual ICC, with the assumed values frequently higher than the observed values. Thus, many of the trials may have been overpowered. This could lead to research waste (Alonzo & Pepe, 2007). Moreover, an overpowered trial might show a statistically significant difference but no clinically relevant change (Bhardwaj *et al.*, 2014). Given that the safety and efficacy of an intervention are not fully known and documented pre-trial, some studies have also discussed the ethical implications of overpowered trials, especially for human and animal subjects (Altman, 1980; Faber & Fonseca, 2014; Rothwell *et al.*, 2018). Conversely, there is a loss in statistical power to detect a true effect if the sample size is small as a result of underspecified parameter estimates. Therefore, the importance of accuracy in these parameter estimates cannot be overemphasised.

We highlight a few HTA trials (Campbell *et al.*, 2015; Adab *et al.*, 2018) that did provide good justification for the assumed ICC. In particular, Campbell *et al.* (2015) utilised a 95% CI for the ICC based on estimates from a pilot study. This provides a range of plausible values for which the true ICC could lie with some degree of certainty. It is also worth noting in the case of Gates *et al.* (2017) that the ICC was monitored at interim analyses for sample size re-estimation. I discuss this adaptive design in detail in Chapter 5.

In conclusion, the discussed key design parameters have an impact on both the sample size and the interpretation of the primary outcome of a CRT. The statistical issues unearthed from the review, coupled with the fact that CRT sample sizes can be highly sensitive to variations in the ICC or the choice of the SD or effect size heightens the need for accuracy in these parameters. The significance of using precise estimates of the ICC during sample size computation is further discussed by Ukoumunne *et al.* (1999), and amplified in the CONSORT extension to CRTs (Campbell *et al.*, 2012). These statistical issues, if not addressed, may have implications on the validity of trials. Some methods to account for uncertainty regarding the ICC are proposed in Chapters 4 and 5.

Chapter 4

A hybrid (Bayesian-frequentist) approach to designing parallel-group and stepped-wedge cluster randomised trials

As discussed in Chapter 1, the design of a CRT requires consideration of the fact that individuals within a cluster may have similarities that need to be accounted for during sample size calculation. The difficulties in obtaining precise estimates of the degree of such similarities (ICC), and the associated statistical issues, were outlined in Chapter 3. This chapter provides a possible solution to such difficulties by making explicit allowance for the uncertainty around the ICC. In addition to the uncertainty in the ICC, I also evaluate how uncertainty in other key parameters (e.g., the treatment effect and SD) impact the required sample size in hybrid design.

4.1 Introduction

The increasing popularity of CRTs has seen the emergence of more established and commonly used designs that have had associated strong methodological development in recent times. One example is the SW-CRT design, which is relatively new, at least compared to the more standard PG-CRT design that has now long stood as the most commonly used CRT design. With interest in SW-CRTs ever expanding, the work in this chapter is restricted to comparing PG-CRT and SW-CRT designs, to contribute to the literature on when a SW-CRT may be preferable. As a reminder, Figure 4.1 provides a brief schematic of the functionality of PG-CRT and SW-CRT designs.

Expanding on the above, in a PG-CRT, clusters are randomised to either control or

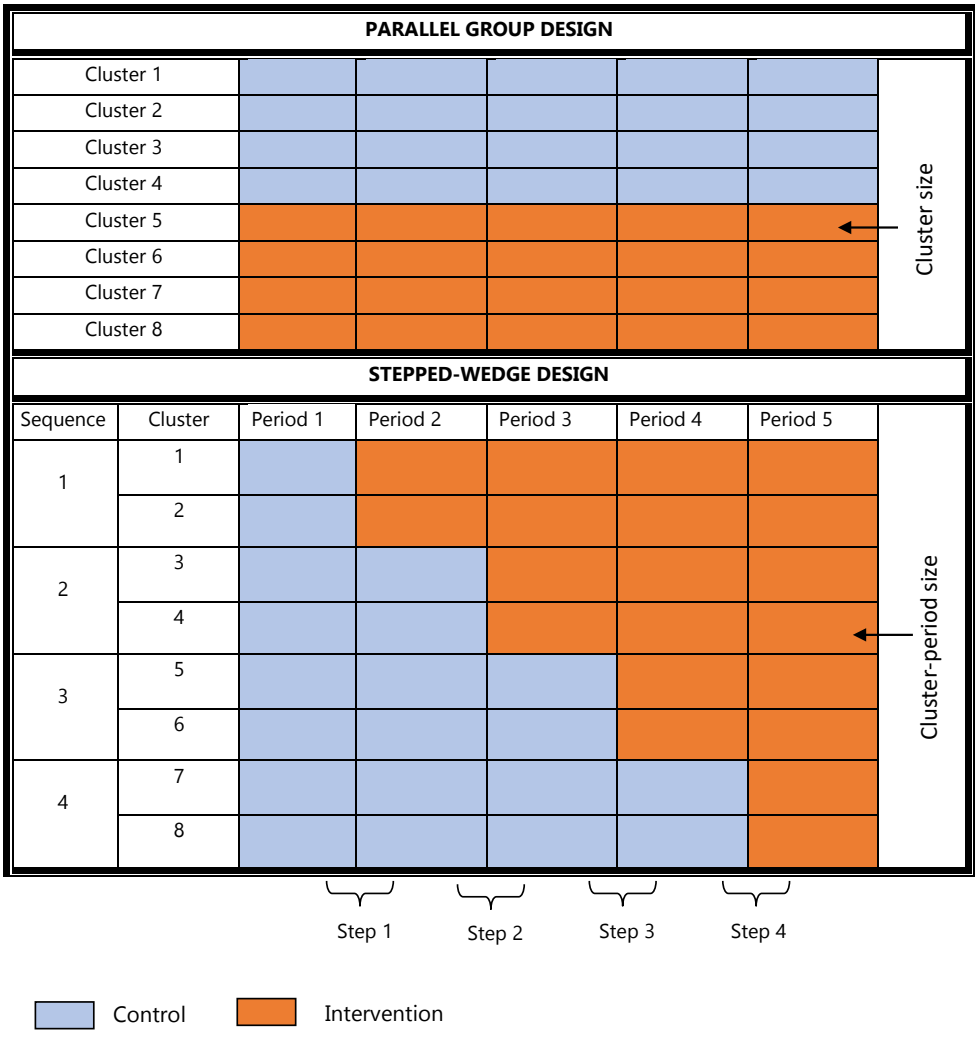

Figure 4.1: Schematic of the parallel group (PG) and stepped-wedge (SW) cluster randomised trial designs. The example PG-CRT comprises 8 total clusters where 4 clusters are randomised to the control and intervention respectively. The example stepped-wedge design also comprises 8 clusters where 2 clusters are randomly allocated to each of the 4 sequences and measurements were taken over 5 time periods in 4 steps.

the intervention and stay in this condition throughout the trial. This conventional design has regularly been used for CRTs, with wide application in health services and sometimes animal research. A recent review of CRTs by Offorha *et al.* (2022) found that 85% (73/86) of included studies had used a two-arm PG design, which is consistent with the results of an earlier review by Eldridge *et al.* (2004).

A key justification for PG-CRT usage is the simplicity of the design, which amongst other advantages, reduces the complexity of data analysis. In particular, as is discussed by

Hemming & Taljaard (2020), PG-CRTs are not susceptible to time-varying confounding, since outcomes from individuals are collected in both arms on the same follow-up schedule. Additionally, contamination across arms in a PG-CRT is less likely to occur if cluster units are selected appropriately (Hemming & Taljaard, 2020). Given that the intervention is withheld from a proportion of the total clusters in the PG design, robust evidence can be generated if randomisation is justified and clinical equipoise holds (Hemming & Taljaard, 2020). Typically, it is also more logistically and practically simple to conduct a CRT trial with a PG design.

As CRTs typically seek approval from decision-makers like GP managers for cluster involvement (Edwards *et al.*, 1999; Taljaard *et al.*, 2013), these key stakeholders are often hesitant to join the trial unless they are assured the chance of receiving the intervention (Prost *et al.*, 2015; Hargreaves *et al.*, 2015). Considering the fact that CRTs are mostly unblinded, this may diminish participation or pose difficulties in recruitment in the PG design when clusters become aware that they might not receive the intervention. Although time-varying confounding is unlikely to occur in a PG-CRT, there is an increased risk of uneven distribution of confounding variables between the treatment arms if only a small number of clusters are available for randomisation. These imbalances in both known and unknown prognostic characteristics could pose challenges in attributing outcome differences to the treatment. This may have consequences on the internal validity of the trial. However, such issues can often be corrected with matching or stratification to increase balance across the treatment arms on likely confounders (Donner & Klar, 2000; Campbell & Walters, 2014).

In a SW-CRT, all clusters typically start in the control condition and switch in one direction to the intervention at different time points (steps). This is, at prespecified time intervals, one or more clusters move to the intervention condition. Therefore, randomisation in this design pertains to when cluster(s) begin to receive the intervention. The number of clusters that switch to the intervention at each time point is often the same. This trial design can be likened to a cluster randomised cross-over design where each cluster receives an intervention for a specified duration, followed by a washout period, and subsequently receives the alternative intervention. The similarity in both designs is that the clusters can act as their own control. In the case of SW-CRTs, however, clusters cannot revert to being in a control condition once the intervention has been administered; this is the principal difference compared to a cluster-randomised crossover trial.

The Gambia Hepatitis Study, which sought to assess the incidence of chronic liver disease and liver cancer over a period of 30 to 40 years following the administration of the hepatitis B vaccine, is likely the earliest SW-CRT and a well-discussed example (Hall *et al.*, 1987). In this trial, researchers divided the entire country into 17 areas and a vaccination team was assigned to each area. The hepatitis B vaccine was administered

to all newborns and infants in an area by a randomly selected team every three months, and after four years the entire country had been vaccinated. The factors that generally account for the adoption of a SW-CRT design, and in particular applied to the Gambia Hepatitis Study, include

- **Ethical reasons:** SW-CRTs have been argued to be useful in situations where it is unethical to withhold the intervention or treatment from a cluster (Hemming & Taljaard, 2020). For example, the safety and importance of vaccines to public health have been well documented by the WHO and NHS (World Health Organisation, 2022; NHS, 2019). Hence, it may be deemed unethical to withhold a vaccine believed/known to be more beneficial than harmful from a cluster. Although some rebuttals have been offered by authors against this justification, on the basis of clinical equipoise and the time lags in administering a supposed beneficial intervention (Prost *et al.*, 2015; Binik, 2019), it remains the main rationale for adopting a SW-CRT in the literature (Grayling *et al.*, 2017; Eichner *et al.*, 2019).

- **Logistical reasons:** The authors noted that the high cost of the vaccines and their limited supply precluded an immediate universal rollout. Therefore, using a stepped wedge design to administer the vaccine in batches was the most efficient way to ensure the universal rollout of the intervention at the end of the study. Here, individual randomisation in such a large trial with sizable immunization teams would have posed severe logistical challenges.

- **Availability of comparison groups from the same period:** Another justification for the adoption of the stepped wedge design is the desire to have comparison groups from the same time period. This becomes a rich source of information for the researchers by using between- and within-cluster comparisons to estimate the treatment effect. In settings where the entire population does not form the scope of the intervention rollout, this can result in a lower required sample size for high statistical power since clusters act as their own control.

Regardless of the advantages of the SW-CRT design outlined above, it does have limitations. First, the unidirectional aspect of treatment crossover does complicate the requisite data analysis since the treatment effect must typically be estimated using both within- and between-cluster comparisons (Campbell & Walters, 2014). Second, the staggered implementation may introduce temporal trends or other confounding factors that need to be carefully considered in the analysis. There is also a greater risk of bias in SW-CRTs when the secular trends are misspecified at the analysis stage (Hemming & Taljaard, 2020).

Given the respective advantages and disadvantages of PG-CRT and SW-CRT designs, previous works have sought to compare them to outline the potential conditions under

which a particular design may be the preferred choice (Hemming & Taljaard, 2020). In this chapter, I seek to contribute to this literature by comparing the efficiency of PG-CRT and SW-CRT designs, as related to their required sample size under a particular design framework. Principally, such sample size comparisons have been conducted by previous authors (Woertman *et al.*, 2013; Baio *et al.*, 2015; Hemming *et al.*, 2015), but in a frequentist framework that did not account for uncertainty in key parameters required for sample size calculation.

In the frequentist framework, the sample size for both the PG and SW designs is typically calculated conservatively by assuming a higher ICC or variability within clusters. This ensures that the calculated sample size is sufficient to achieve the desired power even in scenarios where the ICC or variability is higher than anticipated. Additionally, conservative assumptions may be made regarding other parameters such as effect size or dropout rates to ensure that the calculated sample size provides adequate power under various scenarios.

Therefore, these previous works determined the sample sizes for each CRT design by utilising the standard frequentist approach of calculating an IRT sample size and inflating it by their respective CRT design effect to account for clustering. By fixing the number of clusters and time-period for each value of the ICC, they evaluated how the magnitude of each design effect under the varied ICC impacted the CRT sample size. The varied ICC values in their studies were point estimates, which neglected any consideration of uncertainty inherent in the parameter. Considering the sensitivity of the ICC in CRT sample size determination, methods that fail to take into account uncertainty in the ICC may lead to trials that are either underpowered or overpowered (Grayling *et al.*, 2017).

Therefore, particular focus is given to the uncertainty regarding the value of the ICC. This is accounted for using a hybrid design framework, a perspective which has received little attention in a CRT context. In the hybrid framework, the uncertainty in the ICC is expressed through a parametric prior distribution. This allows researchers to simply and directly account for any uncertainty in key design parameters in their sample size calculation. The most relevant work to CRTs designed within the hybrid framework is that of Lewis & Julious (2021) who, based on results from Ukoumunne (2002), leveraged confidence intervals characterising a plausible range for the ICC to incorporate uncertainty in its value into the sample size calculation. This is similar to a hybrid approach but does not associate a particular prior density to each possible ICC value.

This chapter describes how to determine the minimal sample size required to achieve a desired Expected Power (EP), one of the quantities primarily controlled in the hybrid literature (Kunzmann *et al.*, 2021). This is done for a setting in which a prior is placed not only on the ICC but also on the treatment effect and SD, for which there may also be substantial uncertainty at the design stage. Furthermore, case studies of the PG and SW

CRTs which assumed 'conservative' values for the ICC in their sample size determination are presented; the required sample sizes from the conventional frequentist approach are compared against the sample sizes obtained when uncertainty in the ICC is accounted for through prior specification. Specifically, I select priors with hyperparameters whose mean is equivalent to the point estimate of the PG or SW CRTs' frequentist counterpart. This selection enables an examination of how varying levels of variability, as captured by the SD on the prior mean, affect the required sample size. Subsequently, I provide a critical evaluation of how placing a prior on the ICC, where both the PG and SW CRT priors assume the same values for the prior means and SDs impacts whether a PG- or SW-CRT design is more efficient, extending previous comparisons under a fixed ICC, such as those by previous authors.

I now proceed by briefly discussing some possible approaches to account for uncertainty in the ICC before arguing for the method that will be developed later in this chapter.

4.1.1 Possible approaches to account for uncertainty in the ICC

As stipulated above, failing to formally account for uncertainty in sensitive parameter estimates such as the ICC increases the likelihood of parameter misspecification. This could have implications for the sample size, the statistical power, and the validity of the trial. To guard against these implications, some studies have established the usefulness of allowing for the impression in the ICC within the sample size calculation (Ukoumunne *et al.*, 1999; Campbell *et al.*, 2012). Such methods characterise likely values of the nuisance parameter(s) in the form of a range or a parametric distribution, hoping to provide an improvement over a point estimate assumption or being conservative. I first describe a few such methods in this section.

The *confidence interval (CI)* method provides a plausible range of values within which the 'true' value of the ICC, ρ, may lie with some degree of probability. Constructing a confidence interval for ρ typically requires knowledge of its variance and by extension, standard error. Some analytical formulae to estimate the variance of the ICC exist in the literature (Swiger *et al.*, 1964; Fisher, 1970; Searle, 1971). A few studies have employed these techniques; prominent among them is Pagel *et al.* (2011), who utilised Fisher's formula to construct a 95% confidence interval for the ICC when assessing perinatal outcomes from five randomised-controlled trials in LMICs. Donner & Wells (1986), and Ukoumunne (2002), further provide a comprehensive coverage of these formulae and compare their performance with simulated clustered data.

A common limitation of these formulae is that they are approximate and depend on knowledge of the total sample size, the number of clusters, cluster size(s), and also the 'true' ICC, which is fundamentally unknown. Hence, erroneously specifying the 'true' ICC in the formulae may give investigators a false sense of confidence. Another limitation is

that although the 95% CI provides a range of values for the ICC, the upper limit value is often used in practice to be sufficiently conservative. This may substantially diminish the estimate of power for a fixed sample size (Campbell *et al.*, 2004; Copas & Malley, 2008). As a result, some authors have argued against the use of such methods for planning purposes (Turner *et al.*, 2004; Campbell *et al.*, 2007).

Alternatively, employing a *Bayesian approach* improves upon the upper limit of the confidence interval method as it accounts for the entire assumed distribution of the ICC. This approach provides a favourable alternative for the imprecision in the ICC to be naturally incorporated. In the Bayesian setting, a prior distribution is placed on the ICC (and/or other key design parameters such as the standard deviation or the treatment effect). Through prior specification, the relative likelihood of parameter values is captured which leads to a distribution of the projected power (posterior distribution).

In CRTs, previous work on incorporating uncertainty about the ICC in a Bayesian framework has focused on how to formally quantify uncertainty based on estimates from past studies, compute an associated power distribution, and use an informative prior for the ICC in a trial's analysis (Turner *et al.*, 2004, 2005). Building upon Turner's work, Tishkovskaya *et al.* (2023) recently estimated an ICC using the same methodology for a planned CRT. Gary (2022) also utilised information borrowing from historical data to construct power priors for sample size calculation in the context of CRTs. Application of Bayesian methods to the design and analysis of CRTs has further been reviewed by Jones *et al.* (2021).

In spite of the methodological advancement within the Bayesian paradigm, there are still some limitations. First, issues such as the subjectivity of Bayesian priors make this approach less appealing to a proportion of the research community. The primary concern arises when employing an informative prior for the analysis. Some studies differentiate between a design prior utilised for determining the sample size and an analysis prior used for the actual analysis. In most cases, the former can more reasonably incorporate an informative prior. Additionally, it becomes conceptually and computationally intensive if a closed-form posterior distribution (conjugacy) is not achieved. In a regulated setting, it also remains the case that regulatory agencies tend to favour trials designed and analysed within a frequentist framework over those with a Bayesian paradigm although CRTs are rarely within a regulated setting.

4.1.2 Proposed solution to sample size calculation under parameter uncertainty: Hybrid design

The hybrid approach involves the use of Bayesian techniques for sample size estimation, while always retaining an assumption of a purely frequentist analysis. Spiegelhalter & Freedman (1986) first proposed this concept, before it was later referred to as a hybrid

Bayesian-frequentist approach (Spiegelhalter *et al.*, 2004). Theoretically, using a hybrid approach, we can incorporate uncertainty in particular parameters within the trial design, mitigate the risk of overly optimistic power calculations, and satisfy most regulatory agency guidelines (as well as standard trials community preferences) by maintaining a frequentist framework for the final analysis (Ciarleglio *et al.*, 2016).

In the hybrid literature, the quantities typically controlled for sample size determination are the probability of success (PoS), expected power (EP), and assurance. The relationship between these quantities is highly dependent on the definition of a successful trial, how uncertainty in the unknown parameter(s) is utilised in sample size calculation, and which parameter(s) uncertainty is accounted for. To better understand and appreciate the subtle differences between these quantities, let us first consider a traditional two-arm individually randomised superiority trial where the efficacy of a new drug is compared with a placebo. Here, we may define 'success' as correctly rejecting the null hypothesis (H_0). If the null hypothesis suggests no positive treatment effect ($H_0 \leq \delta$), then the conventional frequentist power may be defined as the probability of rejecting the null hypothesis given some assumed parameters such as the particular positive treatment effect, variance, etc. Based on the above definition of success, the traditional frequentist power can be thought of as similar to the PoS. However, the traditional frequentist power is conditioned on assumed values of unknown parameters which may be imprecise.

The PoS on the other hand is an unconditional probability of correctly rejecting H_0. In the early 2000s, O'Hagan and colleagues referred to this concept as assurance and defined it as the unconditional probability of a trial showing a positive outcome (O'Hagan & Stevens, 2001; O'Hagan *et al.*, 2005). The unconditional probability is obtained by placing a (prior) distribution on the unknown parameter(s) and integrating over the conditional (traditional) frequentist power with a weighting factor given by the prior density. Because of the averaging over the parameter space of the unknown parameter(s), several authors have referred to this concept as a weighted average or expected power (Gillett, 1994; Spiegelhalter *et al.*, 1994; O'Hagan *et al.*, 2005; Chuang-Stein, 2006).

In a recent work by Kunzmann *et al.* (2021), a high-level definition of PoS was proposed that encompasses unconditional rejection of the null hypothesis and a relevant underlying effect. Thus, they considered a fixed assumed treatment effect as a minimal clinically important difference (say $\delta_{MCID} = 0.3$), and selected a prior for the treatment effect which is truncated to only consider cases where the treatment effect is greater than the assumed MCID. Based on their definition of success, the EP was expressed as the quotient of the PoS and the truncated prior for the treatment effect. In the Methods section below, I elaborate on these quantities with notations.

Observe that while the PoS and assurance proposed by previous authors provide an unconditional value that requires just a positive treatment effect ($\delta > 0$ or $\delta_{MCID} > 0$),

Kunzmann's PoS provides an unconditional value that guarantees that the treatment effect will be greater than some MCID ($\delta > \delta_{MCID}$). Hence, the key difference between these quantities is how success is defined in terms of the a priori likelihood of the treatment effect. Therefore, if our interest lies in accounting for the uncertainty in another parameter (for example, the ICC) instead of the treatment effect, then these hybrid quantities become equivalent. In such a scenario they can be used interchangeably. Kunzmann further defines other scenarios in which these quantities become equivalent in the presence of uncertainty in the treatment effect.

For consistency throughout the thesis, I shall utilise the definition of success set forth by Kunzmann and colleagues and adopt the EP quantity for the sample size determination. This is because the EP quantity is known to typically take values more compared to the frequentist power. This will produce some degree of fair comparison between the hybrid and frequentist frameworks in the results.

4.2 Methods

To assess the effectiveness of an intervention in a CRT, we typically need to determine the number of participants required by the trial for a certain level of power. Thus, we now proceed to describe how sample size calculation for a PG- or SW-CRT can be performed in both the frequentist and hybrid frameworks. For brevity as indicated in Chapter 3, the methodology is limited to the setting with one-level clustering and not nested clustering. I also restrict my attention throughout to the case where the outcome data is assumed to be normally distributed. Additionally, when addressing SW-CRTs, I focus solely on 'cross-sectional' design. I will, however, return to comment later on extensions to our work.

4.2.1 Analysis models

We consider the comparison of a control (indexed P, for placebo even though this need not a placebo arm, to avoid confusion with the number of clusters, C) and experimental intervention (indexed E). For PG-CRTs, the following linear-mixed model is assumed to be used for the analysis

$$Y_{ij} = \mu_P + X_j\mu + c_j + e_{ij}.$$

Here, Y_{ij} is the outcome from patient $i = 1, \ldots, N$ (thus we assume N participants per cluster) in cluster $j = 1, \ldots, C$ (thus we assume C clusters); μ_P is an intercept term (or the mean outcome of the control group); $X_j = 1$ if cluster j is allocated to the intervention arm and $X_j = 0$ otherwise; μ is the effect of the intervention relative to the control;

$c_j \sim N(0, \sigma_c^2)$ is a random effect for cluster j, which allows for a non-zero correlation between outcomes from participants within the same cluster; and $e_{ij} \sim N(0, \sigma_e^2)$ is the individual-level error.

As proposed by Hussey & Hughes (2007), we extend this for SW-CRTs to

$$Y_{ijk} = \mu_P + \beta_j + X_{jk}\mu + c_j + e_{ijk}.$$

Here, Y_{ijk} is the outcome from patient $i = 1, \ldots, n$ (thus we assume n participants per cluster-period) in cluster $j = 1, \ldots, C$, in time period $k = 1, \ldots, T$ (thus we assume T time periods; which means there are $N = nT$ measurements per cluster in total); $X_{jk} = 1$ if cluster j is allocated to the experimental arm in time period k and $X_{jk} = 0$ otherwise; and β_j is a fixed effect for period j (i.e., due to the sequential rollout of the intervention, the model adjusts for the time period of collection, setting $\beta_1 = 0$ for identifiability). All other parameters are interpreted as above.

Note that the sample sizes of the PG- and SW-CRT designs are both then given by NC, with $N = nT$ in the SW-CRT case. Furthermore, in both instances, the ICC is defined as $\rho = \sigma_c^2/(\sigma_c^2 + \sigma_e^2) = \sigma_c^2/\sigma^2$. This is the ratio between the variation between clusters (σ_c^2) and the total (between and within cluster) variation (σ^2).

4.2.2 Power and sample size calculation within the frequentist framework

Recall that sample size determination in a CRT amounts to calculating the sample size for a standard IRT and multiplying it by a design effect (DE) to account for the particular CRT design used. Therefore, we first consider an IRT where individuals are randomised to either a control (P) or experimental intervention (E), with data y_{ij} gathered for individual $i = 1, \ldots, N_{IRT}$ and arm $j = P, E$. For simplicity and practical relevance, we further assume that $Y_{ij} \sim N(\mu_j, \sigma^2)$, and that the null hypothesis is

$$H_0 : \mu = \mu_E - \mu_P \leq 0. \tag{4.1}$$

I.e., we assume our interest lies in assessing whether there is a positive treatment effect in the intervention arm.

We perform a test for H_0 using the test statistic

$$z = \frac{\hat{\mu}}{\sqrt{Var(\hat{\mu})}} = \frac{\frac{1}{N_{IRT}} \sum_{i=1}^{N_{IRT}} y_{iE} - \frac{1}{N_{IRT}} \sum_{i=1}^{N_{IRT}} y_{iP}}{\sqrt{\frac{2\sigma^2}{N_{IRT}}}}. \tag{4.2}$$

It can be shown that $Z \sim N(\mu\sqrt{N_{IRT}/(2\sigma^2)}, 1)$, and thus to control the type-I error-rate to level α, we can reject H_0 when $z > \Phi^{-1}(1 - \alpha)$. In turn, this gives that the

probability H_0 is rejected for a particular value of μ is

$$P(\mu) = \mathbb{P}(\text{Reject } H_0|\mu) = \Phi\left\{\mu\sqrt{\frac{N_{IRT}}{2\sigma^2}} - \Phi^{-1}(1-\alpha)\right\}.$$

Thus, if we require power of $1-\beta$ when $\mu = \delta > 0$ (i.e., require $P(\delta) = 1-\beta$), this results in the following sample size calculation formula:

$$N_{IRT} = \frac{2\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2\sigma^2}{\delta^2}.$$

Given the above IRT formula, all that is then required to compute the sample size required by the considered types of PG- and SW-CRT is their respective DEs (Woertman *et al.*, 2013; Hemming & Taljaard, 2016). These are, for the analysis models specified in the previous subsection, as follows

$$\begin{aligned} DE_{PG-CRT} &= 1 + (N-1)\rho, \\ DE_{SW-CRT} &= \frac{(1+\rho(nT+n-1)}{(1+\rho(\frac{nT}{2}+n-1))} \times \frac{3(1-\rho)}{2(T-\frac{1}{T})}. \end{aligned}$$

For complete clarity, the total sample size (SS) required by frequentist PG- and SW-CRT designs are respectively given by

$$\begin{aligned} SS_{PG-CRT} &= 2N_{IRT}DE_{PG-CRT}, \\ SS_{SW-CRT} &= 2N_{IRT}DE_{SW-CRT}. \end{aligned}$$

Here, the factor of 2 arises as N_{IRT} is the sample size required per arm in an IRT.

In the CRT case, the hypothesis in Equation 4.1 is again the one of interest and the test statistic used is still

$$z = \frac{\hat{\mu}}{\sqrt{Var(\hat{\mu})}}.$$

For both the PG-CRT and SW-CRT designs, under our given model assumptions we have that $\mathbb{E}(\hat{\mu}) = \mu$. Specifying a general formula for the frequentist power thus depends on knowledge of $Var(\hat{\mu})$ for a given CRT design, which can be directly related to the above design effects. Specifically, in the PG-CRT case, assuming 1:1 cluster-level allocation to the two arms, it can be shown that (Turner *et al.*, 2005):

$$Var(\hat{\mu}) = \frac{4\{1+(N-1)\rho\}\sigma^2}{CN}.$$

Thus, the probability H_0 is rejected for a PG-CRT design (i.e., the frequentist power) is

$$\Phi\left[\mu\sqrt{\frac{CN}{4\{1+(N-1)\rho\}\sigma^2}}-\Phi^{-1}(1-\alpha)\right].$$

Similarly, it can be shown for a SW-CRT, under the analysis model given earlier, that (Hussey & Hughes, 2007; Lawrie *et al.*, 2015)

$$Var(\hat{\mu})=\frac{C\sigma^2(1-\rho)[1+\rho(nT-1)]}{n\{[1-\rho(nT-1)](CU-W)+n\rho(U^2-CV)\}},$$

where

$$\begin{aligned}U&=\sum_{j=1}^{C}\sum_{k=1}^{T}X_{jk},\\ W&=\sum_{k=1}^{T}\left(\sum_{j=1}^{C}X_{jk}\right)^2,\\ V&=\sum_{j=1}^{C}\left(\sum_{k=1}^{T}Xjk\right)^2.\end{aligned}$$

Thus, frequentist power for a SW-CRT is

$$\Phi\left[\mu\sqrt{\frac{n\{[1+\rho(nT-1)](CU-W)+n\rho(U^2-CV)\}}{C\sigma^2(1-\rho)[1+\rho(nT-1)]}}-\Phi^{-1}(1-\alpha)\right].$$

We will denote the probability of rejecting H_0 for both designs by $P(\mu, N, X, \alpha, \sigma, \rho)$. The parameter X is the matrix of binary treatment indicators; $C \times 1$ in the case of a PG-CRT and $C \times T$ for a SW-CRT. For SW-CRTs, it is also implicitly assumed that $n = N/T$, with T given through X.

In the frequentist framework, a target difference $\mu = \delta > 0$ is allocated and the study is designed to ensure power is at least $100(1-\beta)\%$ in this instance, i.e., $P(\delta, N, X, \alpha, \sigma, \rho) \geq 1-\beta$. Here, β is the nominated type-II error-rate. Thus, in a conventional frequentist power calculation, the parameters δ, ρ, and σ take fixed specified values. As discussed, this negates consideration of any uncertainty around their nominated values. This can be addressed in a hybrid framework by placing priors on these parameters which I describe in the next section.

4.2.3 Sample size calculation within the hybrid framework

Priors on the treatment effect μ, SD, and the ICC ρ, may be used to allow us to capture uncertainty in their values. We will denote these respectively by $\psi_{Mu}(\delta|\theta_{Mu})$, $\psi_{SD}(\sigma|\theta_{SD})$, $\psi_{ICC}(\rho|\theta_{ICC})$. Here, θ_{Mu}, θ_{SD}, and θ_{ICC} give parameters that describe the shape of the prior densities (e.g., its mean value and variance around this). We discuss specific choices for these priors later. Note that the use of the word 'prior' here may cause some confusion; ψ_{Mu}, ψ_{SD}, and ψ_{ICC} capture the (prior beliefs about the) relative likelihood of different values of μ, σ, and ρ. They are not 'priors' in the fully Bayesian sense of the word (i.e., they will not be updated to posterior distributions).

Furthermore, in the hybrid framework, the usual frequentist power requirement is replaced by consideration of the value of the EP. We consider three scenarios: (i) when priors are placed on the treatment effect and the ICC, (ii) when priors are placed on the ICC and the SD, and (iii) when a prior is placed only on the ICC. In all three scenarios, we explicitly list the EP as functions of N and C to reflect the fact that sample size calculation is often performed for CRTs by varying one or both of the parameters N and C. Computing a sample size in the hybrid framework will then amount to ensuring $EP(N, C) \geq 1-\gamma$, by suitable choice of N and/or C through a numerical search. Here, γ need not be equal to the value of β in the traditional frequentist framework, though this is a pragmatic and often assumed approach in the hybrid literature; we will therefore set $\gamma = \beta$ throughout this chapter.

Priors on the treatment effect and the intra-cluster correlation

In this case, we have

$$EP(N, C) = \int_0^1 \int_{\delta_{MCID}}^{\infty} P(\mu, N, X, \alpha, \sigma, \rho)\psi_{Mu}(\mu|\theta_{Mu}, \mu \geq \delta_{MCID})\psi_{ICC}(\rho|\theta_{ICC})\ \mathrm{d}\mu\mathrm{d}\rho.$$

Note that the extra condition in $\psi_{Mu}(\mu \mid \theta_{Mu}, \mu \geq \delta_{MCID})$ means that the prior is truncated to only consider cases where $\mu \geq \delta_{MCID}$. This is defined in terms of the original $\psi_{Mu}(\mu \mid \theta_{Mu})$ as

$$\psi_{Mu}(\mu|\theta_{Mu}, \mu \geq \delta_{MCID}) = \frac{\mathbf{1}\{\mu \geq \delta_{MCID}\}\psi_{Mu}(\mu|\theta_{Mu})}{\int_{\delta_{MCID}}^{\infty} \psi_{Mu}(\mu|\theta_{Mu})\ \mathrm{d}\mu},$$

where $\mathbf{1}\{X\}$ is an indicator function on event X.

This formula introduces the parameter δ_{MCID}, which represents a minimal clinically important difference (MCID) that needs to be observed between the arms for the intervention to be considered better. This may or may not be equal to the value of δ assumed in the standard frequentist sample size calculation above.

As a proof of concept, I have introduced the prior on the treatment effect in a simplified manner drawing from Kunzmann's definition of the prior. However, it is important to exercise greater caution when considering a prior on the treatment effect in practice. I later elaborate on this in the Discussion.

Observe that in the above, σ takes a fixed value, as in the frequentist sample size calculation.

Priors on the standard deviation and intra-cluster correlation

In this case, we instead have

$$EP(N,C) = \int_0^\infty \int_0^1 P(\delta, N, X, \alpha, \sigma, \rho)\psi_{SD}(\sigma|\theta_{SD})\psi_{ICC}(\rho|\theta_{ICC})\ \mathrm{d}\sigma\mathrm{d}\rho.$$

Here, we assume μ takes the fixed value δ, as in the frequentist sample size calculation.

Prior on the intra-cluster correlation only

Finally, to address a specific question of interest later, we consider the scenario in which a prior is placed only on the ICC. In this case, the EP reduces to

$$EP^{ICC}(N,C) = \int_0^1 P(\delta, N, X, \alpha, \sigma, \rho)\psi_{ICC}(\rho|\theta_{ICC})\ \mathrm{d}\rho.$$

As above, we here assume μ takes the fixed value δ, while the standard deviation takes a fixed value σ.

4.2.4 Choice of priors

What remains to be explained is logical choices for the priors ψ_{Delta}, ψ_{SD}, and ψ_{ICC}. We highlight that as these are not priors in the usual Bayesian sense of the word, there are less logical restrictions on the distributional form of the priors to adopt (i.e., we need not concern ourselves with conjugacy).

For the ICC, we may reasonably choose any distribution with support $[0, 1]$ and for the SD and treatment effect, any distributions with support $(0, \infty)$ and $(-\infty, \infty)$ respectively. If two distinct priors have similar values across the range of ρ, the resultant EPs should also be similar. For this reason, our choices below are not unique ones, nor should they be considered best practice; the best distribution for a particular trial will be one that results in prior densities most accurately reflecting beliefs about that parameter.

We explore normal and non-normal priors for the ICC and assess how they impact design. As in Turner *et al.* (2004), we first assume a Truncated normal distribution is used for the ICC, truncated on $[0, 1]$. If we denote the prior mean and its SD as m and s respectively, then the PDF is given as:

$$\psi_{ICC}\{\rho|(m,s)\} = \frac{\phi(\frac{\rho-m}{s})}{s\{\Phi(u)-\Phi(l)\}},$$

where $l = (0-m)/s$ and $u = (1-m)/s$. Note that the mean of ψ_{ICC} is then

$$m + s\frac{\phi(u)-\phi(l)}{\Phi(u)-\Phi(l)},$$

while its variance is

$$s^2\left\{1 + \frac{l\phi(l)-u\phi(u)}{\Phi(u)-\Phi(l)} - \frac{\phi(l)-\phi(u)}{\Phi(u)-\Phi(l)}^2\right\}.$$

Corollary, the truncated normal distribution is expressed as $TN(0,1,m,s)$. In practice, the values of m and s could either be formed using methodology such as that provided by Turner *et al.* (2004) or elicited based on expert opinion.

Next, we assume a beta prior for the ICC, since its support $[0,1]$ is consistent with the range of the ICC. If we denote the prior by $Beta(a,b)$, then

$$\psi_{ICC}\{\rho|(a,b)\} = \frac{x^{a-1}(x-1)^{(b-1)}}{B(a,b)},\ a,b>0,$$

where B is the Beta function. This prior has mean $a/(a+b)$ and variance

$$\frac{ab}{(a+b)^2(a+b+1)}.$$

Regarding the prior for the SD, ψ_{SD}, a convenient form in practice may be a Gamma distribution since this has support $(0,\infty)$. If we denote this by $Gamma(k,\theta)$, we have

$$\psi_{SD}\{\sigma|(k,\theta)\} = \frac{\theta^k}{\Gamma(k)}\sigma^{k-1}e^{-\theta\sigma},\ k,\theta\geq 0,$$

which has mean k/θ and variance k/θ^2.

Finally, for the prior on the treatment effect, ψ_{Mu}, a convenient form in practice may be a normal distribution, as the familiarity of this distribution may make expert elicitation more feasible. Thus, using an $N(m,s^2)$ distribution, we set

$$\psi_{Mu}\{\mu|(m,s)\} = \phi\left(\frac{\mu-m}{s}\right).$$

4.2.5 Motivating examples

We motivate assumed parameters for PG-CRT examples based on Surr *et al.* (2020), a PG-CRT that sought to use Dementia Care Mapping to reduce agitation in care home residents with dementia. Hence, agitation at 16 months was the primary outcome, measured by

the Cohen-Mansfield Agitation Inventory. This study was powered at 90% ($\beta = 0.1$) with a 2.5% one-sided significance level ($\alpha = 0.025$) to detect a clinically important difference of 3 points ($\delta = 3$) with a SD of 7.5 points ($\sigma = 7.5$). An ICC of $\rho = 0.1$ was assumed, leading to 50 care homes ($C = 50$) being recruited with $n = 11$ participants per cluster.

We use O'Grady *et al.* (2021) as motivation for assumed parameter values in SW-CRT examples. This SW-CRT aimed to implement a model that would improve outpatient substance use disorder treatment outcomes. The design had $T = 7$ time periods, randomising 5 clinics to begin the intervention in each of time periods 2 through 7 (i.e., $C = 30$). The assumption was that there would be $n = 132$ participants per clinic per time period. The study was powered at 80% ($\beta = 0.2$) for $\alpha = 0.005$ and a clinically important difference of $\delta = 0.0278$. It assumed $\rho = 0.2$ and $\sigma = 0.426$.

These motivating examples were selected because they are recent trials which may offer insights into emerging trends or issues that could impact the design and interpretation of the study. Additionally, they provided clear reporting of the parameters required for calculating the sample size.

4.3 Results

4.3.1 Example trials designed within the hybrid framework

First, I provide a simple example of how the EP varies in the hybrid framework as a function of the number of clusters (Figure 4.2). This relationship is illustrated under three scenarios: (a) when priors are placed on both the ICC and the treatment effect, while holding all other parameters from the motivating examples fixed; (b) when priors are placed on both the ICC and the SD, while holding all other parameters from the motivating examples fixed; and (c) when a prior is placed only on the ICC, while holding all other parameters from the motivating examples fixed.

Parameters for the priors were selected such that the mode of the prior is always equal to the point estimate assumed in the motivating example. In what follows, we term priors whose mode matches the corresponding frequentist assumptions as 'correctly centrally specified priors'. In this case, I used for the PG-CRT example

$$\begin{aligned}\psi_{Mu} &\sim N(3, 0.01),\\ \psi_{ICC} &\sim TN(0, 1, 0.1, 0.01),\\ \psi_{SD} &\sim Gamma(75, 10).\end{aligned}$$

While, for the SW-CRT example, I utilised

$$\psi_{Mu} \sim N(0.028, 0.01),$$
$$\psi_{ICC} \sim TN(0, 1, 0.2, 0.01),$$
$$\psi_{SD} \sim Gamma(17.32, 40).$$

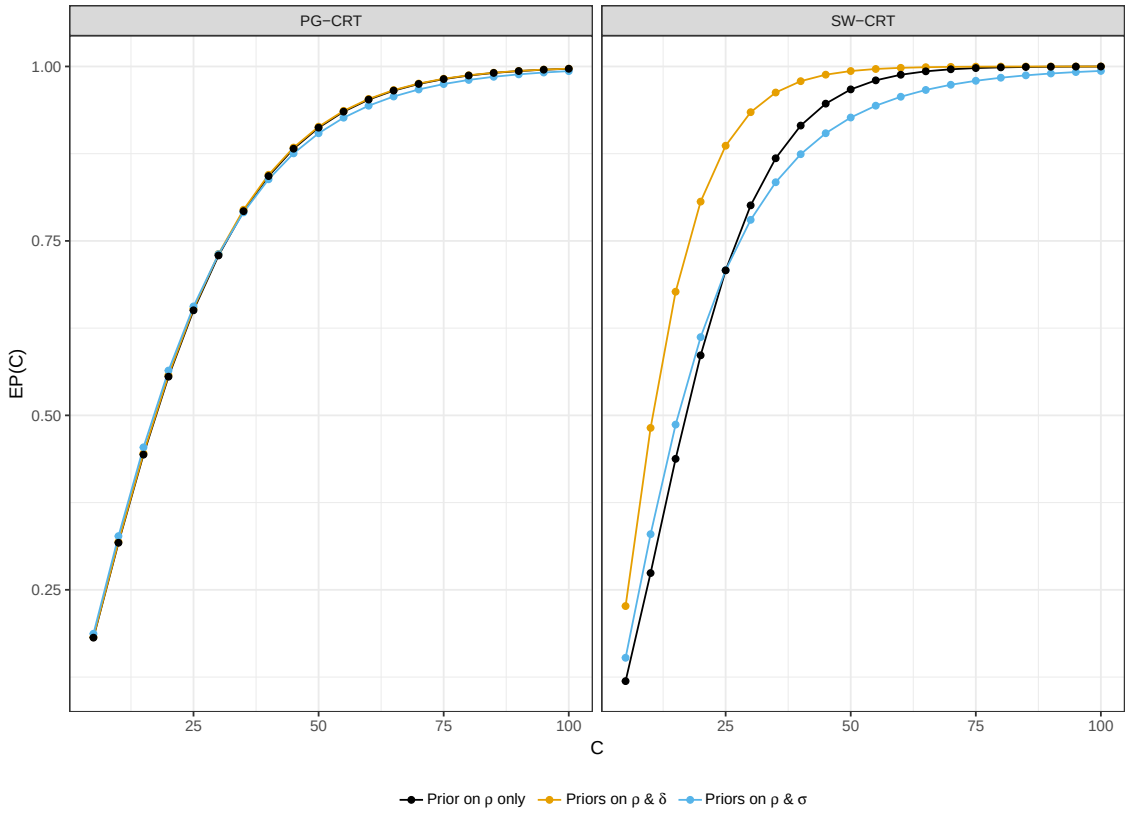

Figure 4.2: Plot of the Expected Power as a function of the total number of clusters (C) based on the above priors for the PG-CRT and SW-CRT examples. The fixed parameter assumptions for both designs were drawn from their respective motivating examples.

Similar to the classical frequentist power, the EP increases as the number of clusters increases, including to 1 as the number of clusters is made very large. As a result of the small SD on the priors making them highly informative, no significant distinction is observed between the prior lines in all three scenarios, particularly in the case of the PG-CRT. Thus, like in a traditional sample size calculation, trials designed within the hybrid framework would simply require the determination of the minimal number of clusters required to achieve the desired EP.

For both designs, having priors on the treatment effect and ICC means the EP curve increases to 1 more quickly. This is a consequence of the prior on the treatment effect being

truncated to only consider cases of effects greater than the MCID. I.e., effects larger than that assumed in the traditional frequentist approach are captured in the hybrid approach, thus resulting in a higher EP.

Equally, versus including priors on the ICC and SD, a prior only on the ICC expedites the convergence of EP towards 1. This is because including a prior on the SD incorporates consideration of power for large values of σ, which will be low.

All of the above implies that the EP of a design, and hence the required sample size, may be highly dependent on the robustness with which a design (PG-CRT vs. SW-CRT) handles the prior weights. This is made clear in subsequent sections, where I investigate the impact on EP by incrementally modifying the values of the prior parameters, transitioning from highly informative to non-informative settings.

4.3.2 Comparison between the frequentist and hybrid approaches

Next, to expand on the above, we compare the two approaches (frequentist and hybrid) to sample size determination in CRTs in more depth, discussing the implications of choosing (a) a particular framework and (b) particular priors in the trials designed within the hybrid framework. For a fairer comparison, we compute required sample sizes based on control of the EP to the same level as that in the frequentist framework (i.e., $\gamma = \beta$) and employ correctly centrally specified priors. This then leaves free choice of the SD of the priors. I, therefore, demonstrate how a 'small' (highly-informative), 'moderate' (weakly informative), or 'large' (non-informative) prior SD affects the sample size required to achieve a desired EP for each design. A plot of all utilised priors is given in Figure 4.3.

Accounting for uncertainty in the ICC and SD

Required sample sizes in the frequentist and hybrid frameworks for a selection of possible priors are presented in this section. In considering Beta and Truncated normal priors for the ICC, matching SDs are selected, so that a fairer comparison can be made between utilising a Beta or Truncated normal distribution. For the comparison in this Chapter, the 'true sample size' is defined as the sample size that would have been selected in the frequentist framework if all assumed parameters were accurate or true, amounting to 550 and 3960 in the PG and SW-CRT respectively.

Observe from Table 4.1 that when priors are correctly centrally specified for both the ICC and the SD, the number of participants required to achieve the desired EP is often higher than when a prior is correctly centrally specified to only the ICC. As discussed above in relation to Figure 4.2, the magnitude of the increase or decrease is determined by the SD of the prior. In particular, Table 4.1 shows that small prior SD for Truncated normal and Gamma priors are the only considered scenarios where the sample size under

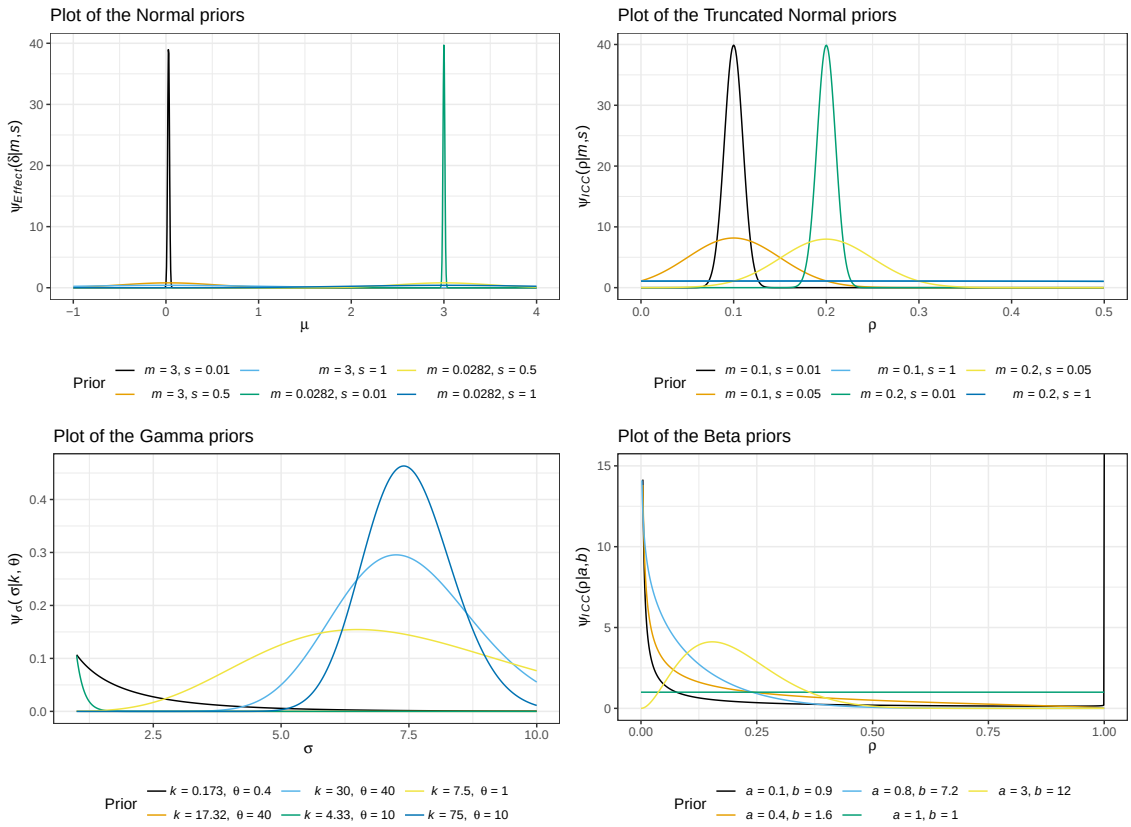

Figure 4.3: Plot of the Gamma, Truncated normal, Beta, and normal correctly centrally specified priors

used in Table 4.1

the hybrid approach is smaller than the frequentist approach for the PG-CRT design. It is worth noting that while large prior SD on the PG-CRT designs requires more participants to achieve the desired EP compared to the frequentist framework, the SW-CRT design required fewer participants due to its capacity to handle large variance as noted above.

These findings highlight the sensitivity of the PG-CRT to variability in the ICC and the SW-CRT design's known efficiency for higher ICC values. Specifically, a high ICC means that the clusters themselves are responsible for most of the outcome variance, hence the within-cluster comparisons facilitated by a SW-CRT become a rich source of information. A consequence is that, perhaps counterintuitively, incorporating larger uncertainty in the ICC can lower the required sample size for a SW-CRT compared to a frequentist approach.

The choice of prior distribution and the level of uncertainty arising from its weightings are critical in the hybrid framework. For example, the uniform prior ($Be(1,1)$), increases the sample size under the PG-CRT relative to the SW-CRT. I later discuss the implications of such priors. Unsurprisingly, when a Beta or Truncated normal distribution with similar densities are used as the prior, the resultant required sample size is similar. Dif-

ferences arise when the desire for a small prior mode results in a Beta distribution with an undefined density at zero. In some settings, it may be the case that an extremely small ICC is a reasonable assumption. In general, though, it is for this reason (along with non-statisticians greater familiarity with the normal distribution) that I prefer the use of a Truncated normal prior for the ICC. In subsequent comparisons, therefore, I shall utilise the Truncated normal prior for the ICC in all instances, and not consider Beta priors further.

	Frequentist		**Hybrid**						
	Parameters	*Sample size*	*Parameters*	$\Psi_{ICC\ only}$	*Sample size*	*Parameters*	Ψ_{ICC}	Ψ_{SD}	*Sample size*
$PG-CRT$									
(Surr et al)	$\alpha = 0.025$		$\alpha = 0.025$	$Be(0.80, 7.20)$	539	$\alpha = 0.025$	$Be(0.80, 7.20)$	$Gamma(75, 10)$	561
	$\beta = 0.1$		$EP = 0.9$	$Be(0.10, 0.90)$	517	$EP = 0.9$	$Be(0.10, 0.90)$	$Gamma(30, 4)$	550
	$N = 0.1$		$\sigma = 7.5$	$Be(1.00, 1.00)$	1738	$N = 11$	$Be(1.00, 1.00)$	$Gamma(7.5, 1)$	2046
	$\sigma = 7.5$	550	$N = 11$	$TN(0, 1, 0.1, 0.01)$	517	$\delta = 3$	$TN(0, 1, 0.1, 0.01)$	$Gamma(75, 10)$	539
	$\rho = 0.1$		$\delta = 3$	$TN(0, 1, 0.1, 0.05)$	539		$TN(0, 1, 0.1, 0.05)$	$Gamma(30, 4)$	572
	$\delta = 3$			$TN(0, 1, 0.1, 1.00)$	1650		$TN(0, 1, 0.1, 1.00)$	$Gamma(7.5, 1)$	1936
$SW-CRT$									
(O'Grady et al)	$\alpha = 0.005$		$\alpha = 0.005$	$Be(3.00, 12.00)$	3930	$\alpha = 0.005$	$Be(3.00, 12.00)$	$Gamma(17.32, 40)$	4230
	$\sigma = 0.426$		$EP = 0.8$	$Be(0.40, 1.60)$	3750	$EP = 0.8$	$Be(0.40, 1.60)$	$Gamma(4.33, 10)$	4380
	$\delta = 0.028$		$\sigma = 7.5$	$Be(1.00, 1.00)$	2550	$N = 11$	$Be(1.00, 1.00)$	$Gamma(0.0173, 0.4)$	1290
	$\beta = 0.2$	3960	$N = 11$	$TN(0, 1, 0.2, 0.01)$	3930	$\delta = 3$	$TN(0, 1, 0.2, 0.01)$	$Gamma(17.32, 40)$	4230
	$\rho = 0.2$		$\delta = 3$	$TN(0, 1, 0.2, 0.05)$	3930		$TN(0, 1, 0.2, 0.05)$	$Gamma(4.33, 10)$	4710
	$T = 7$			$TN(0, 1, 0.2, 1.00)$	2730		$TN(0, 1, 0.2, 1.00)$	$Gamma(0.173, 0.4)$	1410

Table 4.1: Comparison between the frequentist and hybrid approaches for example parameters motivated by Surr et al. (PG-CRT) and O'Grady et al. (SW-CRT); priors correctly centrally specified. Here, priors are placed on the ICC only, and the ICC-and-SD only.

Accounting for uncertainty in the treatment effect and the ICC

In what follows, I compare the required sample sizes in the frequentist and hybrid frameworks when priors are placed on the treatment effect and the ICC.

Observe from 4.2 that as the SD on the treatment effect prior increases, the minimal number of clusters needed to achieve the desired EP in the hybrid approach becomes smaller than in the frequentist approach. Recall that the prior densities on the treatment effect are conditioned on there being a sufficiently large effect (categorised as values greater than the MCID). Here, we set the MCID to be equal to the assumed target effect in the frequentist framework. Thus, when the prior variance is large, the distribution flattens, extending the likely values of μ. Therefore, effects larger than those assumed in the frequentist approach are taken into consideration in the sample size calculation. In turn, fewer clusters are needed to achieve the desired EP compared to that required to achieve the traditional frequentist power. This is consistent with the plot in the Figure 4.2 and echoes the findings for IRTs by Kunzmann and colleagues (Kunzmann *et al.*, 2021).

	Frequentist		**Hybrid**			
	Parameters	*Sample size*	*Parameters*	Ψ_μ	Ψ_ρ	*Sample size*
$PG-CRT$						
(Surr et al)	$\alpha = 0.025$		$\alpha = 0.025$	$N(3, 0.01)$	$TN(0, 1, 0.1, 0.01)$	517
	$\beta = 0.1$		$EP = 0.9$	$N(3, 0.01)$	$TN(0, 1, 0.1, 0.05)$	539
	$N = 11$		$\sigma = 7.5$	$N(3, 0.01)$	$TN(0, 1, 0.1, 1.00)$	1639
	$\sigma = 7.5$	550	$N = 11$	$N(3, 0.05)$	$TN(0, 1, 0.1, 0.01)$	506
	$\rho = 0.1$		$\delta_{MCID} = 3$	$N(3, 0.05)$	$TN(0, 1, 0.1, 0.05)$	528
	$\delta = 3$			$N(3, 0.05)$	$TN(0, 1, 0.1, 1.00)$	1606
	$\delta_{MCID} = 3$			$N(3, 1.00)$	$TN(0, 1, 0.1, 0.01)$	352
				$N(3, 1.00)$	$TN(0, 1, 0.1, 0.05)$	363
				$N(3, 1.00)$	$TN(0, 1, 0.1, 1.00)$	1111
$SW-CRT$						
(O'Grady et al)	$\alpha = 0.005$		$\alpha = 0.005$	$N(0.028, 0.01)$	$TN(0, 1, 0.1, 0.01)$	2610
	$\beta = 0.2$		$EP = 0.80$	$N(0.028, 0.01)$	$TN(0, 1, 0.1, 0.05)$	2610
	$C = 30$		$\sigma = 0.433$	$N(0.028, 0.01)$	$TN(0, 1, 0.1, 1.00)$	1770
	$T = 7.0$	3960	$C = 30$	$N(0.028, 0.05)$	$TN(0, 1, 0.1, 0.01)$	1140
	$\rho = 0.2$		$T = 7.0$	$N(0.028, 0.05)$	$TN(0, 1, 0.1, 0.05)$	1140
	$\sigma = 0.433$		$\delta_{MCID} = 3$	$N(0.028, 0.05)$	$TN(0, 1, 0.1, 1.00)$	720
	$\delta = 0.028$			$N(0.028, 1.00)$	$TN(0, 1, 0.1, 0.01)$	30
	$\delta_{MCID} = 0.028$			$N(0.028, 1.00)$	$TN(0, 1, 0.1, 0.05)$	30
				$N(0.028, 1.00)$	$TN(0, 1, 0.1, 1.00)$	30

Table 4.2: Comparison between the frequentist and hybrid approaches for example parameters motivated by Surr et al. (PG-CRT) and O'Grady et al. (SW-CRT); priors correctly centrally specified. Here, priors are placed on the treatment effect and the ICC.

4.3.3 Sensitivity analysis: Robustness of trials designed within the frequentist and hybrid frameworks to prior misspecification

A major motivation for this work was the dearth of available ICC values and their possible misspecification in practice. Hence, a sensitivity analysis is conducted to assess how misspecification of the assumed frequentist ICC and hybrid prior at the design stage impacts the sample sizes in both frameworks. To illustrate this, I first assume that the parameters from our motivating examples are the 'true' values of the ICC (PG-CRT: $\rho = 0.1$, SW-CRT: $\rho = 0.2$). I then assess how frequentist and hybrid framework-designed trials fare if the assumed ICC (prior modal ICC in the hybrid case) is greater/less than the true value.

The Truncated normal distribution will be used in the hybrid framework to easily enable the ICC prior mode to take on values that are equivalent to the frequentist's misspecified values. In terms of the prior SD, we employ a quantile approach such that (a) all the distribution of the misspecified prior is above/below the true ICC value (Q4), (b) 75% of the distribution is above/below the true ICC value (Q3), (c) 50% of the distribution is above/below the true ICC value (Q2), and (d) 25% of the distribution is above/below the true ICC value (Q1). Here, our emphasis is on the prior for the ICC, since the SD can more often be obtained through a pilot trial. This sensitivity analysis is presented in Table 4.3 and show a plot of the assumed priors in Figure 4.4.

For the PG-CRT, when the prior is misspecified such that all of the distribution is either above or below the true ICC, the hybrid framework requires a smaller sample size than the frequentist framework. This may be advantageous when the prior mode is larger than the true ICC, but would likely result in greater power loss when the prior mode is smaller than the true ICC compared to the frequentist framework. For the remaining hybrid priors, the magnitude of the increase or decrease in sample size compared to the frequentist approach is highly dependent on the percentage of the distribution that is above or below the true ICC. For example, for a misspecified ICC prior with a mode of 0.05, 506 participants are required when 75% of the distribution is below the true ICC, whereas 1056 participants are required to achieve the same desired EP when 25% of the prior is below the true ICC. The key message from this is that choosing the SD of the prior for the ICC in the case of a PG-CRT would be critical; it could rescue significant power compared to misspecification in the frequentist setting, or result in an even larger waste of resources.

With respect to the SW-CRT, the value of the SD is seemingly less critical, and provided a large SD is avoided (Q1) the study sample size would not be far from that truly required. Again, this is a consequence of the SW-CRT design's robustness across possible values of the ICC.

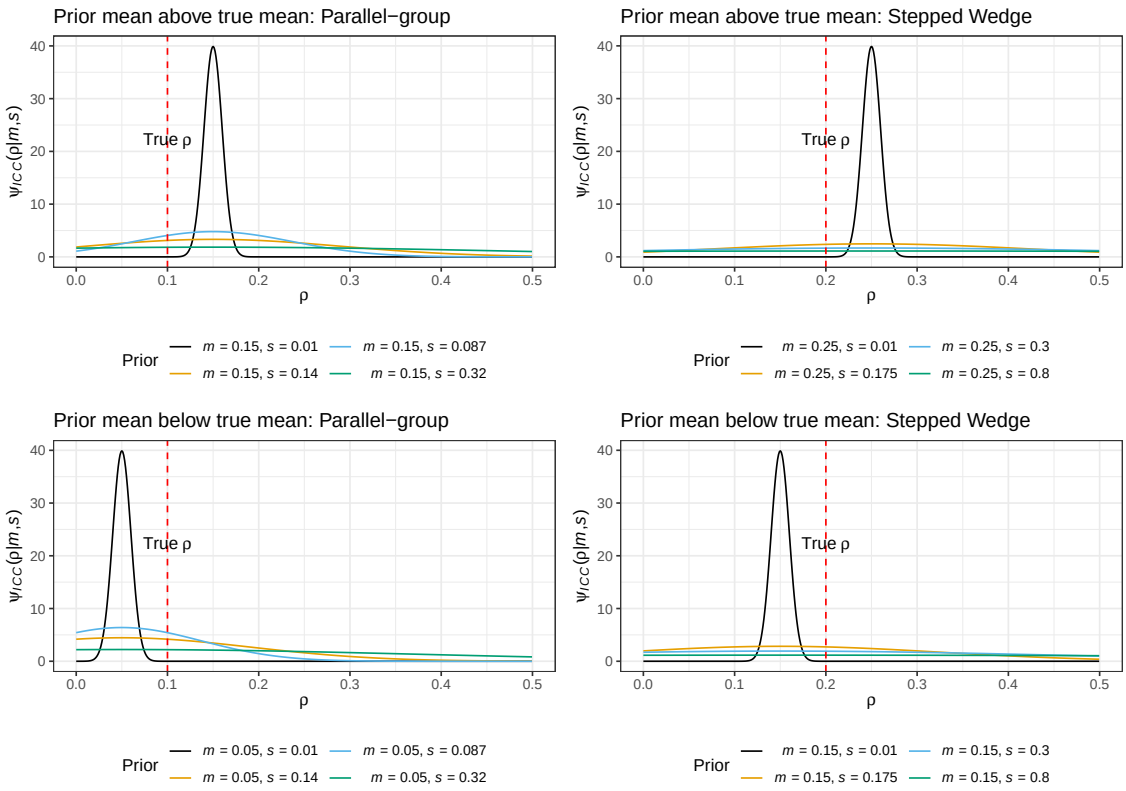

Figure 4.4: Truncated normal prior showing the spread of the assumed ICC misspecifications from the true ICC values.

	Frequentist			**Hybrid**		
	Parameters	*Assumption*	*Sample size*	*Parameters*	Ψ_ρ	*Sample size*
$PG-CRT$						
(Surr et al)	$\alpha = 0.025$	$\rho = 0.05$	396	$\alpha = 0.025$	$TN(0, 1, 0.1, 0.05^2)$	528
	$\beta = 0.1$			$EP = 0.9$	$TN(0, 1, 0.1, 0.05^2)$	429
	$N = 0.1$			$\sigma = 7.5$	$TN(0, 1, 0.1, 0.05^2)$	363
	$\sigma = 7.5$	$\rho = 0.15$	660	$N = 11$	$TN(0, 1, 0.1, 0.1^2)$	605
	$\rho = 0.1$			$\delta = 3$	$TN(0, 1, 0.1, 0.1^2)$	550
	$\delta = 3$				$TN(0, 1, 0.1, 0.1^2)$	418
$SW-CRT$	$\alpha = 0.005$					
(O'Grady et al)	$\sigma = 0.426$	$\rho = 0.25$	3696	$\alpha = 0.005$	$TN(0, 1, 0.2, 0.05^2)$	2610
	$\delta = 0.028$			$EP = 0.8$	$TN(0, 1, 0.2, 0.05^2)$	1980
	$n = 132$			$\sigma = 7.5$	$TN(0, 1, 0.2, 0.05^2)$	1590
	$\beta = 0.2$	$\rho = 0.15$	3960	$\delta = 3$	$TN(0, 1, 0.2, 0.1^2)$	2580
	$\rho = 0.2$			$n = 132$	$TN(0, 1, 0.2, 0.1^2)$	1950
	$T = 7$				$TN(0, 1, 0.2, 0.1^2)$	1590

Table 4.3: Sensitivity analysis of priors for example parameters motivated by Surr et al. (PG-CRT) and O'Grady et al. (SW-CRT); parameter misspecifications. Recall that the true required sample size for the PG-CRT example is 550 participants, while for the SW-CRT it is 3960 participants.

4.3.4 Comparison of the Expected Power provided by PG-CRT and SW-CRT designs

To conclude, I include an important comparison of the EP provided by PG-CRT and SW-CRT designs when a prior is placed only on the ICC. This then serves to extend previous comparisons of which design is more efficient to the case where there is uncertainty in the ICC's value.

The EP is, in this case, dependent on: the number of clusters (C); the number of measurements per cluster (N and nT for the PG-CRT and SW-CRT designs respectively); the number of time periods in the SW-CRT design (T); the standardised effect size (δ/σ); and the prior parameters m and s.

To make the comparison fair, I assume each cluster provides a common number of measurements, setting $n = N/T$ in the SW-CRT designs for specified N. I then provide a comparison of the EP for various combinations of the design parameters C, N, T, δ/σ, m, and s. Figure 4.7 assumes $C = 50$; results, shown to be similar, for other values of C are given in Figures 4.5, 4.6, and 4.8. A black curve is added to each sub-panel to indicate the (m, s)-contour across which the two designs (PG-CRT and SW-CRT) have equal EP.

It is observed that the PG-CRT has larger EP only for very small m and very small s. The maximal values of m and s at which the PG-CRT has larger EP across Figure 4.7 are $m = 0.105$ and $s = 0.163$; both occurring when $N = 30$ and $T = 3$. As the SD of the prior (s) increases, the prior places a larger likelihood on a high ICC, which leads to a SW-CRT design becoming more efficient, even when $m \approx 0$. Of 34 reviewed HTA trials that reported assumed ICC values (see Chapter 3), 90% of these trials assumed an ICC below 0.105. Thus, whether a PG-CRT design would be more efficient than a SW-CRT in practice would have heavily depended on the uncertainty around the ICC's value.

Observe also that the results are sensitive to the values of N and T. Specifically, the region in which the SW-CRT has larger EP increases in size as (i) N is increased for fixed T and δ/σ, or (ii) T is increased for fixed N and δ/σ. The pattern as δ/σ is increased for fixed N and T is more complex, though most often increasing the standardised effect size leads to more comparable performance between the two designs, as both transitions towards a very high EP.

While the SW-CRT design might be more powerful in cases of very high ICC compared to the PG-CRT design, it could face significant challenges if certain sequences lack clusters that contribute substantially to treatment effect estimation. This is because the basic SW assumes that all clusters must be randomised simultaneously. On the other hand, the PG-CRT design is easier to salvage such situations, as it allows for balancing through later randomisation. Considering these practical strengths and weaknesses of the PG and SW designs, I comment in the Discussion on how limiting the choice of a CRT design to only a sample size requirement oversimplifies the process of choosing an optimal CRT design.

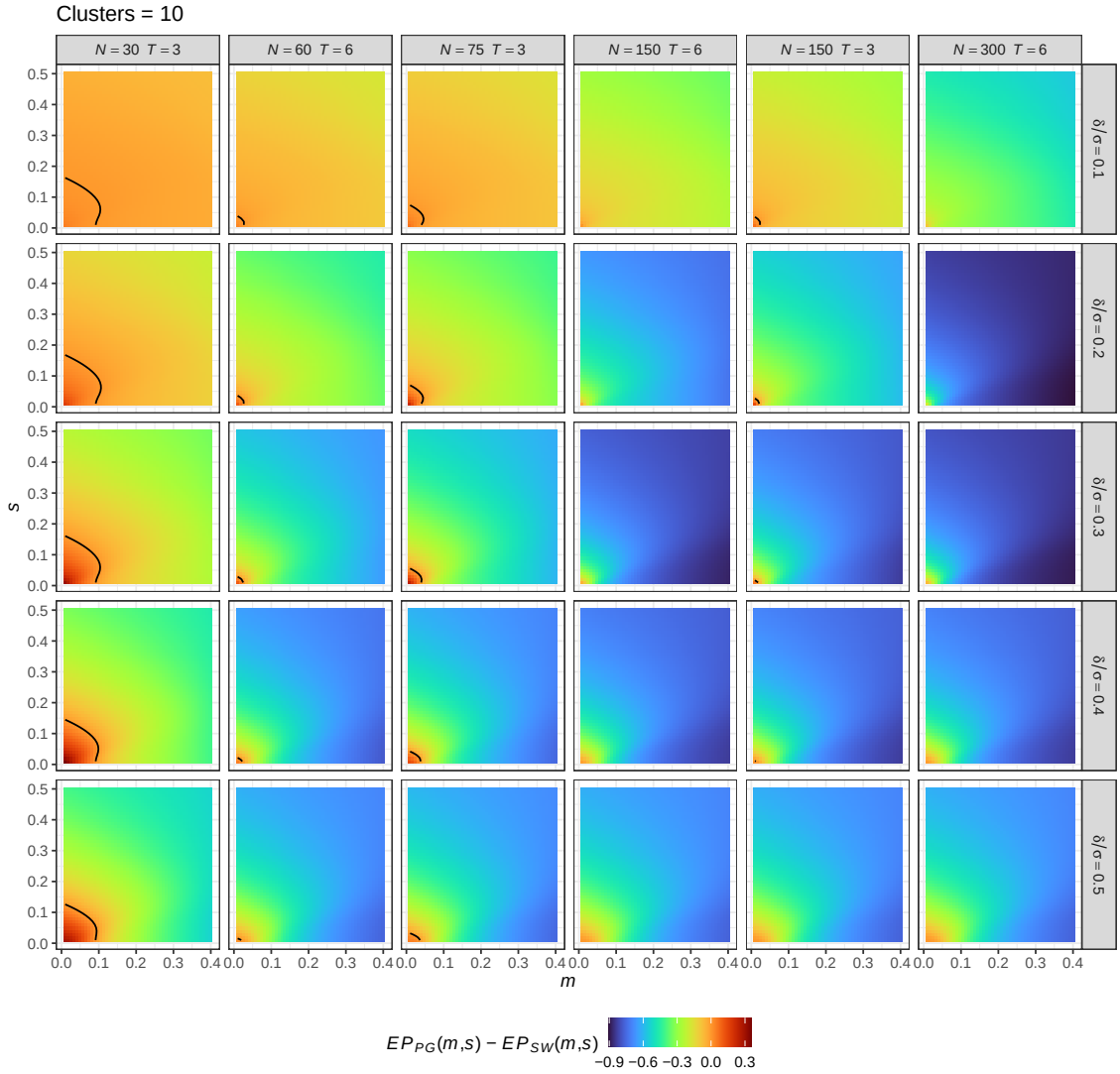

Figure 4.5: Comparison of the Expected Power (EP) provided by PG-CRT and SW-CRT designs for different values of the Truncated normal prior parameters m and s, faceted by the assumed effect size (δ/σ) and assumed values of N (number of participants per cluster) and T (number of time periods in the SW-CRT design). The black curves indicate the point at which the EP is equal for the two designs. Sub-plots without a black curve indicate negative values within the entire region. All results here assume that $C = 10$.

4.4 Discussion

The significance of the ICC to sample size determination and the challenges associated with pre-specification at the design stage have long been discussed in literature (Hade *et al.*, 2010; Korendijk *et al.*, 2010; Pagel *et al.*, 2011). Motivated by this problem, we, therefore, presented the detailed calculations required to take a hybrid approach to sample size calculation that allows for direct incorporation of uncertainty in the ICC, the target effect,

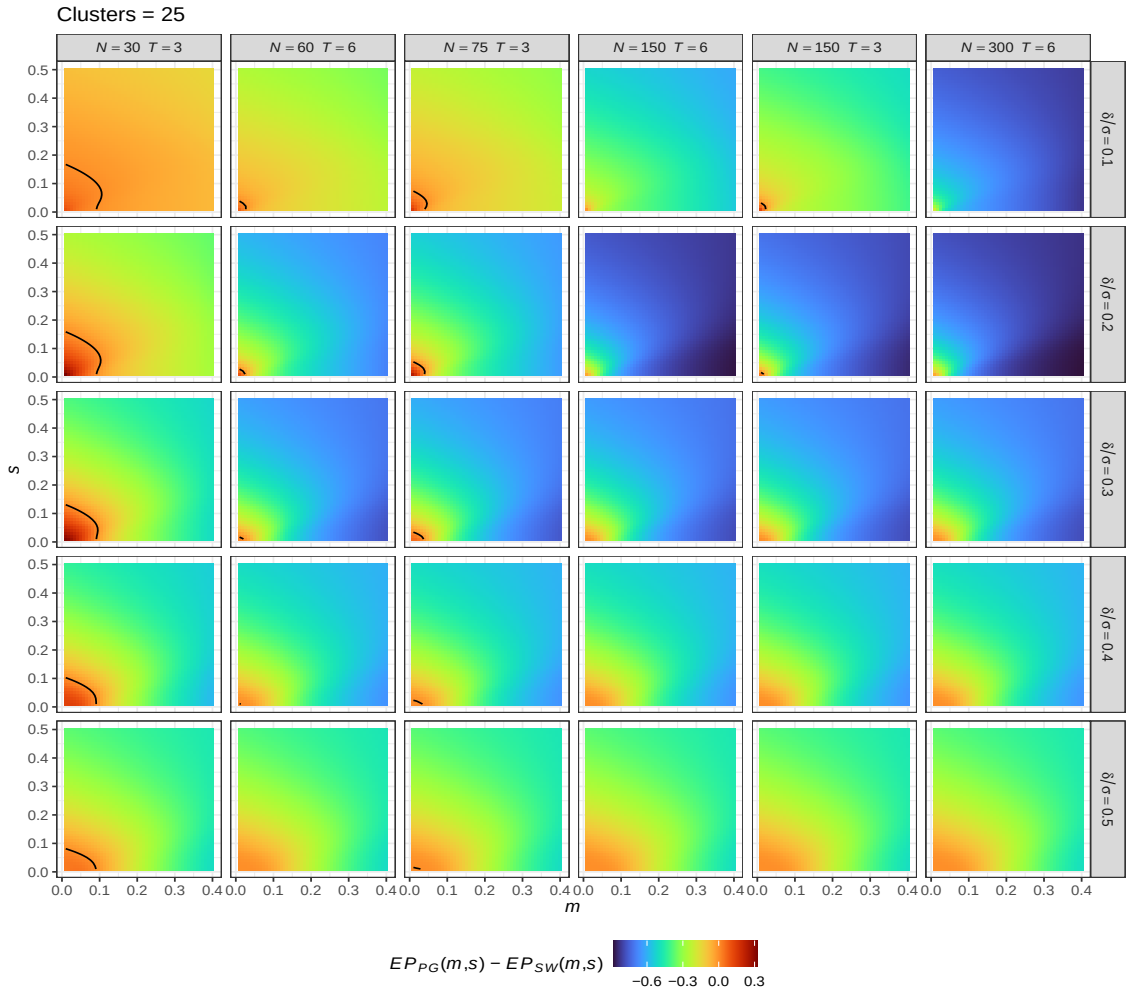

Figure 4.6: Comparison of the Expected Power (EP) provided by PG-CRT and SW-CRT designs for different values of the truncated normal prior parameters m and s, faceted by the assumed effect size (δ/σ) and assumed values of N (number of participants per cluster) and T (number of time periods in the SW-CRT design). The black curves indicate the point at which the EP is equal for the two designs. Sub-plots without a black curve indicate negative values within the entire region. All results here assume that C=25.

and/or the SD. This approach may be advantageous in circumstances where obtaining an accurate ICC estimate during the design stage is problematic, and is more consistent with CONSORT guidance on accounting for ICC uncertainty.

To assess the efficiency of our proposed framework, a sensitivity analysis was performed to show how a high degree of ICC misspecification had less impact on the EP and sample size in the hybrid framework. Conversely, various priors indicating varied levels of uncertainty were also presented to demonstrate when a frequentist approach may also be optimal. Thus, in a real-world trial, a simulation study can be utilised to determine when a hybrid approach may be desirable for a given set of assumptions.

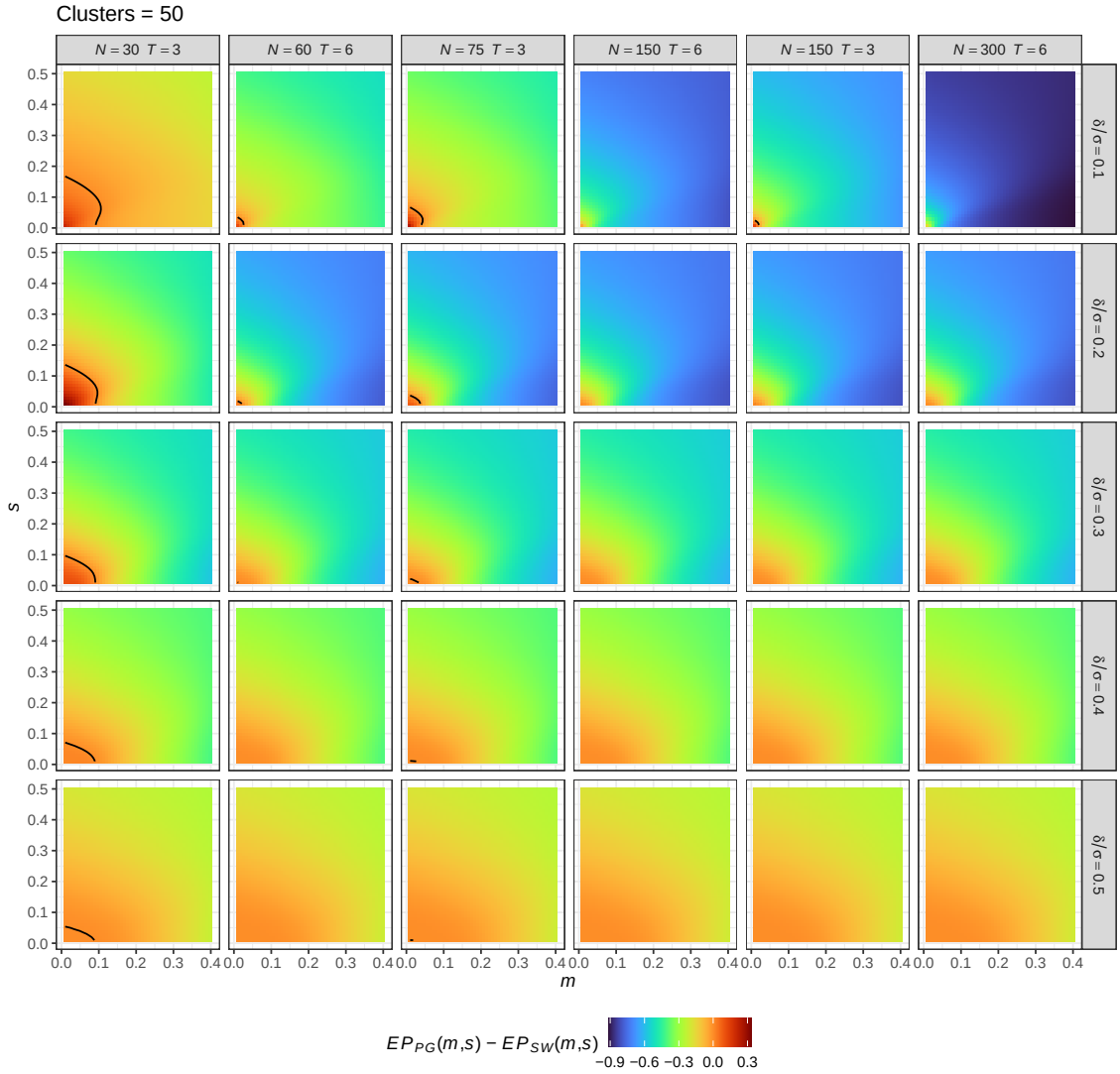

Figure 4.7: Comparison of the Expected Power (EP) provided by PG-CRT and SW-CRT designs for different values of the truncated normal prior parameters m and s, faceted by the assumed effect size (δ/σ) and assumed values of N (number of participants per cluster) and T (number of time periods in the SW-CRT design). The black curves indicate the point at which the EP is equal for the two designs. Sub-plots without a black curve indicate negative values within the entire region. All results here assume that C=50.

An attempt was made to clarify the ambiguity associated with the hybrid quantities (PoS and EP) for sample size determination and scenarios concerning when they can be used interchangeably were proposed. Considering the cost, time consumption, and the high failure rate of trials, robust statistical methods such as those espoused in this study are critical for trial success. This can be more practical when tighter definitions of the probability of success with commensurate formulae are selected by funders and the trial team.

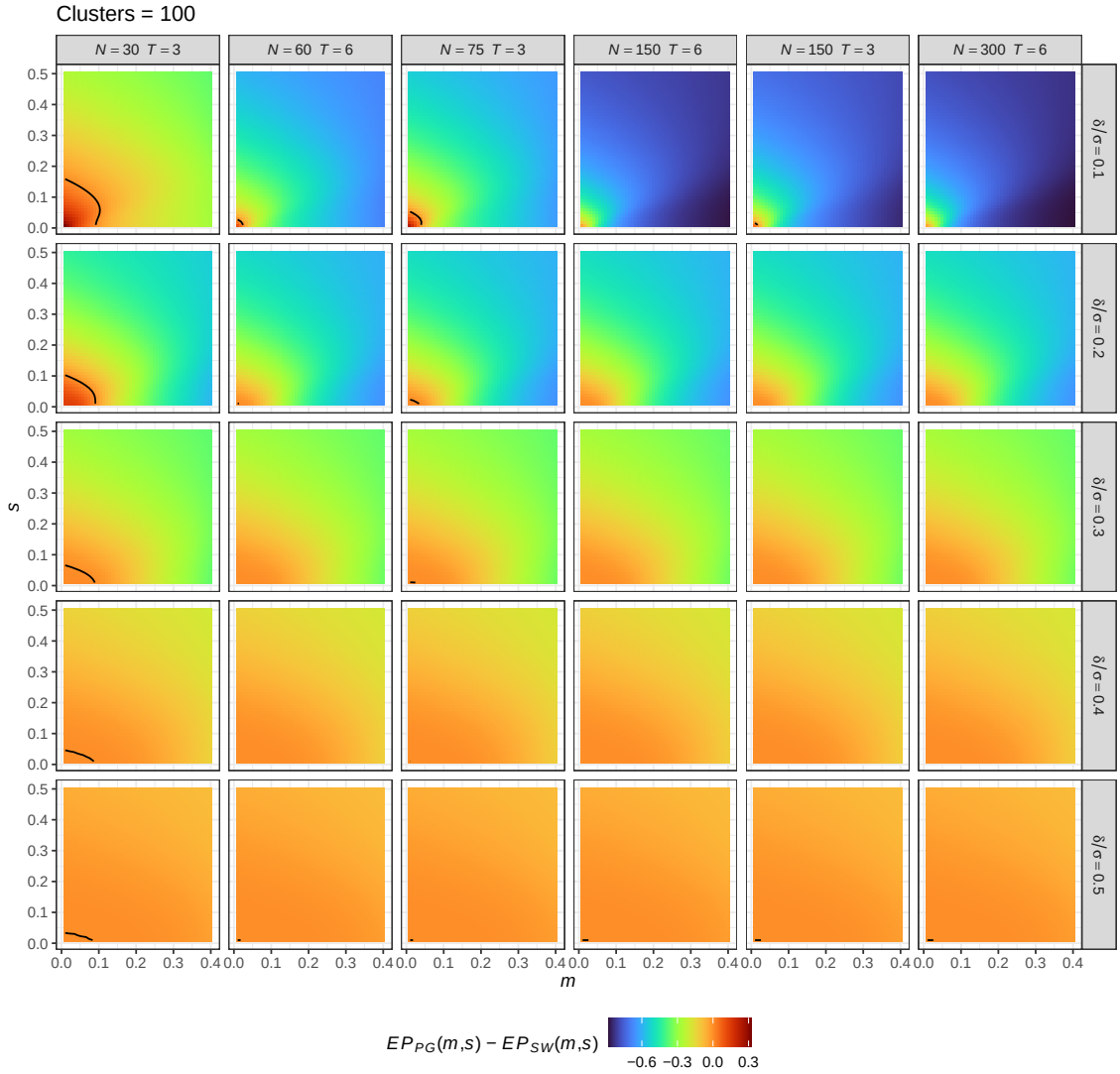

Figure 4.8: Comparison of the Expected Power (EP) provided by PG-CRT and SW-CRT designs for different values of the truncated normal prior parameters m and s, faceted by the assumed effect size (δ/σ) and assumed values of N (number of participants per cluster) and T (number of time periods in the SW-CRT design). The black curves indicate the point at which the EP is equal for the two designs. Sub-plots without a black curve indicate negative values within the entire region. All results here assume that C=100.

Like Kunzmann *et al.* (2021), we argue for the control of the EP in designing and determining the sample size of a trial under the hybrid framework since it typically takes values more comparable to the frequentist power. As others have identified in an IRT setting (Lan & Wittes, 2012; Chen & Ho, 2017), we demonstrated the monotonic relationship between the number of clusters (sample size) and the EP; thus, an increase in sample size increases the EP, and sample size calculation under a hybrid framework for a CRT functions very similarly to the more familiar frequentist approach.

Additionally, it was observed that the EP may lead to lower required sample sizes in certain cases, particularly when a prior is introduced on the treatment effect. Consequently, this could be deemed an efficient and cost-effective tool for trial design, given the routine high costs associated with CRTs. This may have a considerable implication, especially for trials in LMICs, if they have more limited resources. While this is a positive finding, greater caution must be exercised when introducing a prior on the treatment effect as stated earlier. If one chooses to do this in practice, consulting the Cook Report (Cook *et al.*, 2018) for guidance on selecting the effect size could serve as a first step. Also, if a second study is initiated based on promising results from a first study to inform the design of the second study, there could be a risk of bias in the impression of the effect size, which must be taken into account (Rothwell *et al.*, 2022). Given that all these factors were not taken into account when defining the prior on our treatment effect, considering them in future may lead to sample sizes that deviate from those observed in the Results.

A critical consideration of the hybrid method, however, is the choice of prior. In this context, expert opinion could be used to develop appropriate priors, or methodology such as that presented by Turner *et al.* (2004) could be used to form an informative prior distribution. We discourage the use of uninformative priors such as the uniform distribution since they can be informative in some settings. Having observed from our review of HTA trials in Chapter 3, as was also found by Offorha *et al.* (2022), that ICCs in health services research are typically small (≤ 0.1), a uniform prior that places equal weight on all values of the ICC might not be ideal. A corollary to this is that all priors are inherently subjective and possible misspecification cannot be overlooked. Of course, parameter misspecification is also a problem in frequentist design, and effective prior construction may be reasonably anticipated to mitigate the problem of under- or over-powering on average compared to choosing specific parameter values to assume.

When incorporating uncertainty in the ICC, the SW-CRT appears to almost always be a more efficient design relative to the PG-CRT. Specifically, this study showed that the SW-CRT is more efficient when there is higher uncertainty in the ICC ($s \leq 0.16$), even for a small modal ICC assumption ($m \leq 0.1$). This is because a SW-CRT is typically less sensitive (i.e., more efficient, with a lower design effect) for higher values of the ICC, owing to the clusters acting as their own control and its ability to leverage both within and between cluster comparisons. However, it is notable that the region in which the performance between the designs was similar, in terms of the value of m, does correspond for certain N and T to more commonly assumed values for the ICC. Thus, the uncertainty in the ICC, as captured by s, could be a key determinant of which design is more efficient in practice when using a hybrid approach.

Although a comparison between PG-CRT and SW-CRTs when using the hybrid approach is presented in this study, we agree with Hemming & Girling (2013) that one design

cannot be a panacea to all of the issues and complexities of CRTs. While sample size, the measure of efficiency in this paper, is a key determinant of the probability of detecting a significant effect (Van Breukelen & Candel, 2012; Kunzmann *et al.*, 2021), the choice of design to use in a particular context must take into consideration a wide array of factors. In this sense, this study's focus on the efficiency of a CRT design through the required sample size is a simplification of choosing an optimal design in practice.

However, we believe that the implications of an erroneous sample size on statistical power provide reasonable justification for focusing solely on sample size in this study. As cost is also typically a function of sample size, this further key consideration in design choice is arguably also well captured by our focus on sample size (Baio *et al.*, 2015). Thus, the significance of our comparison of PG-CRT and SW-CRT under uncertainty should not be downplayed.

The motivating examples used in this study had continuous outcomes and as such our conclusions cannot be directly extended to settings with binary outcomes. Nonetheless, we do not foresee any theoretical reason that would make the results with a binary outcome (e.g., placing a prior on the control arm response rate) considerably different from the continuous outcome in this study. The standard Hussey and Hughes model was also assumed, and we limited our focus to cross-sectional SW-CRT. Therefore, conclusions cannot be made on closed-cohort SW-CRT designs or for more complex modelling strategies based on our findings. Nonetheless, our approach could be readily extended for such design/analysis scenarios by placing priors on the additional parameters required for closed cohort SW-CRT designs or on, e.g., the autoregressive parameter of a more complex correlation structure (Hooper *et al.*, 2016; Kasza *et al.*, 2019).

Some researchers argue that not all trials require consideration of the ICC during the planning stages, particularly when analysing at the cluster level. An example is a village survey (see, e.g., Mosha *et al.*, 2022), where the ICC was not specified by the researchers during the design stage but later analysed for the outcome data. Whether this approach constitutes best practice is debatable since the CONSORT extension for CRTs and other authors emphasise the importance of considering the ICC in both the sample size calculation and analysis for all trials (Ivers *et al.*, 2011; Campbell *et al.*, 2012). In my opinion, failure to clearly report the ICC at the planning and analysis stages raises questions about the validity of the findings. Hence, the methods proposed here could be employed in situations where it is challenging to obtain precise values of the ICC during the planning of a specific survey, as they may offer some utility.

Chapter 5

A hybrid approach to sample size reestimation in cluster randomised trials

In Chapter 4, the usefulness of allowing for uncertainty in the ICC within sample size calculations for CRTs was established. Nonetheless, a limitation with approaches of the discussed kind is that their utility can be highly dependent on the choice of prior. To address this limitation, an adaptive design is introduced in this chapter where pre-trial knowledge about the ICC is captured by placing a prior upon it, which is then updated at an interim analysis using the study data to reestimate the sample size.

5.1 Introduction

As discussed in the previous chapters, desired precision of information on key design parameters may not always be available at the planning stage of a trial. Evidence of the gravity of this issue can be seen through a review of trials which revealed significant disparities between values utilised in determining the sample size at the planning stage and the estimates obtained from the trial itself (Charles *et al.*, 2009). This suggests that a significant number of trials may either have excessive statistical power or lack the required power.

In CRTs, one critical nuisance parameter that has been much discussed in this thesis is the ICC. Numerous authors have discussed the difficulty in obtaining reliable estimates of the ICC (Campbell *et al.*, 2004; Ip *et al.*, 2011), and the implications of its misspecification on the statistical power of a trial (Murray *et al.*, 2004; Wu *et al.*, 2012). To mitigate against misspecification of the ICC, I established the usefulness of allowing for uncertainty in its value within the sample size calculation in Chapter 4. I discussed in detail the application

of frequentist and Bayesian methods to address uncertainty, highlighted their respective drawbacks, and advocated for a hybrid method as the preferred approach. However, it is important to note that the hybrid approach also has some drawbacks. A particular limitation of note is that their utility can be highly dependent on the choice of prior, with the possibility always present that it may poorly reflect the data from the intended trial (Lan & Wittes, 2012).

Owing to this, a potentially appealing solution is to conduct a pilot study, i.e., a preliminary or small-scale investigation could be conducted before a full-scale clinical trial in order to inform the prior (Friedman *et al.*, 2015). Although pilot studies are primarily designed to assess the feasibility, safety, and potential efficacy, they can additionally be used to calculate nuisance parameters to inform the planning and design of the larger trial (Hawk, 2013; Shanyinde *et al.*, 2011). As pilot studies utilise real data from an initial phase of a trial, they may be expected to provide a better estimate of parameter values than any pre-trial guess, even guesses that account for uncertainty such as those represented through a prior in a hybrid approach (Lake *et al.*, 2002).

There are two types of pilot study, internal and external. External pilot studies are independent studies that are specifically planned and executed separately from the main study (Lancaster *et al.*, 2004). As participants from an external pilot are not incorporated into the main study, the overall cost can increase significantly. Considering the added expense and the routine ethical concerns of medical trials, it might not be reasonable to conduct a large pilot study without incorporating its data into the ultimate inferential process (Bauer & Kohne, 1994). An internal pilot study is integrated into the overall design of the trial (Lancaster *et al.*, 2004). Thus, the final analysis of the results includes all data, without distinguishing that some of the data originated from the internal pilot study.

A key challenge in inference is to then evaluate how any sample size modifications emanating from the internal pilot affect the statistical analysis. The impact of changes in sample size depends on the inference approach, Bayesian or frequentist. If the sample size is adapted based on a data-driven method, this would not raise concerns in a Bayesian analysis. However, to a frequentist, the impact on the statistical analysis would hinge on the specific rule employed for the adjustment. For example, if the variability within groups is used for the sample size modification, then it will have little impact on the type I error rate, with a likely benefit in power, whereas the type I error rate may be inflated if the difference between treatment means is utilised for the sample size adjustment (Wittes & Brittain, 1990; Bauer & Kohne, 1994).

In the context of adaptive design, sample size reestimation (SSRE) can be likened to a trial with an internal pilot. The study uses data gathered during the trial to reestimate sample size parameter(s) at an interim analysis. It is worth noting that there are blinded

and unblinded approaches to SSRE (Grayling *et al.*, 2018). Unblinding of an ongoing trial can carry considerable risk of introducing subsequent biases. Therefore, it is essential to take specific measures, such as establishing an independent data monitoring committee, to mitigate this risk (Kieser & Friede, 2000). Issues in data monitoring and interim analysis of trials are discussed in detail by Grant *et al.* (2005).

Previous work has investigated the potential of SSRE within the context of CRTs, in the parallel-group (PG) (Lake *et al.*, 2002; Schie & Moerbeek, 2014) and stepped-wedge domains (Grayling *et al.*, 2018). Methods proposed in these publications reestimated the total outcome variance, the ICC, the target effect powered for, or some subset of these parameters. The investigations used interim point estimates of the nuisance parameters to update the required sample size, as is typical in the frequentist literature. Though the methods performed well on average, the variability in their reestimated required sample size could undermine their utility in practice (Hemming *et al.*, 2021). This is a consequence of challenges with precision of estimation in CRTs, particularly in relation to estimating the ICC, which can be difficult to estimate even on completion of a large trial. That is, existing frequentist methods for SSRE in CRTs neglect any uncertainty in the interim point estimates of the nuisance parameters. I demonstrated in Chapter 4 that the hybrid approach may provide utility at the design stage of a CRT. Thus, a natural question of interest is whether a hybrid approach could also be useful for SSRE.

Consequently, in this chapter, I develop a hybrid approach to SSRE for PG-CRTs. This is achieved by assuming a prior for the ICC at the design stage of the trial. This prior is then updated at an interim analysis, to a posterior, based on available data. The posterior is then used in determining the reestimated required sample size to control the EP to a desired level. Following accrual of the reestimated sample size, the final analysis uses all available data in a conventional frequentist analysis to determine whether the null hypothesis can be rejected. Such a hybrid approach to SSRE seems intuitively appealing, as it may directly account for uncertainty in the ICC through both a pre-trial prior and also by considering the variability of the interim ICC estimate. To ascertain whether this approach is useful in practice, we explore both blinded and unblinded SSRE methods. A blinded procedure in this context implies that the treatment status of an observation is undisclosed, but there is awareness of which observations belong to the same cluster while treatment allocation is known in unblinded procedure (Grayling *et al.*, 2018). Subsequently, I perform a comparison between the existing frequentist and our proposed hybrid approach.

In the next section, the methodology underpinning the SSRE in CRT design is introduced by initially describing the setting, along with the notations and the analysis model. Section 5.2.2 offers a high-level summary of how SSRE is performed in both the frequent and Hybrid paradigms. Following that, Section 5.2.5 outlines the simulation study, detail-

ing the parameters derived from the motivating example that will serve as the foundation for the results. The outcomes of the simulations, comparing the hybrid and frequentist SSRE, are presented in Sections 5.3, with a comprehensive discussion of the results in Section 5.4.

5.2 Methods

First, we describe how SSRE can be performed in both the frequentist and hybrid frameworks. As increasing the number of clusters typically has a bigger impact on power than increasing the cluster size does (Campbell *et al.*, 2012), we focus on interim updating of the required number of clusters throughout. Updating the required cluster size, or updating both the cluster size and number of clusters, could be treated similarly.

5.2.1 Setting and notation

We consider the case of a PG-CRT where clusters are randomised to receive an experimental or a control treatment in a 1:1 manner. We assume the primary outcome is continuous and normally distributed with variance σ^2. While we acknowledge that sample sizes can vary between clusters, we restrict our attention to assume the same number of participants are present in each cluster. Accordingly, let Y_{ij} be the outcome from patient $i = 1, \ldots, n$ in cluster $j = 1, \ldots, C$. Then, due to non-independence in the data, we fit a linear mixed model to the data at both interim and final analyses. At the final analysis, and at the interim analysis in unblinded SSRE procedures, the model is given by

$$Y_{ij} = \theta + X_j\mu + c_j + e_{ij}. \tag{5.1}$$

Here, θ is an intercept term (here, the mean in the control arm), $X_j = 1$ if cluster j is allocated to the experimental arm and $X_j = 0$ otherwise, $c_j \sim N(0, \sigma_c^2)$ is a random effect for cluster j, and $e_{ij} \sim N(0, \sigma_e^2)$ is the individual-level error. Note that $\sigma^2 = \sigma_c^2 + \sigma_e^2$, and that the ICC $\rho = \sigma_c^2/\sigma^2$. Then, μ is the treatment effect of interest and we specify our one-sided null hypothesis as $H_0 : \mu \leq 0$. The test statistic for H_0 is

$$t = \frac{\hat{\mu}}{\sqrt{Var(\hat{\mu})}},$$

which can be computed using, e.g., REML estimation. The degrees of freedom for the test statistic will be assumed to be that in a corresponding balanced ANOVA analysis. That is, $df = nC - C - 1$. We note that in the case where C is small, alternative degrees of freedom specifications may be needed. For a target type I error of α, H_0 will be rejected if t is greater than the $(1-\alpha)$-quantile of a $t_{df}(0)$ distribution.

At the interim analysis in blinded SSRE procedures, where cluster assignment is unknown, the model fitted is instead (with parameters interpreted above):

$$Y_{ij} = \theta + c_j + e_{ij}.$$

In either case, the interim analysis results in estimates that we denote by $\hat{\sigma}^2_{c,int}$ and $\hat{\sigma}^2_{e,int}$, which can be combined into an interim estimate of the ICC $\hat{\rho}_{int} = \hat{\sigma}^2_{c,int}/(\hat{\sigma}^2_{c,int} + \hat{\sigma}^2_{e,int})$.

5.2.2 Sample size reestimation procedure

A high-level summary of how SSRE functions (independent of which statistical framework it is conducted in) is as follows.

First, a sample size is chosen for when the interim analysis will occur. As we assume n is fixed, this corresponds to selecting a certain number of clusters, C_{int}, from which data will have been collected at the interim analysis. This could be achieved by utilising some proportion of an initially calculated sample size based on assumed values for required parameters. Alternatively, a pragmatic sample size could be selected, e.g., based on the number of clusters required to achieve a sufficiently precise estimate of the ICC.

The trial is then conducted until the interim required sample size is achieved and the ICC estimated (i.e., $\hat{\rho}_{int}$ is computed). Given the value of $\hat{\rho}_{int}$ (and using other selected design parameters, e.g., the target type I error rate), the required sample size is re-estimated. That is, a value for the final target number of clusters, C_{reest} is computed. It is the interim estimation of ρ (blinded or unblinded) and the method of utilising $\hat{\rho}_{int}$ to compute C_{reest} (frequentist or hybrid) that will differ between the compared methods.

Next, if $C_{reest} \leq C_{int}$, the study terminates and the final analysis is conducted. Otherwise, the trial continues until data from C_{reest} clusters has been accrued, with the final analysis then conducted using data from both stages. This final analysis is conducted using the approach outlined above (i.e., without adjustment for the inclusion of the interim analysis); thus consideration of the potential for type I error inflation will be important.

5.2.3 Sample size reestimation in the frequentist framework

The classical method of sample size estimation for a PG-CRT in the frequentist framework is to first calculate the sample size required for a corresponding IRT, and then multiply it by a 'design effect' (or 'variance inflation factor') to account for clustering. The sample size for the IRT (N_{IRT}), assuming power of $1 - \beta$ is desired when $\mu = \delta > 0$, is obtained as

$$N_{IRT} = \frac{4(z_{1-\alpha} + z_{1-\beta})^2\sigma^2}{\delta^2}.$$

Then, the design effect for the considered type of PG-CRT is given by

$$DE(\rho) = 1 + (n-1)\rho. \tag{5.2}$$

Hence, if there are n measurements per cluster, the required number of clusters is, for a particular value of ρ

$$C(\rho) = N_{IRT} DE(\rho)/n \tag{5.3}$$

In SSRE within the frequentist framework, the interim estimated ICC ($\hat{\rho}_{int}$) is simply inserted into Equation 5.3. That is, the method sets $C_{reest} = C(\hat{\rho}_{int})$.

5.2.4 Sample size reestimation in the hybrid framework

Sample size calculation in the hybrid framework amounts to averaging the frequentist power over any uncertainty in nuisance parameters by placing priors on these parameters. In this framework, two quantities are commonly used for sample size determination: the EP and the PoS. In this work, where a prior is placed only on the ICC, the standard definitions of the PoS and EP become equivalent (as demonstrated in Chapter 4) and can be expressed as:

$$EP(\psi, \mu, C) = \int_0^1 P(\mu, n, C, \alpha, \sigma, \rho)\psi_{ICC}(\rho|\theta)\ \mathrm{d}\rho,$$

where $P(\mu, n, C, \alpha, \sigma, \rho)$ is the probability of rejecting H_0 under a PG-CRT design, given by

$$P(\mu, n, C, \alpha, \sigma, \rho) = \Phi\left\{\mu\sqrt{\frac{Cn}{4\{1+(n-1)\rho\}\sigma^2}} - \Phi^{-1}(1-\alpha)\right\},$$

and $\psi(\rho|\theta)$ is the prior density for an ICC of ρ, which is dependent on parameters θ. Assuming an EP of $1-\gamma$ is then desired when $\mu = \delta > 0$, sample size calculation is performed by numerically searching for the minimal C such that $EP(\psi, \delta, C) \geq 1-\gamma$. In practice, γ is often set to be equal to the value of β from the frequentist framework.

Since the ICC (typically) ranges between 0 and 1, we select a prior distribution with support on $[0, 1]$. Note that for simplicity and efficiency, a conjugate prior distribution may be desirable. However, conjugacy cannot be achieved in this study since the likelihood for the ICC is complex: this requires approximation using JAGS to make sampling from the non-conjugate posterior distribution possible. We achieve this through the rjags package in R Plummer (2021). We select a truncated normal distribution, truncated on $[0, 1]$, as the prior for this study. We acknowledge that other distributions, such as the Beta distribution, could similarly be used. Nonetheless, an evaluation of alternative priors in Chapter 4 revealed no significant sensitivity to the exact choice of prior, given they held

(approximately) the same mean and variance. Therefore, for simplicity, we consider only a truncated normal prior and denote the choice by $TN(0, 1, m, s)$, where m and s are the mean and SD parameters of the normal distribution before truncation.

For normally distributed outcome data, Ukoumunne (2002) proposed an approximation for the variance of an ICC estimate using Fisher (1970). Using this, it is assumed that:

$$\hat{\rho} \sim N\left\{\rho, \frac{2(1-\rho)^2[(1+(n-1)\rho]^2}{n(n-1)C}\right\}.$$

Adopting this as the likelihood, the posterior for ρ can be determined at the interim analysis. Consequently, on calculating $\hat{\rho}_{int}$ at the interim analysis, the prior $TN(0, 1, m, s)$ can then be updated to a posterior that is a function of $\hat{\rho}_{int}$, n, C_{int}, m, and s, which we denote for brevity as $\psi(\rho|m, s, \hat{\rho}_{int})$. The posterior is then substituted into the EP to update the sample size. That is, the method sets C_{reest} as the minimal value such that $EP\{\psi(\rho|m, s, \hat{\rho}_{int}), \delta, C_{reest}\} \geq 1 - \gamma$.

5.2.5 Simulation study

The parameters for our simulation study are based on the study by Hankonen *et al.* (2016) which sought to reduce adolescent sedentary behaviour by improving physical activity. The study assumed no ICC at the beginning of the trial. At an interim analysis, they calculated the ICC to be $\hat{\rho}_{int} = 0.059$ and specified the sample size using frequentist methodology. An internal pilot of 25 clusters with an average cluster size of 17 was used to estimate the ICC at the interim analysis. Accordingly, we set $C_{int} = 26$ (to allow equal allocation of clusters to the control and experimental treatments) and $n = 17$. The study desired 80% power ($\beta = 0.2$) to detect a difference of $\delta = 0.3$ for an SD of $\sigma = 1.3$ and $\alpha = 0.025$.

To evaluate the performance of the SSRE techniques, we conducted a thorough simulation study. Specifically, we wanted to evaluate how varied values of the prior parameters m and s impacted the operating characteristics. We consider $m = 0.01, 0.059, 0.1$ to give a range of possible concordances of the prior densities in relation to $\hat{\rho}_{int} = 0.059$. The value of the standard deviations s were selected as $s = 0.01, 0.1, 1$, such that the priors were highly informative, weakly informative, and approximately non-informative respectively. These priors are shown in Figure 5.1.

We consider the performance of the hybrid approach for these m and s, alongside the performance of the frequentist method, for both blinded and unblinded SSRE, for a range of possible values of ρ. We also complete this work under the null ($\tau = 0$) and the alternative ($\tau = \delta$), in order to empirically estimate the type I error rate and power of the various SSRE procedures. For simplicity, we follow previous works in setting $\gamma = \beta$.

For each combination of assumed parameters and particular SSRE approach, we empirically compute three measures to assess performance, based on the results of 10,000

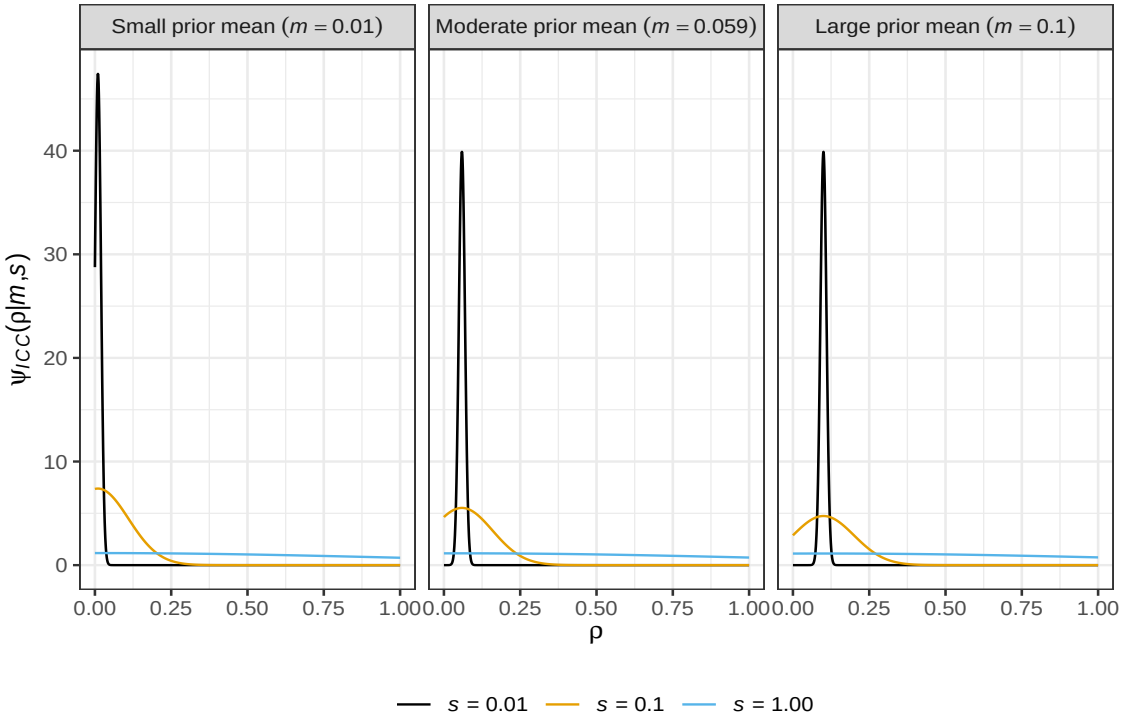

Figure 5.1: Plot of utilised truncated normal prior distributions. Plots are faceted by the use of $m = 0.01, 0.059, 0.1$ and all combinations of $s = 0.01, 0.10, 1$ are considered.

simulation runs. Firstly, the probability of rejecting H_0 is estimated. Our other two metrics relate to the ability of the SSRE procedures to reliably specify the 'correct' required number of clusters. That is, we think of the goal of each SSRE procedure being to 'estimate' the sample size that would have been used if the true values of the design parameters were known. Thus we consider the re-estimated sample size as an estimator, with the target estimand being the 'oracle' sample size that would have been chosen in the frequentist framework if all parameters were known. Natural measures of the performance of these estimators are then its bias and mean square error (MSE); it is these we also compute. A SSRE that performs well will have a bias close to 0 and a low MSE. The selection of the MSE as a metric is based on its ability to penalise designs with both incorrect mean sample sizes and larger variances in sample sizes. If the values of the re-estimated required number of clusters from the simulation replicates are $C_{reest,1} \dots C_{reest,10000}$, the empirical bias and MSE are given by:

$$Bias = \frac{1}{10000}\sum_{n=1}^{10000} C_{reest,i} - C(\rho),$$
$$MSE = \frac{1}{10000}\sum_{n=1}^{10000} [C_{reest,i} - C(\rho)]^2.$$

Here, $C(\rho)$ is the 'oracle' required number of clusters for the particular value of ρ assumed in the simulations that generated $C_{(reest,1)} \dots C_{(reest,10000)}$.

5.3 Results

To provide some intuition on how the choice of prior can influence the reestimated required number of clusters in the hybrid framework, we present a plot of posterior modes in Figure 5.2. I.e., modal values of $\psi(\rho|m, s, \hat{\rho}_{int})$, over $\rho \in [0, 1]$, are given as a function of m, s, and $\hat{\rho}_{int}$.

Figure 5.2 shows an interplay between $\hat{\rho}_{int}$ and the posterior mode, given the prior mean and SD. Generally, as the interim estimate of the ICC increases ($\hat{\rho_{int}} > 0.13$), a monotonic relationship between the variables is observed where large prior mean and SD values result in a large posterior mode and vice versa. While the rate of increase appears constant, the impact of the SD on the posterior mode diminishes as the SD becomes large. We note how the lines for the posterior mode practically merge into a single line when the prior is non-informative ($s = 1$), in contrast to the multiple lines that are clearly visible when the prior is highly informative ($s = 0.01$). This implies that the final sample size determination is not heavily dependent on the prior mean when the prior is non-informative and vice versa. That is, an inaccurate prior mean will have less impact on the final sample size if the prior is non-informative. Although the posterior mode lines for the 'weakly' informative prior ($s = 0.1$) appear somewhat distinct, they are not as widely separated as for the informative priors.

5.3.1 Reestimated sample size, power, and type I error rates for correctly specified priors

Next, we evaluate the distribution of the reestimated required number of clusters, the power, and the type I error rate, for selected priors in the hybrid framework. Specifically, we assume that $\rho = 0.059$, and that the priors are 'correctly specified' (i.e., $m = \rho$). Then, we explore how on average, a highly informative prior, a weakly informative prior, and a non-informative prior impact the performance of the SSRE procedure in the hybrid framework. Our results are stratified by the use of blinded or unblinded SSRE, and are also

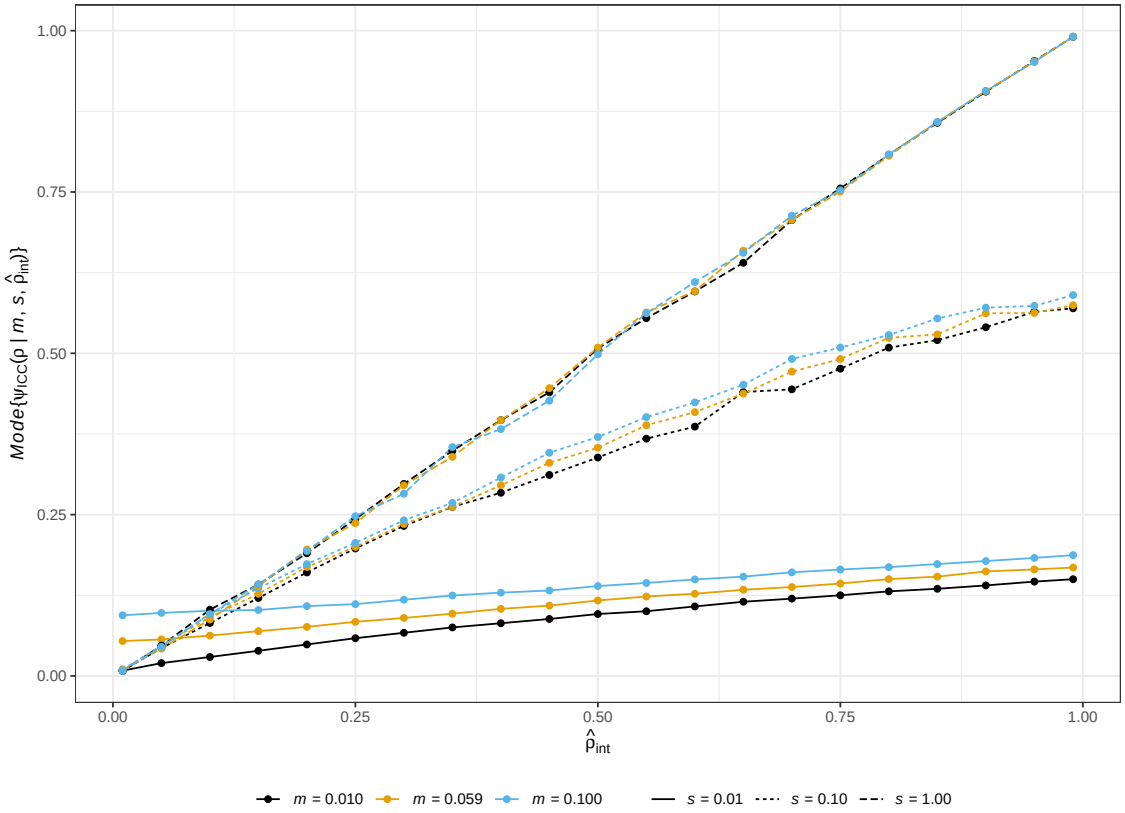

Figure 5.2: Plot of the posterior mode as a function of $\hat{\rho}_{int}$, given the prior mean and SD for all combinations of $m = 0.010, 0.059, 0.100$ and $s = 0.01, 0.10, 1.00$.

compared to the performance of the frequentist approach. A summary of the performance measures for the SSRE procedures is given in Table 5.1, while the distribution of the re-estimated required number of clusters is presented in more detail in Figure 5.3. Note that for the parameters from the motivating example, when $\rho = 0.059$, 68 clusters are required for a frequentist power of approximately 80%.

On average, both the hybrid and frequentist approaches can mitigate against the implications of misspecifying the ICC at the trial's design stage, as indicated by their mean interim estimates of ρ. As the SSRE techniques leverage the interim estimate of the ICC to make a final determination of the sample size, the likelihood of obtaining an accurate sample size is dependent on the closeness of the interim estimate to the truth. However, unlike the frequentist approach whose final sample size is dependent only on ρ_{int}, the final sample size in the hybrid framework is a function of the mean interim estimates and other parameters which include the prior SD. Thus, although the frequentist and hybrid approaches yield the same mean interim ICC estimates, their final average reestimated required number of clusters sample sizes may differ, as seen in Table 5.1.

As a result of the monotonic relationship in Figure 5.2, the required sample size in

the hybrid framework increases as s increases. For $s = 1$, the difference in sample sizes between the hybrid and frequentist frameworks is relatively small, with this phenomenon expected based on results from Chapter 4. Explicitly, a maximum increase of 10% in sample size is observed between the frequentist approach and the use of a non-informative prior. In this setting, this small increase in sample size may be considered beneficial if it translates into power being more reliably above the desired level.

Of importance in the SSRE procedure is the control of the type I error rate. There is some evidence to suggest in Table 5.1 that this is better controlled in the hybrid framework, for both the blinded and unblinded models, though the differences are small when allowing for the simulation error.

In both frameworks, the interim ICC estimates from the blinded model are biased when there is a non-zero treatment effect (i.e., where $\tau = \delta$). Thus, the model overestimates the interim ICC on average and reestimates a larger required number of clusters. Although regulatory agencies prefer blinded models (Friede & Kieser, 2013; FDA, 2019), this may be less necessary in CRTs as cluster allocations are not always blinded.

Interim Model	**Framework**	**Assumed Prior** ψ	**Mean of** $\hat{\rho_{int}}$		**Mean of** C_{reest}		**Power**	**Type I error rate**
			$\tau = 0$	$\tau = \delta$	$\tau = 0$	$\tau = \delta$		
Blinded	Frequentist	N/A	0.0583	0.0711	68	75	0.80	0.029
Blinded	Hybrid	$TN(0, 1, 0.059, 0.01^2)$	0.0583	0.0711	68	69	0.80	0.026
Blinded	Hybrid	$TN(0, 1, 0.059, 0.10^2)$	0.0583	0.0711	73	79	0.83	0.026
Blinded	Hybrid	$TN(0, 1, 0.059, 1.00^2)$	0.0583	0.0711	75	82	0.84	0.026
Unblinded	Frequentist	N/A	0.0583	0.0583	68	68	0.80	0.033
Unblinded	Hybrid	$TN(0, 1, 0.059, 0.01^2)$	0.0583	0.0583	68	68	0.80	0.027
Unblinded	Hybrid	$TN(0, 1, 0.059, 0.10^2)$	0.0583	0.0583	73	73	0.83	0.031
Unblinded	Hybrid	$TN(0, 1, 0.059, 1.00^2)$	0.0583	0.0583	75	75	0.84	0.030

Table 5.1: A summary of the performance of several sample size reestimation procedures is shown for the case where $m = \rho = 0.059$.

Having demonstrated from Table 5.1 that there is no considerable difference in the average value of C_{reest} for both frameworks, we next examine the variability in this quantity. Particularly, our interest lies in whether the hybrid framework has lower variability than the frequentist framework, which is known to have high variability in terms of the reestimated required sample size. Concerning the hybrid framework, we again evaluate how the selected prior SD impacts the performance. Our findings are shown in Figure 5.3.

In comparison to the hybrid approach, the frequentist approach results in a higher variability in the reestimated required number of clusters. This is evidenced by the lower variance and lower interquartile ranges recorded in the hybrid framework compared to the frequentist. Furthermore, the variability is significantly lower for a highly informative prior, with variance increasing as the prior becomes less informative. However, despite the relative variability increase that results from a large prior SD, the variability in these scenarios is still lower than in the corresponding frequentist approach.

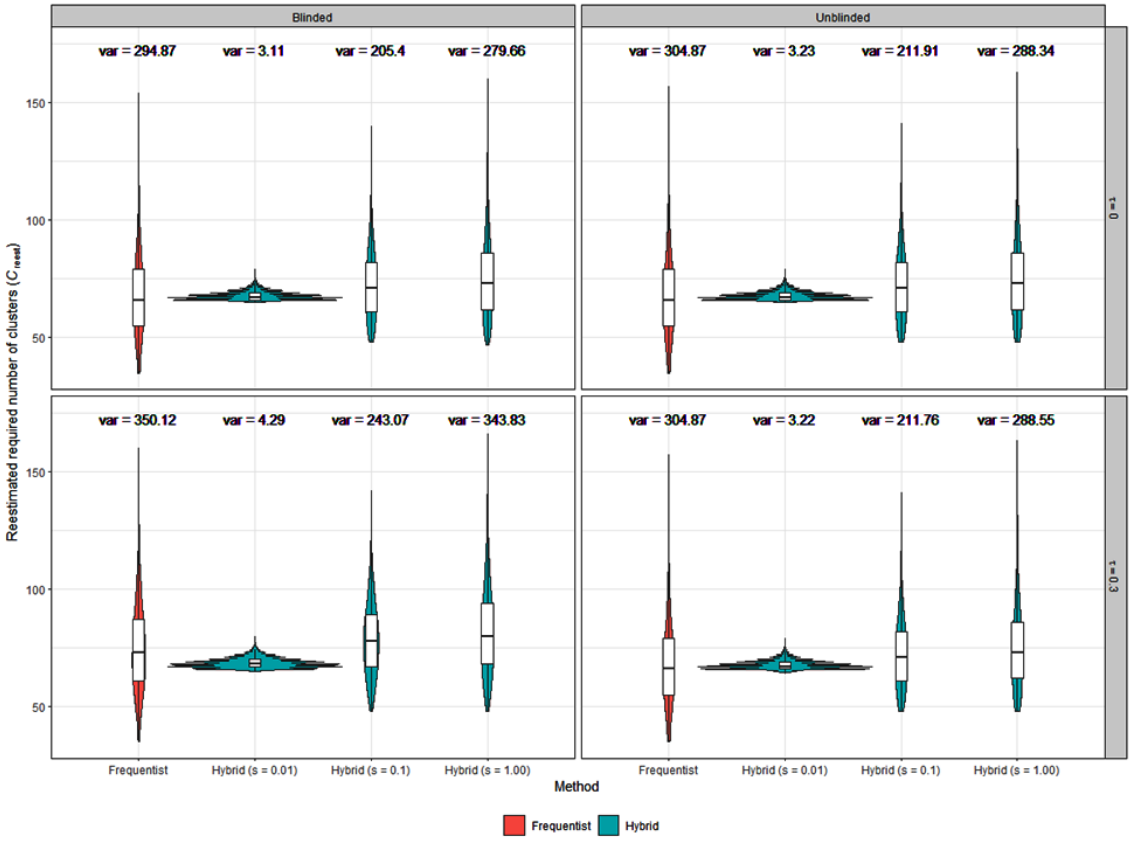

Figure 5.3: Violin and boxplots showing the variability in the reestimated sample sizes (C_{reest}) for the frequentist and hybrid methods ($s = 0.01, 0.1, 1$), with the respective variances ($Var(C_{reest})$) also displayed. Results are faceted by the use of blinded vs. unblinded sample size reestimation and the value of the treatment effect. In all cases, $m = \rho = 0.059$ is assumed.

5.3.2 Impact of prior misspecification on SSRE performance

The results above correspond to when $m = \rho$. In practice, this is unlikely to be the case as SSRE is specifically utilised in scenarios where the ICC is subject to considerable uncertainty. Accordingly, in what follows, we evaluate the performance of a range of SSRE methods across possible values of the ICC, specifically $\rho \in [0.01, 0.2]$. We present our findings in Figures 5.4-5.6.

The frequentist framework seems relatively stable in terms of the performance measures across the considered values of ρ. As expected, the final sample sizes in the frequentist approach are unbiased for the unblinded model, and subject only to small bias for the blinded model.

For highly informative priors, the hybrid approach is only approximately unbiased in terms of the reestimated sample size when the prior mean is equal to ρ, for both blinded and unblinded models. The final sample size in the hybrid SSRE is considerably underestimated when the value of the ICC is larger than the prior mean. Given that an underestimated sample size results in an underpowered trial and vice versa, the negative relationship between the bias and ICC is also observed in the power. When compared to the frequentist method, the hybrid techniques offer a lower MSE if the ICC is within a specific range, with the range dependent on the values of m. For example, when $m = 0.01$, the hybrid method has lower MSE if the ICC is less than 0.05 for both blinded and unblinded models. Whereas, when $m = 0.059$, a lower MSE is observed in the hybrid framework if $\rho \in [0.03, 0.12]$ in the blinded model and if $\rho \in [0.03, 0.1]$ in the unblinded model. When $m = 0.1$, the hybrid method can reduce the MSE if $\rho \in [0.06, 0.17]$ in the blinded model and if $\rho \in [0.06, 0.15]$ in the unblinded model.

When using weakly informative priors, the interplay between bias and power is also exhibited in the same way as when using highly informative priors. However, the power curves for weakly informative priors are not too steep unless the ICC is very small. As a result, there is a maximum 5% loss or gain in power compared to the desired power over a wide range of ICCs, specifically, when $\rho \in [0.025, 0.2]$. In terms of the MSE, when $m = 0.01$, the hybrid framework performs better when the ICC is greater than 0.03 in the blinded model and greater than 0.025 in the unblinded model. With $m = 0.059$, the hybrid framework is better than the frequentist in terms of MSE when the ICC is greater than 0.05 in the blinded model and greater than 0.03 in the unblinded model. When a larger prior mean is selected ($m = 0.1$), the hybrid method becomes more powerful when the ICC is greater than 0.07 in the blinded model and greater than 0.05 in the unblinded model.

For completely non-informative priors, the MSE is slightly better in the unblinded model and slightly worse in the blinded model for hybrid designs when compared to the frequentist approach. Observed gains over the frequentist approach in terms of power are

a result of an increased positive bias in the hybrid approach.

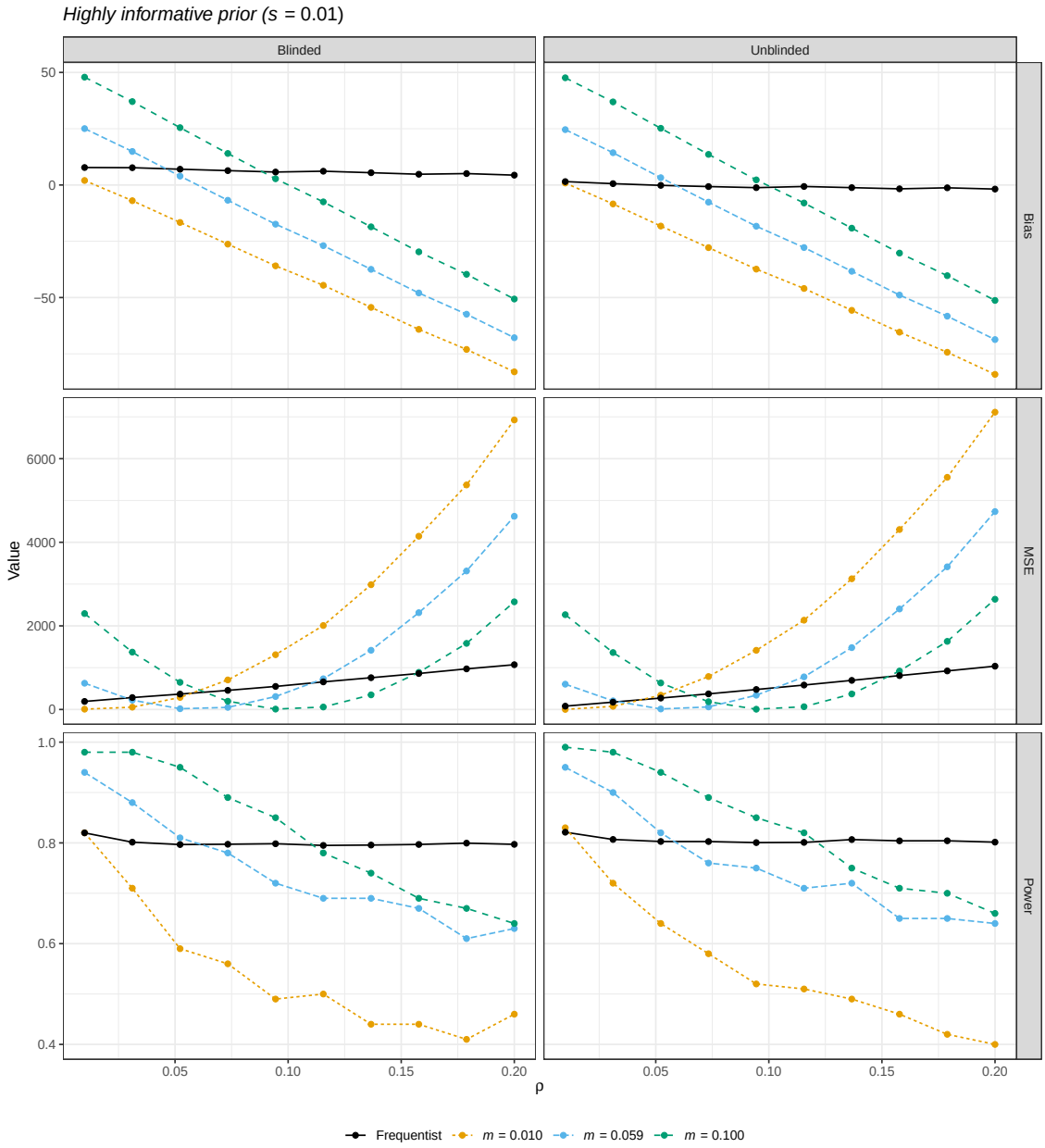

Figure 5.4: The bias, mean square error (MSE), and power of the frequentist and hybrid methods is shown as a function of the intra-cluster correlation (ρ). Results are faceted by the use of blinded vs. unblinded sample size re-estimation. For the hybrid approach, all combinations of $m = 0.01, 0.059, 0.1$ and $s = 0.01$ are considered.

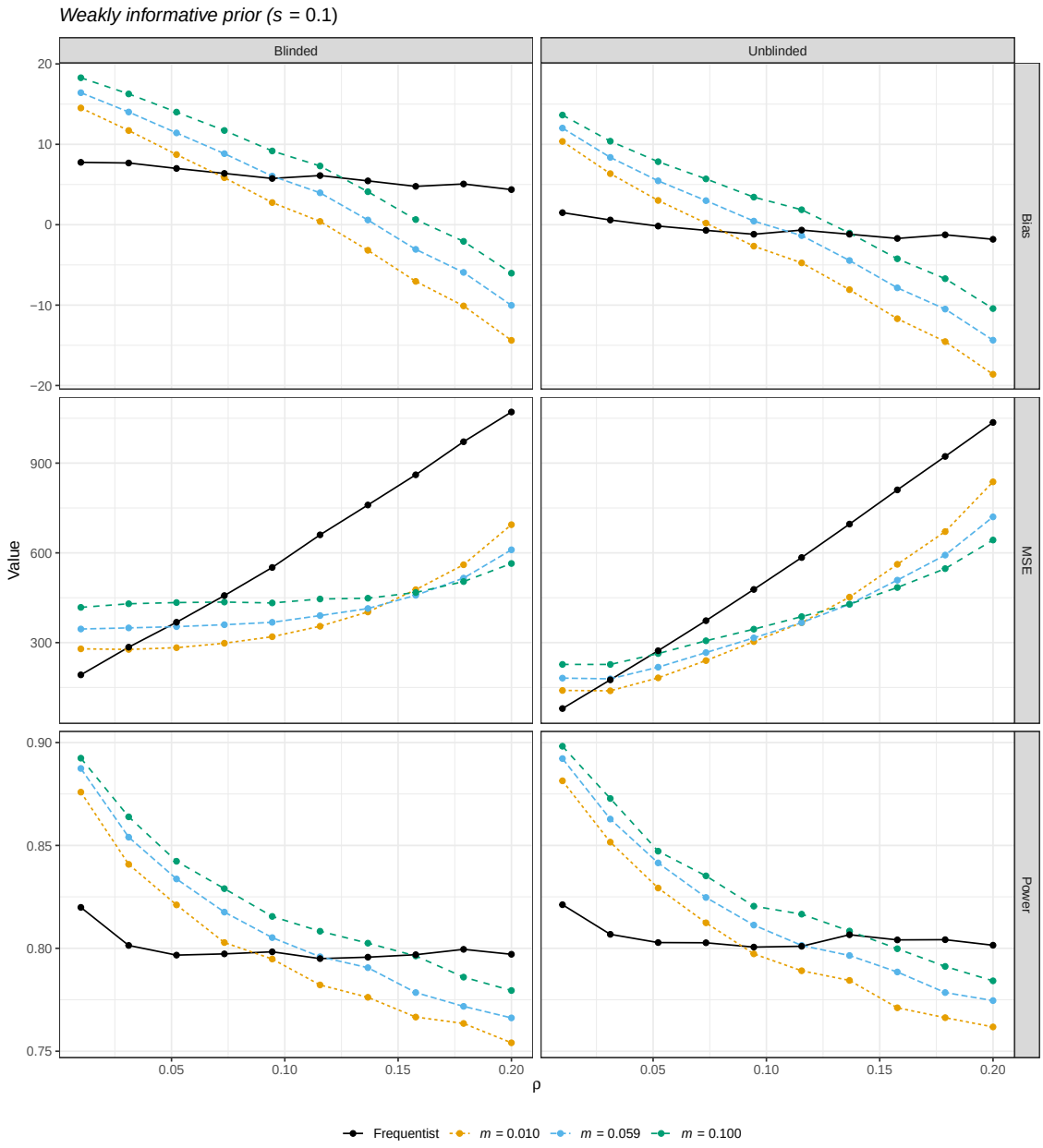

Figure 5.5: The bias, mean square error (MSE), and power of the frequentist and hybrid methods is shown as a function of the intra-cluster correlation (ρ). Results are faceted by the use of blinded vs. unblinded sample size re-estimation. For the hybrid approach all combinations of $m = 0.01, 0.059, 0.1$ and $s = 0.1$ are considered.

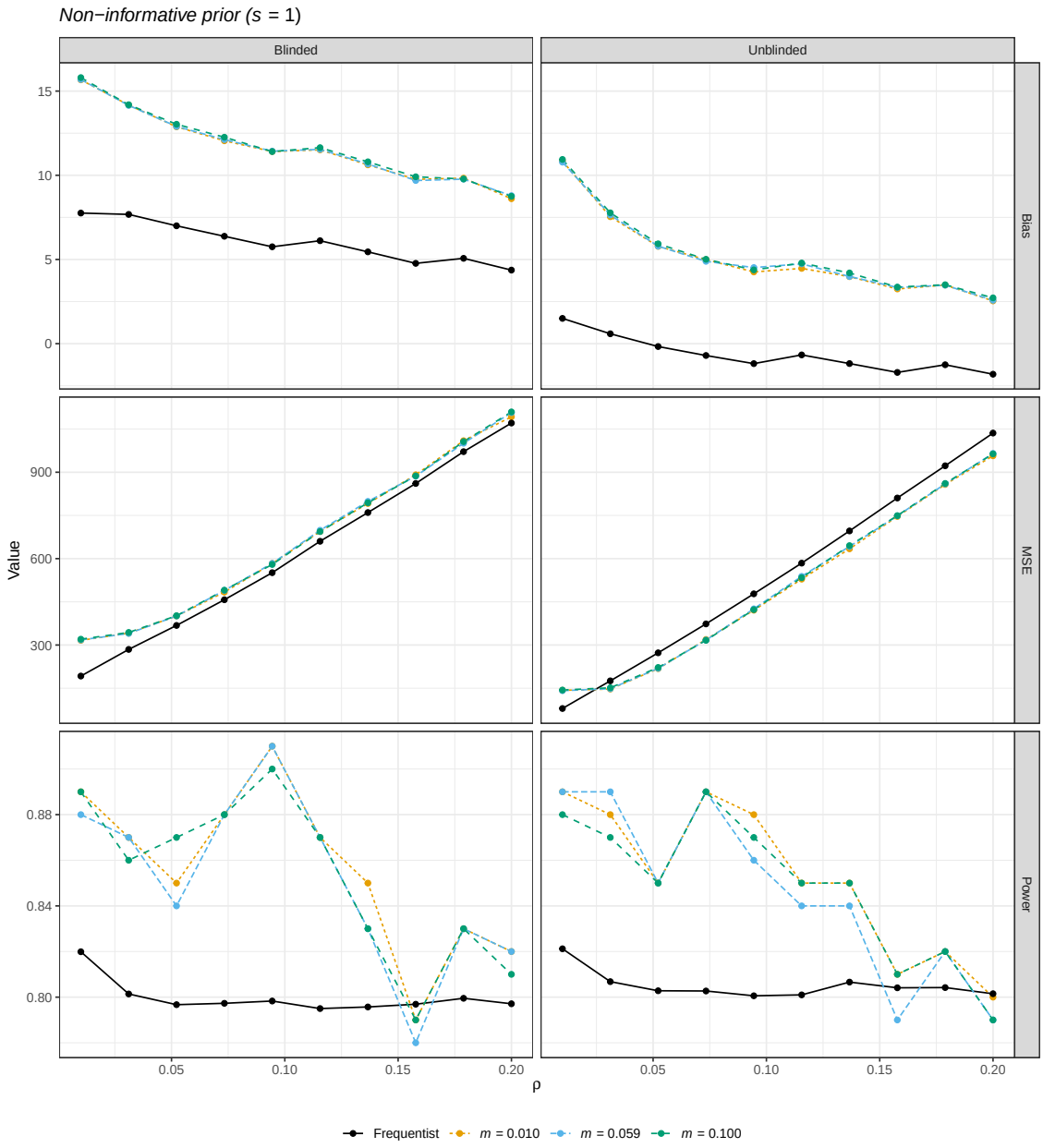

Figure 5.6: The bias, mean square error (MSE), and power of the frequentist and hybrid methods is shown as a function of the intra-cluster correlation (ρ). Results are faceted by the use of blinded vs. unblinded sample size re-estimation. For the hybrid approach, all combinations of $m = 0.01, 0.059, 0.1$ and $s = 1$ are considered.

5.4 Discussion

SSRE using a frequentist approach mostly abates the difficulties of obtaining precise estimates of the ICC during the trial design stage; yet, it has some practical issues. Notable among these issues is a large variation in the reestimated sample size, a consequence of uncertainty around the reestimated ICC (Hemming *et al.*, 2021). In this chapter, we have demonstrated how a hybrid approach to SSRE could address this known limitation within the frequentist framework whilst effectively controlling the type I error rate. We have also demonstrated when a hybrid approach may be useful in comparison to the frequentist.

Regarding the impact of prior on the final sample size, the hybrid framework is notably more efficient in terms of MSE and power when the prior is weakly informative, and substantially more efficient when a highly informative prior has a mean close to the true ICC. The finding that a weakly informative prior effectively facilitates the SSRE process is a highlight-worthy benefit of the hybrid technique, given that some information in the form of routinely gathered data or expert opinion may exist in practice for prior construction. Furthermore, a recent review of ICCs in health services research revealed that the median ICC for sample size calculation at the planning stage is typically 0.05 whereas the observed ICC for analysed primary outcome had an IQR of $(0.001, 0.060)$ (Offorha *et al.*, 2022). This evidence further indicates the utility of the proposed hybrid methods in practice since a weakly informative prior when $m = 0.01$, for example, offered a MSE below the frequentist design while retaining a higher power when $\rho \in [0.026, 0.9]$.

Results from this study are consistent with the widely held assertion that blinded SSRE is generally better because there is a small positive bias in the estimate of the variances and that translates into a slightly higher power (Kieser & Friede, 2003; Grayling *et al.*, 2018). However, higher power isn't necessarily optimal since there is often a cost in terms of the MSE. This relationship was mostly evident in non-informative priors where higher power was mostly observed but at the cost of a large positive bias and MSE. This indicates that the method is overpowering the study, which is not desirable. Nonetheless, there are some specific ICC ranges for which there is no trade-off between the MSE and power as shown in the example above. Thus, it is imperative to determine through simulation an acceptable trade-off or the optimality of the hybrid design in future trials.

Though some studies have defined SSRE in the context of updating cluster sizes (Hemming *et al.*, 2021), we have updated the number of clusters instead, as this generally has a higher impact on power. This is consistent with studies which sought increases in power through increasing the number of clusters rather than cluster size (Koepsell *et al.*, 1992; Thompson *et al.*, 1997; Van Breukelen & Candel, 2012). Hemming *et al.* (2017) further discusses diminishing returns in power and precision of a CRT as cluster size increases. We acknowledge that due to some logistical constraints and funding, it may be more difficult

to add more clusters than to increase the number of participants per cluster (Van Breukelen & Candel, 2012). In such scenarios, the methods proposed here could be extended accordingly.

A practical consideration for SSRE designs is the choice of sample size for the interim analysis. Studies have shown that estimates from pilot studies, which typically employ small sample sizes are frequently imprecise (Ip *et al.*, 2011; Eldridge *et al.*, 2016). To some researchers, 40 clusters seem inadequate to yield precise estimates of the ICC in the frequentist framework (Leyrat *et al.*, 2018). Therefore, when there is such uncertainty around the assumed ICC, the hybrid approach might be preferred over the frequentist if SSRE is to be performed. This is because even if the best guess estimate (m) based on an existing data is inaccurate, a weakly informative prior may still be advantageous since the performance does not depend heavily on the accuracy of the prior mean.

We note that this study has some limitations. First, we limited the choice of a prior distribution. While we believe that an extension of this approach to different priors and outcome data might yield similar results, future studies detailing the results and complexities of such models could be helpful. Another limitation was the consideration of only one CRT design (the parallel group). Hence, inferences from this study cannot be used to generalise to other CRT designs. For example. re-estimating the number of clusters in a stepped-wedge CRT might be more difficult if the original roll-out was planned in a particular way.

In conclusion, the hybrid approach is similar to the frequentist approach when using a completely uninformative prior, whereas a highly informative prior does better if the prior is correct (and is poor otherwise). utilising a weakly informative prior performs well, demonstrating robustness in terms of the MSE over a wide range of ICC values seen in practice. Owing to the range of ICC for which one can achieve a low MSE and higher power, a simulation study can be useful to assess when a hybrid approach may offer utility in terms of low MSE and higher power as well as help overcome known issues with the frequentist approach.

Chapter 6

Discussion and Conclusions

This chapter offers an overview of the thesis, highlighting the innovative methodologies introduced within it, and discusses the practical implications of their application. It also outlines the advantages and limitations of the proposed approaches and explores potential avenues for further advancement in the research area.

6.1 Motivation and overview of the thesis

Over the years, there have been prominent calls for global action, notably by the WHO, emphasising the importance of all nations being both producers and consumers of health research (WHO, 2013). Within this period, the number of countries partaking in pivotal trials filed to permit medication registration has almost doubled, yet the diversity of clinical trial populations has not increased significantly (Gross *et al.*, 2022). Despite some advancements (Jones *et al.*, 2007; Ijsselmuiden *et al.*, 2012), the majority of clinical research is still spearheaded by HICs, and many LMICs lack the capacity to independently conduct clinical research and implement findings into policy (Drain *et al.*, 2018). The lack of progress in health research capacity by LMICs may be attributed to a plethora of issues. Hence, an initial search was conducted to unearth the key barriers contributing to the low number of bespoke trials in LMICs. The majority of the studies cited funding or the high cost of running trials as the main barrier to conducting more trials in LMICs (Aboulghar, 2011; Schlaff, 2011; Franzen *et al.*, 2013*b*; Seruga *et al.*, 2014; Cardoso *et al.*, 2015). Subsequently, a relationship between sample size and the cost of trials was also uncovered in the above literature.

Given that ADs are recognised for their ability to generate clinical evidence regarding the most effective and cost-efficient interventions while maximising the use of limited resources, there is the need to empower LMICs to utilise such methodologies (McMichael *et al.*, 2005; Rosala-Hallas *et al.*, 2018). Nonetheless, there are barriers and complexities to

using ADs. Motivated by this background, the thesis sought to develop methods for overcoming the barriers to using ADs in LMICs. To address the sample size component of the overall cost of trials, the study in particular focused on the development of methods that provide robust and cost-effective sample sizes for both IRT and CRT designs. Although these methods were developed with LMICs in mind, their application is anticipated to extend to HICs as well.

In the subsequent sections, I summarise each segment of work presented in the thesis, emphasising the underlying motivation, employed methodologies, implications of the findings on trials in LMICs, as well as the general strengths and limitations. Subsequently, I conclude by making some remarks on non-methodological barriers and delve into potential future directions for methodological research in this area.

6.1.1 Chapter 2: Optimal drop-the-loser trials when an intermediate endpoint is used for interim selection

The first objective proposed an integration of seamless phase II/III and *drop-the-loser* designs into one framework. Within this design, two treatments were compared to a shared control in phase II and the least effective arm was dropped from the subsequent phase III stage. The primary aim of the design was to speed up the drug development process, by combining two trials traditionally conducted independently. To make good use of limited resources, an intermediate endpoint, envisioned to be cheaper and faster to use, was utilised at the first stage of the trial to inform adaptations before the definitive outcome was evaluated in the second stage. The methodology by Wason *et al.* (2017) was extended to accommodate the intermediate and definitive endpoints at the first and second stages respectively. Normal outcomes were assumed in both stages. Consequently, if a researcher intends to utilise a binary outcome for one of the endpoints, they must employ a normal approximation to align it with the established framework.

Considering that ADs rely on data available at the interim analysis to inform adaptations, an optimal timing for the interim analysis was proposed. The key finding was that conducting the interim analysis when 65% ($\theta = 0.65$) of the data had been collected was optimal for cases where the correlation between the endpoints is treated as unknown in the FWER control requirement. Conversely, the optimal timing of the interim analysis reduced from $\theta = 0.75$ when $\rho = 0.25$ to $\theta = 0.65$ when $\rho = 1$ in the case of treating the correlation as known for the FWER control requirement. Thus, the larger the assumed value of ρ, the earlier the optimal timing of the interim analysis. It was also established that a strong correlation between the intermediate and definitive outcome translated into a reduced required sample size. Furthermore, the results indicated that having precise knowledge of the assumed correlation was not really essential, as the variation in sample size from assuming no correlation to assuming a perfect correlation was minimal.

These findings have some implications for trials in LMICs, and by extension HICs. First, the impact of the limitation related to delayed outcomes in ADs is mitigated because the intermediate endpoint offers a quick measure to guide adaptations in the definitive endpoint. This feature of the proposed design, combined with the integration of two trial phases into a single trial, will accelerate the conduct of regulated trials within the region. The second benefit stems from its cost-effectiveness. While the seamless design results in a relatively small reduction in the required sample size, they significantly decrease the number of patients for whom the definitive endpoint needs to be measured. Thus, if the definitive endpoint involves an expensive piece of equipment, even the smallest reduction in sample size could lead to substantial cost savings, particularly for a large-scale trial conducted in an LMIC where resources are limited. Therefore, the reduced sample size in comparison to traditional non-ADs, and the use of a less expensive endpoint to identify the treatment arm for subsequent data collection make the design more cost-effective. Whether this is true, of course, depends on whether it is considered acceptable to collect the definitive endpoint only for those patients on the control and selected treatment arms.

6.1.2 Chapter 3: A review of approaches to specifying the intra-cluster correlation and other design parameters

To motivate the subsequent chapters' development of CRT methodologies, this chapter examined various approaches to specifying the ICC and other parameters influencing sample size. The key objectives included: i) unveiling the complexities related to the selection and justification of the assumed values for sample size estimation, ii) identifying the values commonly observed in the analysed primary outcomes, and iii) evaluating adherence to the CONSORT extension in terms of reporting. Trials reported in the HTA journal were used for this review based on the comprehensive nature of such reports and their lack of publication bias. Of the 54 articles that were identified from a search of PubMed, 34 of them met the eligibility criteria. Specifically, we (myself, James Wason, and Michael Grayling) were interested in trials that performed power calculations for which the parameters of interest could be extracted.

Contrary to expectations from the HTA journal, the ICC and other design parameters were poorly reported. Many HTA reports did not adhere to the CONSORT guideline of indicating the uncertainty around the assumed ICC, while others did not justify the assumed ICC or SD. Of those that reported the ICC, the study's findings highlighted a significant disparity between assumed and observed ICCs, with assumed values often surpassing observed values. The indication that many trials may have been overpowered raises concerns about potential research waste and ethical considerations, particularly concerning human subjects. While trialists generally favour an overpowered trial over an underpowered one, the extent of overestimation plays a critical role in this preference. However, there was no

standardised method identified in the review for inflating the required sample size to address the risk of being underpowered or to accommodate patient dropouts. Some studies allow for a 10% increase in the estimated sample size, while others adjust more or less. These informal approaches may inflate the sample size unnecessarily and could potentially expose patients to drugs whose full efficacy and safety profiles are not fully understood before the trial commences. These findings underscored the importance of using accurate parameter estimates for sample size calculations and laid the foundation for the methods developed in the subsequent chapters where uncertainty in the sample size parameters is formally captured using a prior distribution to estimate the required sample size.

6.1.3 Chapter 4: A hybrid approach to designing parallel-group and stepped-wedge cluster randomised trials

The review in Chapter 3 led to an evaluation of the impact of uncertainty in the ICC on the efficiency of sample size requirements for both parallel-group and stepped-wedge CRT designs. These two CRT designs have wide utility in practice, therefore methods developed around them could be useful in practice. To enable uncertainty at the design stage to be incorporated into the design specification, I described how sample size calculation can be performed for both types of CRT design in the 'hybrid' framework, which places priors on design parameters and controls the expected power in place of the conventional frequentist power. A comparison of the PG-CRT and SW-CRT designs was conducted by placing Beta or truncated normal priors on the ICC, and a Gamma prior on the standard deviation.

The findings suggested that, in instances where the mean ICC is greater than 0.1 and the uncertainty regarding the ICC captured by the SD is greater than 0.16, a SW-CRT design tends to be more efficient than a PG-CRT design. Even for a prior ICC distribution with a small mode, moderate prior densities on high ICC values can lead to a SW-CRT being more efficient because of the degree to which a SW-CRT is more efficient for high ICCs. Moreover, with careful specification of the priors, the designs in the hybrid framework can become more robust to, for example, an unexpectedly large value of the outcome variance. Therefore, when there is difficulty obtaining a reliable value for the ICC to assume at the design stage, the proposed methodology offers an appealing approach to sample size calculation.

In the context of disease epidemics prevalent in LMICs such as ebola, HIV, tuberculosis, and others, vaccine trials are crucial for developing effective remedies. In the case of vaccine trials in LMICs, where the burden of infectious diseases is high, the SW-CRT design emerges as a pragmatic choice. That is, the SW-CRT design's efficiency in sample size and its flexible, gradual implementation make it a potentially advantageous choice for trials in LMICs. It also aligns with the practical constraints and ethical considerations often encountered in these settings, contributing to the feasibility and success of trials

aimed at improving health outcomes. Thus the aforementioned findings could be useful ones. Nonetheless, we note again that its design and analysis are complex relative to the PG-CRT. Therefore, although the required sample size is an important consideration in selecting a trial design, the right balance must be struck based on the objectives and other considerations of the trial before selecting a CRT design.

6.1.4 Chapter 5: A hybrid approach to sample size reestimation in cluster randomised trials

An inherent limitation of the approach in Chapter 4 is its reliance on the choice of prior at the design stage. The design provides no opportunity to evaluate whether the selected prior correctly reflects the data from the intended trial until the trial reaches completion. To mitigate against this limitation, an AD was next introduced wherein pre-trial knowledge about the ICC is captured by placing a prior on it and updated at an interim analysis using study data. In this chapter, only the PG-CRT design was considered since it is more responsive to variations in the ICC compared to the SW-CRT design.

In the proposed methodology, I began by describing how SSRE can be performed in both the frequentist and hybrid frameworks. As increasing the number of clusters typically has a bigger impact on power than increasing the cluster size does, I focused on interim updating of the required number of clusters throughout. The primary outcome of the motivating example was continuous and normally distributed with known variance σ^2. The analysis utilised a linear mixed model. A truncated normal distribution was selected for the prior placed on the ICC. Given the prior, constructing the posterior for the ICC required knowledge of the likelihood; Fisher's approximation for the variance component of the ICC was used as the likelihood, with MCMC methods then employed to sample from the posterior. Here, the SSRE design can be conceptualised as an attempt to determine the sample size required under the assumption that the true parameters are known. The metrics used to evaluate the performance of the SSRE 'estimators' were thus their bias and MSE in the final re-estimated required sample size. A well-performing SSRE method was expected to exhibit a bias close to 0 and a low MSE.

On average, both the hybrid and frequentist approaches mitigated against the implications of misspecifying the ICC during the trial's design stage. Furthermore, both frameworks yielded SSRE designs with approximate control of the type I error rate to the desired level. The study clearly illustrated how the hybrid approach can minimise the significant variability in the reestimated sample size observed within the frequentist framework, contingent on the informativeness of the prior. This implies that the hybrid approach could provide benefits for CRTs employing SSRE, especially when there is available data or expert opinion to guide the choice of ICC prior. It was further shown in the results that the greatest utility of the hybrid approach is likely observed in scenarios with

low-quality evidence available for informing the choice of prior, as SSRE is less likely to be utilised when substantial data are available.

These findings also have notable implications for trials in LMICs. Most prominently, given that LMICs face challenges in routinely collecting high-quality data due to limited infrastructure and resources, an approach that proves beneficial in scenarios of low-quality data becomes particularly valuable.

Similarly, this study also had some limitations. First, the consideration of only the PG-CRT design restricts the generalisability of the findings to other CRT designs; different CRT designs may exhibit distinct characteristics, rendering the results not universally applicable. For example, the SW-CRT design is robust against variations in the ICC. Consequently, changes in the prior following the interim analysis may not have a significant impact on a SW-CRT design when compared to a PG-CRT. Secondly, the study focused on normal outcomes, which might not capture the diversity of outcome data types encountered in various research settings. Future studies could enhance their scope by exploring different types of outcome data with varied prior distributions. This would contribute to a more comprehensive understanding of the proposed methods across a broader range of scenarios.

6.2 Non-methodological recommendations

In addition to the methodological issues considered in this thesis, I recognise that other non-methodological issues need to be addressed to obtain a holistic solution to the barriers to conducting trials in LMICs. In what follows, I comment on a few of these issues. They include:

- **Capacity building:** LMICs need to provide training and education for researchers, healthcare professionals, and ethics committees to enhance their understanding of AD methodologies, ethics, and regulatory requirements. It often happens that trials conducted within LMICs lack locally-led professionals such as PIs, data managers, lead statisticians, etc. (Pai, 2011; Zegers-Hochschild, 2011; Franzen *et al.*, 2013*b*,*a*).

- **Building of research infrastructure:** Despite the WHO statement in 2005 emphasising the international priority of establishing African-owned research centres capable of conducting their own clinical trials (Matsoso *et al.*, 2005), progress has been limited due to a lack of infrastructure (Cardoso *et al.*, 2015; Zegers-Hochschild, 2011). Therefore, LMICs should invest in building and strengthening research infrastructure, including research centres, laboratories, and data management systems. This can improve the capacity to conduct high-quality research.

- **Ethical oversight and regulatory framework:** Another major setback to conducting trials in LMICs is the delay and complex ethical approval by regulatory

bodies (Seruga *et al.*, 2014; Pai, 2011; L. Gómez *et al.*, 2015; Sulthan, 2015; Aboulghar, 2011). Hence, it is vital to strengthen ethical review processes to ensure they are swift and conform to international standards while addressing cultural sensitivities and community engagement. In addition, clear and transparent regulatory frameworks for research that ensure patient safety and data integrity while facilitating efficient research processes should be established.

- **Disease-specific research:** Given that LMICs do not have the needed infrastructural capacity for all trials, it is important that they focus on research areas that are relevant to their specific health challenges, such as diseases disproportionately affecting these regions.

Additionally, researchers within LMICs should advocate for government support and funding for research, increase public awareness about clinical research, and foster international collaborations between LMIC institutions, HICs, and research organisations for technical expertise and knowledge sharing. By implementing these strategies and working collaboratively, LMICs can create an environment conducive to conducting impactful and ethically sound clinical research that addresses their unique healthcare challenges.

6.3 Areas for future work

Measurement error is a challenge that could negatively affect the validity and reliability of results. Hence, the ability to minimise its impact contributes to the overall success of the trial. It is anticipated that measurement error may be more common in an LMIC setting since cheaper data collection and measurement methods may be employed due to limited resources. A possible approach to handling this challenge is to develop an internal pilot design where an intermediate and definitive endpoint is measured at the first stage. During the interim analysis, the measurement error between them is quantified to enable the trialists to choose the sample size for the remainder of the trial. Since we compute measurement error at the interim analysis, we can subsequently not measure the more expensive definitive outcome in the second stage, having controlled the measurement error quantified at the interim analysis. Consequently, the intermediate endpoint is measured during the second stage for the remainder of the trial. This technique could build upon methodology for a single outcome trial (Wason *et al.*, 2014) where the difference between a standard and a biomarker-direct treatment group is quantified, and for a multi-outcome trial (Law *et al.*, 2020) where endpoints are selected at an interim analysis. To the best of my knowledge, there is no study on the impact of measurement error in adaptive trials or how its impact can be mitigated. This research gap leaves numerous avenues for research in this area. Therefore, future work could explore measurement errors within the drop-

the-loser design. Specifically, we could investigate the case where the evaluation of the Type I error rate would depend on the result of quantifying the measurement error and then subsequently testing the intermediate outcome at the final analysis.

Also, an extension to the drop-the-loser design in Chapter 2 could focus on exploring joint distributions for different types of outcome data, such as binary and time-to-event. The importance of these two endpoints in clinical trials lies in their ability to provide clear interpretable outcomes and dynamic time-related information respectively. Together, they contribute to a thorough understanding of intervention effects and patient outcomes. Although a normal approximation could work in this instance, further studies could employ conditional distributions between the two endpoints, such as methods proposed by Stallard (2010), for a higher level of precision. The MIDFUT trial, as documented by Brown *et al.* (2020), could serve as a motivating example for such methods. This was a seamless phase II/III, open, parallel-group, MAMS design of patients with hard-to-heal diabetic foot ulcers. The trial utilised a binary endpoint in phase II (at least 50% reduction in index ulcer area at 4 weeks post-randomisation or otherwise) and a time-to-event endpoint in phase III (time to healing of the index ulcer). Additionally, it will be important to assess whether the methodology regarding the FWER control in this work remains applicable in this particular context.

A notable extension to the most basic PG-CRT design emerges in settings where trialists are interested in assessing the effect of an intervention relative to some baseline measurement; this design is sometimes referred to as a CRT with before and after observations (CRT-BA) (Eldridge & Kerry, 2012; Hemming & Taljaard, 2016). This design, in some cases, can be highly efficient and increasingly becoming widespread in their use. For brevity, however, the thesis did not consider this design. In this regard, future studies could incorporate the hybrid approach into this CRT design and assess its properties relative to the more classical ones which were considered in this study. Considering the similarities between the cluster crossover trial and the SW-CRT designs, the comparison in Chapter 4 could be extended to the cluster crossover trial and the CRT-BA designs. This comparison would aim to evaluate how uncertainty in the ICC influences the determination of efficiency in terms of sample size between the two designs.

Lastly, the use of Bayesian assurance in sample size determination has been gaining popularity recently. Therefore, studies that compare the hybrid and fully Bayesian assurance can add to the existing literature and provide options for clinical researchers.

6.4 Conclusions

The importance of locally-led clinical research, particularly in populations within LMICs is widely acknowledged. This is deemed crucial for accurately identifying problems, propos-

ing culturally appropriate and cost-effective interventions, investigating implementation strategies, and overcoming obstacles to the adoption of recommended approaches. To contribute to the international priority set by the WHO for bespoke trials, the specific focus was on developing methods that address the major challenge of funding in trials, aiming to optimise limited resources. The methods proposed in this thesis are robust, cost-effective, and expedite trials. The hope is that these methods will therefore find utility within LMICs, helping bridge the research gap between LMICs and HICs.

Appendix A

A.1 Software

Code to reproduce the results in each chapter is available from `https://github.com/sks2023/article_codes`.

Bibliography

Aboulghar, M. 2011 Barriers to conducting clinical research in reproductive medicine: Egypt. *Fertility and Sterility* **96** (4), 805–806.

Adab, P., Barrett, T., Bhopal, R., Cade, J. E., Canaway, A., Cheng, K. K., Clarke, J., Daley, A., Deeks, J., Duda, J., Ekelund, U., Frew, E., Gill, P., Griffin, T., Hemming, K., Hurley, K., Lancashire, E. R., Martin, J., McGee, E., Pallan, M. J., Parry, J. & Passmore, S. 2018 The west midlands active lifestyle and healthy eating in school children (Waves) study: A cluster randomised controlled trial testing the clinical effectiveness and cost-effectiveness of a multifaceted obesity prevention intervention programme targeted at. *Health Technology Assessment* **22** (8), 1–644.

Adler, C. H., Singer, C., O'Brien, C., Hauser, R. A., Lew, M. F., Marek, K. L., Dorflinger, E., Pedder, S. & Deptula, D. 1998 Randomized, Placebo-Controlled Study of Tolcapone in Patients With Fluctuating Parkinson Disease Treated With Levodopa-Carbidopa. *ARCH NEUROL* **55**, 1089–1095.

Alemayehu, C., Mitchell, G. & Nikles, J. 2018 Barriers for conducting clinical trials in developing countries- a systematic review. *International Journal for Equity in Health* **17** (1), 1–11.

Alonzo, T. A. & Pepe, M. S. 2007 *Development and evaluation of classifiers.*, , vol. 404. Totowa, New Jersey: Humana Press.

Altman, D. G. 1980 Statistics and ethics in medical research. *British Medical Journal* **281** (6252), 1422–1423.

Alwan, A. 2011 *Global status report on noncommunicable diseases 2010*. World Health Organization.

American Cancer Society 2020 Types and Phases of Clinical Trials.

ASPE 2014 Examination of Clinical Trial Costs and Barriers for Drug Development. *Tech. Rep.*. U.S. Department of Health and Human Services, Washington, DC.

Bacchetti, P., Mcculloch, C. E., Segal, M. R., Hanley, J. A., Shapiro, S. H., Müller, P., Rosner, G. L., Simon, R., Bacchetti, P., Mcculloch, C. E. & Segal, M. R. 2008 Linked references are available on JSTOR for this article : Simple , Defensible Sample Sizes Based on Cost Efficiency. *Biometrics* **64** (2), 577–594.

Baio, G., Copas, A., Ambler, G., Hargreaves, J., Beard, E. & Omar, R. Z. 2015 Sample size calculation for a stepped wedge trial. *Trials* **16** (1), 1–15.

Baker, S. G. & Kramer, B. S. 2003 A perfect correlate does not a surrogate make. *BMC Medical Research Methodology* **3**, 1–5.

Barreto, M. L. 2009 Health research in developing countries. *BMJ* **339**, b4846.

Bauer, P. & Kieser, M. 1999 Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* **18** (14), 1833–1848.

Bauer, P. & Kohne, K. 1994 Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics* **50** (4), 1029–1041.

Benda, N., Branson, M., Maurer, W. & Friede, T. 2010 Aspects of Modernizing Drug Development Using Clinical Scenario Planning and Evaluation. *Therapeutic Innovation & Regulatory Science* **44** (3), 299–315.

Benjamin, P., Zeestraten, E., Lambert, C., Chis Ster, I., Williams, O. A., Lawrence, A. J., Patel, B., Mackinnon, A. D., Barrick, T. R. & Markus, H. S. 2016 Progression of MRI markers in cerebral small vessel disease: Sample size considerations for clinical trials. *Journal of Cerebral Blood Flow and Metabolism* **36** (1), 228–240.

Bentley, C., Cressman, S., van der Hoek, K., Arts, K., Dancey, J. & Peacock, S. 2019 Conducting clinical trials—costs, impacts, and the value of clinical trials networks: A scoping review. *Clinical Trials* **16** (2), 183–193.

Besarab, A., Bolton, W. K., Browne, J. K., Egrie, J. C., Nissenson, A. R., Okamoto, D. M., Schwab, S. J. & Goodkin, D. A. 1998 The effects of normal as compared with low hematocrit values in patients with cardiac disease who are receiving hemodialysis and epoetin. *The New England journal of medicine* **339** (9), 584–590.

Bhardwaj, S. S., Camacho, F., Derrow, A., Fleischer, A. B. & Feldman, S. R. 2014 Statistical Significance and Clinical Relevance. *Arch Dermatol* **140**, 1520–1523.

Bhatt, A. 2010 Evolution of clinical research: a history before and beyond James Lind. *Perspectives in Clinical Research.* **1** (1), 6–10.

Binik, A. 2019 Delaying and withholding interventions: Ethics and the stepped wedge trial. *Journal of Medical Ethics* **45** (10), 662–667.

Bonell, C., Fletcher, A., Fitzgerald-Yau, N., Hale, D., Allen, E., Elbourne, D., Jones, R., Bond, L., Wiggins, M., Miners, A., Legood, R., Scott, S., Christie, D. & Viner, R. 2015 Initiating change locally in bullying and aggression through the school environment (INCLUSIVE): A pilot randomised controlled trial. *Health Technology Assessment* **19** (53).

Bothwell, L. E. & Kesselheim, A. S. 2017 The Real-World Ethics of Adaptive-Design Clinical Trials. *The Hastings Center report* **47** (6), 27–37.

Brakenhoff, T. B., Roes, K. C. & Nikolakopoulos, S. 2019 Bayesian sample size re-estimation using power priors. *Statistical Methods in Medical Research* **28** (6), 1664–1675.

Bratton, D. J., Parmar, M. K. B., Phillips, P. P. J. & Choodari-Oskooei, B. 2016 Type I error rates of multi-arm multi-stage clinical trials: strong control and impact of intermediate outcomes. *Trials* **17** (1), 309.

Bretz, F., Schmidli, H., König, F., Racine, A. & Maurer, W. 2006 Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal* **48** (4), 623–634.

Brown, S., Nixon, J., Ransom, M., Gilberts, R., Dewhirst, N., McGinnis, E., Longo, R., Game, F., Bojke, C., Chadwick, P., Chandrasekar, A., Chetter, I., Collier, H., Fernandez, C., Homer-Vanniasinkam, S., Jude, E., Leigh, R., Lomas, R., Vowden, P., Wason, J., Sharples, L. & Russell, D. 2020 Multiple Interventions for Diabetic Foot Ulcer Treatment Trial (MIDFUT): Study protocol for a randomised controlled trial. *BMJ Open* **10** (4).

Burzykowski, T., Molenberghs, G. & Buyse, M. 2005 *The Evaluation of Surrogate Endpoints*. New York: Springer.

Butler, C. C., Simpson, S. A., Hood, K., Cohen, D., Pickles, T., Spanou, C., McCambridge, J., Moore, L., Randell, E., Alam, M. F., Kinnersley, P., Edwards, A., Smith, C. & Rollnick, S. 2013 Training practitioners to deliver opportunistic multiple behaviour change counselling in primary care: A Cluster randomised trial. *BMJ (Online)* **346** (7901), 1–25.

Buyse, M. & Molenberghs, G. 1998 Criteria for the Validation of Surrogate Endpoints in Randomized Experiments **54** (3), 1014–1029.

Campbell, J. L., Fletcher, E., Britten, N., Green, C., Holt, T., Lattimer, V., Richards, D. A., Richards, S. H., Salisbury, C., Taylor, R. S., Calitri, R., Bowyer, V., Chaplin, K., Kandiyali, R., Murdoch, J., Price, L., Roscoe, J., Varley, A. & Warren, F. C. 2015 The clinical effectiveness and cost-effectiveness of telephone triage for managing same-day consultation requests in general practice: A cluster randomised controlled trial comparing general practitioner-led and nurse-led management systems with usual car. *Health Technology Assessment* **19** (13), 1–212.

Campbell, M. J., Donner, A. & Klar, N. 2007 Developments in cluster randomized trials and Statistics in Medicine. *STATISTICS IN MEDICINE* **26**, 2–19.

Campbell, M. J. & Walters, S. J. 2014 *How to design, analyse and report Cluster Randomised Trials in Medicine and Health Related Research*. Chichester: John Wiley & Sons Ltd.

Campbell, M. K., Grimshaw, J. M. & Elbourne, D. R. 2004 Intracluster correlation coefficients in cluster randomized trials: Empirical insights into how should they be reported. *BMC Medical Research Methodology* **4**, 1–5.

Campbell, M. K., Piaggio, G., Elbourne, D. R. & Altman, D. G. 2012 Consort 2010 statement: Extension to cluster randomised trials. *BMJ (Online)* **345** (7881), 1–21.

Cancer Research UK 2022 How long a new drug takes to go through clinical trials.

Cardoso, A. L., Breugelmans, G., Manville, C., Chataway, J., Cochrane, G., Snodgrass, J., Chataway, M. & Murali, N. 2015 Africa mapping: current state of health research on poverty-related and neglected infectious diseases in Sub-Saharan Africa. *Tech. Rep.*. The European & Developing Countries Clinical Trials Partnership.

Carlos, R. C. & Goeree, R. 2009 Introduction: Health Technology Assessment in Diagnostic Imaging. *Journal of the American College of Radiology* **6** (5), 297–298.

Chakraborty, H., Moore, J., Carlo, W. A., Hartwell, T. D. & Wright, L. L. 2009 A simulation based technique to estimate intracluster correlation for a binary variable. *Contemporary Clinical Trials* **30** (1), 71–80.

Charles, P., Giraudeau, B., Dechartres, A., Baron, G. & Ravaud, P. 2009 Reporting of sample size calculation in randomised controlled trials: Review. *BMJ (Online)* **338** (7705), 1256.

Chataway, J., Nicholas, R., Todd, S., Miller, D. H., Parsons, N., Valdés-Márquez, E., Stallard, N. & Friede, T. 2011 A novel adaptive design strategy increases the efficiency of clinical trials in secondary progressive multiple sclerosis. *Multiple Sclerosis Journal* **17** (1), 81–88.

Chen, D. G. D. & Ho, S. 2017 From statistical power to statistical assurance: It's time for a paradigm change in clinical trial design. *Communications in Statistics: Simulation and Computation* **46** (10), 7957–7971.

Chow, S. C. & Chang, M. 2008 Adaptive design methods in clinical trials - A review. *Orphanet Journal of Rare Diseases* **3** (1), 1–13.

Chow, S. C. & Corey, R. 2011 Benefits, challenges and obstacles of adaptive clinical trial designs. *Orphanet Journal of Rare Diseases* **6** (1), 1–10.

Chow, S.-C. & Liu, J.-P. 2014 *Design and Analysis of Clinical Trials. Concepts and Methodologies*, 3rd edn. New Jersey: Wiley.

Chuang-Stein, C. 2006 Sample size and the probability of a successful trial. *Pharmaceutical Statistics* **5** (4), 305–309.

Ciani, O., Grigore, B., Blommestein, H., de Groot, S., Möllenkamp, M., Rabbe, S., Daubner-Bendes, R. & Taylor, R. S. 2021 Validity of Surrogate Endpoints and Their Impact on Coverage Recommendations: A Retrospective Analysis across International Health Technology Assessment Agencies. *Medical Decision Making* **41** (4), 439–452.

Ciarleglio, M. M., Arendt, C. D. & Peduzzi, P. N. 2016 Selection of the effect size for sample size determination for a continuous response in a superiority clinical trial using a hybrid classical and Bayesian procedure. *Clinical Trials* **13** (3), 275–285.

Cook, J. A., Julious, S. A., Sones, W., Hampson, L. V., Hewitt, C., Berlin, J. A., Ashby, D., Emsley, R., Fergusson, D. A., Walters, S. J., Wilson, E. C. F., Maclennan, G., Stallard, N., Rothwell, J. C., Bland, M., Brown, L., Ramsay, C. R., Cook, A., Armstrong, D., Altman, D. & Vale, L. D. 2018 DELTA 2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a r andomised controlled trial considered to be the best method to. *BMJ* .

Copas, J. B. & Malley, P. F. 2008 A robust P -value for treatment effect in meta-analysis with publication bias. *STATISTICS IN MEDICINE* **27**, 4267–4278.

Cornfield, J. 1978 Symposium on CHD prevention trials: Design issues in testing life style intervention: Randomization by group: A formal analysis. *American Journal of Epidemiology* **108** (2), 100–102.

Coronado, G. D., Vollmer, W. M., Petrik, A., Taplin, S. H., Burdick, T. E., Meenan, R. T. & Green, B. B. 2014 Strategies and Opportunities to STOP Colon

Cancer in Priority Populations: Design of a cluster-randomized pragmatic trial. *Contemporary Clinical Trials* **38** (2), 344–349.

Coventry, P., Lovell, K., Dickens, C., Bower, P., Chew-Graham, C., McElvenny, D., Hann, M., Cherrington, A., Garrett, C., Gibbons, C. J., Baguley, C., Roughley, K., Adeyemi, I., Reeves, D., Waheed, W. & Gask, L. 2015 Integrated primary care for patients with mental and physical multimorbidity: Cluster randomised controlled trial of collaborative care for patients with depression comorbid with diabetes or cardiovascular disease. *BMJ (Online)* **350**, 1–11.

Crowley, J. & Hoering, A. 2012 *Handbook of statistics in clinical oncology*, 3rd edn. CRC Press.

Cundill, B., Mbakilwa, H., Chandler, C. I., Mtove, G., Mtei, F., Willetts, A., Foster, E., Muro, F., Mwinyishehe, R., Mandike, R., Olomi, R., Whitty, C. J. & Reyburn, H. 2015 Prescriber and patient-oriented behavioural interventions to improve use of malaria rapid diagnostic tests in Tanzania: Facility-based cluster randomised trial. *BMC Medicine* **13** (1), 1–16.

Dafni, U. G. & Tsiatis, A. A. 1998 Evaluating Surrogate Markers of Clinical Outcome When Measured with Error. *Biometrics* **54** (4), 1445–1462.

Darrow, J. J., Avorn, J. & Kesselheim, A. S. 2020 FDA Approval and Regulation of Pharmaceuticals, 1983-2018. *JAMA - Journal of the American Medical Association* **323** (2), 164–176.

DeMets, D. L., Psaty, B. M. & Fleming, T. R. 2020 When Can Intermediate Outcomes Be Used as Surrogate Outcomes? *JAMA* **323** (12), 1184 – 1185.

Di Luca, D. G., Luo, S., Liu, H., Cohn, M., Davis, T. L., Ramirez-Zamora, A., Rafferty, M., Dahodwala, N., Naito, A., Neault, M., Beck, J. & Marras, C. 2023 Racial and Ethnic Differences in Health-Related Quality of Life for Individuals With Parkinson Disease Across Centers of Excellence. *Neurology* **100** (21), E2170–E2181.

Dimairo, M., Boote, J., Julious, S. A., Nicholl, J. P. & Todd, S. 2015 Missing steps in a staircase: A qualitative study of the perspectives of key stakeholders on the use of adaptive designs in confirmatory trials. *Trials* **16** (1), 1–16.

DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. 2016 Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics* **47**, 20–33.

Donner, A. & Klar, N. 2000 *Design and analysis of cluster randomization trials in health research*. New York: Oxford University Press Inc.

Donner, A. & Wells, G. 1986 A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics* pp. 401–412.

Drain, P. K., Parker, R. A., Robine, M. & Holmes, K. K. 2018 Global migration of clinical research during the era of trial registration. *PLoS ONE* **13** (2), 1–13.

Dunnett, C. W. 1955 A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association* **50** (272), 1096–1121.

Echt, D. S., Liebson, P. R., Mitchell, L. B., Peters, R. W., Obias-Manno, D., Barker, A. H., Arensberg, D., Baker, A., Friedman, L. & Greene, H. L. 1991 Mortality and morbidity in patients receiving encainide, flecainide, or placebo: the Cardiac Arrhythmia Suppression Trial. *New England journal of medicine* **324** (12), 781–788.

Edwards, S. J., Braunholtz, D. A., Lilford, R. J. & Stevens, A. J. 1999 Ethical issues in the design and conduct of cluster randomised controlled trials. *British Medical Journal* **318** (7195), 1407–1409.

Eichner, F. A., Groenwold, R. H., Grobbee, D. E. & Oude Rengerink, K. 2019 Systematic review showed that stepped-wedge cluster randomized trials often did not reach their planned sample size. *Journal of Clinical Epidemiology* **107**, 89–100.

Eldridge, S. & Kerry, S. 2012 *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Wiley.

Eldridge, S. m., Ashby, D., Feder, G. s., Rudnicka, A. r. & Ukoumunne, O. c. 2004 Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. *Clinical Trials* **1** (1), 80–90.

Eldridge, S. M., Costelloe, C. E., Kahan, B. C., Lancaster, G. A. & Kerry, S. M. 2016 How big should the pilot study for my cluster randomised trial be? *Statistical Methods in Medical Research* **25** (3), 1039–1056.

Eldridge, S. M., Ukoumunne, O. C. & Carlin, J. B. 2009 The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions. *International Statistical Review* **77** (3), 378–394.

Elley, C. R., Chondros, P. & Kerse, N. M. 2004 Randomised trials – cluster versus individual randomisation. *Australian Family Physician* **33** (9), 759–763.

Evans, I., Thornton, H., Chalmers, I. & Glasziou, P. 2011 Testing Treatments: Better Research for Better Healthcare.

Faber, J. & Fonseca, L. M. 2014 How sample size influences research outcomes. *Dental Press Journal of Orthodontics* **19** (4), 27–29.

FDA 2010 Draft Guidance for Industry - Adaptive Design Clinical Trials for Drugs and Biologics. *Tech. Rep.*. The United State Food and Drug Administration, Rockville, Maryland.

FDA 2018 Surrogate endpoint resources for drug and biologic development.

FDA 2019 Placebos and Blinding in Randomized Controlled Cancer Clinical Trials for Drug and Biological Products Guidance for Industry. *Tech. Rep.* August. FDA.

FDA-NIH Biomarker Working Group 2021 BEST (Biomarkers , EndpointS , and other Tools) Resource. *Tech. Rep.*. Food and Drug Administration and National Institutes of Health (US).

Fisher, R. A. 1970 Statistical Methods for Research Workers. *Darien, CT* .

Fleming, T. R. & Powers, J. H. 2012 Biomarkers and surrogate endpoints in clinical trials. *Statistics in Medicine* **31** (25), 2973–2984.

Fogel, D. B. 2018 Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary Clinical Trials Communications* **11**, 156–164.

Food and Drug Administration 2017 Draft guideline on multiple endpoints in clinical trials. *Tech. Rep.* January. Food and Drug Administration.

Frangakis, C. E. & Rubin, D. B. 2002 Principal stratification in causal inference. *Biometrics* **58** (1), 21–29.

Franzen, S. R., Chandler, C., Enquselassie, F., Siribaddana, S., Atashili, J., Angus, B. & Lang, T. 2013*a* Understanding the investigators: A qualitative study investigating the barriers and enablers to the implementation of local investigator-initiated clinical trials in Ethiopia. *BMJ Open* **3** (11), 1–10.

Franzen, S. R., Chandler, C., Siribaddana, S., Atashili, J., Angus, B. & Lang, T. 2017 Strategies for developing sustainable health research capacity in low and middle-income countries: A prospective, qualitative study investigating the barriers and enablers to locally led clinical trial conduct in Ethiopia, Cameroon and Sri Lanka. *BMJ Open* **7** (10), 1–15.

Franzen, S. R. P., Chandler, C., Atashili, J., Angus, B. & Lang, T. 2013*b* Barriers and enablers of locally led clinical trials in Ethiopia and Cameroon: a prospective, qualitative study. *The Lancet* **382**, 14.

Friede, T. & Kieser, M. 2013 Blinded sample size re-estimation in superiority and non-inferiority trials: Bias versus variance in variance estimation. *Pharmaceutical Statistics* **12** (3), 141–146.

Friede, T., Nicholas, R., Stallard, N., Todd, S., Parsons, N., Valdés-Márquez, E. & Chataway, J. 2010 Refinement of the Clinical Scenario Evaluation Framework for Assessment of Competing Development Strategies with an Application to Multiple Sclerosis. *Therapeutic Innovation & Regulatory Science* **44** (6), 713–718.

Friede, T., Parsons, N., Stallard, N., Todd, S., Valdes Marquez, E., Chataway, J. & Nicholas, R. 2011 Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine* **30** (13), 1528–1540.

Friede, T., Stallard, N. & Parsons, N. 2020 Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R. *Biometrical Journal* **62** (5), 1264–1283.

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M. & Granger, C. B. 2015 *Fundamentals of clinical trials*. Springer.

Froggatt, K., Best, A., Bunn, F., Burnside, G., Coast, J., Dunleavy, L., Goodman, C., Hardwick, B., Jackson, C., Kinley, J., Lund, A. D., Lynch, J., Mitchell, P., Myring, G., Patel, S., Algorta, G. P., Preston, N., Scott, D., Silvera, K. & Walshe, C. 2020 A group intervention to improve quality of life for people with advanced dementia living in care homes: The namaste feasibility cluster RCT. *Health Technology Assessment* **24** (6), vii–139.

Gallo, P., Anderson, K., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M. & Pinheiro, J. 2010 Viewpoints on the FDA draft adaptive designs guidance from the PhRMA working group. *Journal of Biopharmaceutical Statistics* **20** (6), 1115–1124.

Gary, B. 2022 Bayesian Methods for the Design and Analysis of Cluster Randomised Controlled Trials. PhD thesis, University of Plymouth.

Gates, S., Lall, R., Quinn, T., Deakin, C. D., Cooke, M. W., Horton, J., Lamb, S. E., Slowther, A. M., Woollard, M., Carson, A., Smyth, M., Wilson, K., Parcell, G., Rosser, A., Whitfield, R., Williams, A., Jones, R., Pocock, H., Brock, N., Black, J. J., Wright, J., Han, K., Shaw, G., Blair, L., Marti,

J., Hulme, C., McCabe, C., Nikolova, S., Ferreira, Z. & Perkins, G. D. 2017 Prehospital randomised assessment of a mechanical compression device in out-of-hospital cardiac arrest (PARAMEDIC): A pragmatic, cluster randomised trial and economic evaluation. *Health Technology Assessment* **21** (11), 1–175.

Gillett, R. 1994 An average power criterion for sample size estimation. *Journal of the Royal Statistical Society: Series D (The Statistician)* **43** (3), 389–394.

Global Forum for Health 2002 The 10/90 Report on Health Research 2001-2002. *Tech. Rep.*. World Health Organisation.

Grant, A. M., Altman, D. G., Babiker, A. B., Campbell, M. K., Clemens, F. J., Darbyshire, J. H., Elbourne, D. R., McLeer, S. K., Parmar, M. K., Pocock, S. J., Spiegelhalter, D. J., Sydes, M. R., Walker, A. E. & Wallace, S. A. 2005 Issues in data monitoring and interim analysis of trials. *Health Technology Assessment* **9** (7).

Gravenstein, S., Dahal, R., Gozalo, P. L., Edward Davidson, H., Han, L. F., Taljaard, M. & Mor, V. 2016 A cluster randomized controlled trial comparing relative effectiveness of two licensed influenza vaccines in US nursing homes: Design and rationale. *Clinical Trials* **13** (3), 264–274.

Grayling, M. J. & Mander, A. P. 2021 Accounting for variation in the required sample size in the design of group-sequential trials. *Contemporary Clinical Trials* **107**, 1–5.

Grayling, M. J., Mander, A. P. & Wason, J. M. 2018 Blinded and unblinded sample size reestimation procedures for stepped-wedge cluster randomized trials. *Biometrical Journal* **60** (5), 903–916.

Grayling, M. J. & Wason, J. M. 2020 A web application for the design of multi-arm clinical trials. *BMC Cancer* **20** (1), 1–12.

Grayling, M. J., Wason, J. M. & Mander, A. P. 2017 Stepped wedge cluster randomized controlled trial designs: A review of reporting quality and design features. *Trials* **18** (1), 1–13.

Grayling, M. J. & Wheeler, G. M. 2020 A review of available software for adaptive clinical trial design. *Clinical Trials* **17** (3), 323–331.

Gross, A. S., Harry, A. C., Clifton, C. S. & Della Pasqua, O. 2022 Clinical trial diversity: An opportunity for improved insight into the determinants of variability in drug response. *British Journal of Clinical Pharmacology* **88** (6), 2700–2717.

Grover, S., Xu, M., Jhingran, A., Mahantshetty, U., Chuang, L., Small, W. & Gaffney, D. 2017 Clinical trials in low and middle-income countries — Successes and challenges. *Gynecologic Oncology Reports* **19**, 5–9.

Hade, E. M., Murray, D. M., Pennell, M. L., Rhoda, D., Paskett, E. D., Champion, V. L., Crabtree, B. F., Dietrich, A., Dignan, M. B., Farmer, M., Fenton, J. J., Flocke, S., Hiatt, R. A., Hudson, S. V., Mitchell, M., Monahan, P., Shariff-Marco, S., Slone, S. L., Stange, K., Stewart, S. L. & Strickland, P. A. 2010 Intraclass correlation estimates for cancer screening outcomes: Estimates and applications in the design of group-randomized cancer screening studies. *Journal of the National Cancer Institute - Monographs* **40**, 97–103.

Haeussler, C. & Assmus, A. 2021 Bridging the gap between invention and innovation: Increasing success rates in publicly and industry-funded clinical trials. *Research Policy* **50**, 1–17.

Hall, A. J., Inskip, H. M., Loik, F., Day, N. E., O'Connor, G., Bosch, X., Muir, C. S., Parkin, M., Munoz, N. & Tomatis, L. 1987 The Gambia hepatitis intervention study. *Cancer Res* **47** (21), 5782–5787.

Hampson, L. V. & Jennison, C. 2013 Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **75** (1), 3–54.

Hampson, L. V. & Jennison, C. 2015 Optimizing the data combination rule for seamless phaseII/III clinical trials. *Statistics in Medicine* **34** (1), 39–58.

Han, X., Lin, J., Xu, J., Wang, M., Zee, B. & Chong, K. C. 2019 A Review of Assumed and Reported Intracluster Correlations in Cluster Randomized Trials. *Research square* pp. 1–17.

Hankonen, N., Heino, M. T., Araujo-Soares, V., Sniehotta, F. F., Sund, R., Vasankari, T., Absetz, P., Borodulin, K., Uutela, A., Lintunen, T. & Haukkala, A. 2016 'Let's Move It' - A school-based multilevel intervention to increase physical activity and reduce sedentary behaviour among older adolescents in vocational secondary schools: A study protocol for a cluster-randomised trial. *BMC Public Health* **16** (1), 1–15.

Hargreaves, J. R., Copas, A. J., Beard, E., Osrin, D., Lewis, J. J., Davey, C., Thompson, J. A., Baio, G., Fielding, K. L. & Prost, A. 2015 Five questions to consider before conducting a stepped wedge trial. *Trials* **16** (1), 1–4.

Harris, T., Kerry, S., Victor, C., Iliffe, S., Ussher, M., Fox-Rushby, J., Whincup, P., Ekelund, U., Furness, C., Limb, E., Anokye, N., Ibison, J., DeWilde, S., David, L., Howard, E., Dale, R., Smith, J., Normansell, R., Beighton, C., Morgan, K., Wahlich, C., Sanghera, S. & Cook, D. 2018 A pedometer-based walking intervention in 45- to 75-year-olds, with and without practice nurse support: The PACE-UP three-arm cluster RCT. *Health Technology Assessment* **22** (37), 1–273.

Hauck, W. W., Gilliss, C. L., Donner, A. & Gortner, S. 1991 Randomization by cluster. *Nursing Research* **40** (6), 356–358.

Hawk, M. 2013 The Girlfriends Project: Results of a pilot study assessing feasibility of an HIV testing and risk reduction intervention developed, implemented, and evaluated in community settings. *AIDS education and prevention : official publication of the International Society for AIDS Education* **25** (6), 519–534.

Hawkins, N. M., Jhund, P. S., McMurray, J. J. & Capewell, S. 2012 Heart failure and socioeconomic status: Accumulating evidence of inequality. *European Journal of Heart Failure* **14** (2), 138–146.

He, F. J., Wu, Y., Feng, X. X., Ma, J., Ma, Y., Wang, H., Zhang, J., Yuan, J., Lin, C. P., Nowson, C. & MacGregor, G. A. 2015 School based education programme to reduce salt intake in children and their families (School-EduSalt): Cluster randomised controlled trial. *BMJ (Online)* **350**.

Heller, S., Lawton, J., Amiel, S., Cooke, D., Mansell, P., Brennan, A., Elliott, J., Boote, J., Emery, C., Baird, W., Basarir, H., Beveridge, S., Bond, R., Campbell, M., Chater, T., Choudhary, P., Clark, M., de Zoysa, N., Dixon, S., Gianfrancesco, C., Hopkins, D., Jacques, R., Kruger, J., Moore, S., Oliver, L., Peasgood, T., Rankin, D., Roberts, S., Rogers, H., Taylor, C., Thokala, P., Thompson, G. & Ward, C. 2014 Improving management of type 1 diabetes in the UK: the Dose Adjustment For Normal Eating (DAFNE) programme as a research test-bed. A mixed-method analysis of the barriers to and facilitators of successful diabetes self-management, a health economic analys. *Programme Grants for Applied Research* **2** (5), 1–188.

Hemming, K., Eldridge, S., Forbes, G., Weijer, C. & Taljaard, M. 2017 How to design efficient cluster randomised trials. *BMJ (Online)* **358**, 1–5.

Hemming, K. & Girling, A. 2013 The efficiency of stepped wedge vs. cluster randomized trials: Stepped wedge studies do not always require a smaller sample size. *Journal of Clinical Epidemiology* **66** (12), 1427–1428.

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J. & Lilford, R. J. 2015 The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. *BMJ (Online)* **350**, 1–7.

Hemming, K., Martin, J., Gallos, I., Coomarasamy, A. & Middleton, L. 2021 Interim data monitoring in cluster randomised trials: Practical issues and a case study. *Clinical Trials* **18** (5), 552–561.

Hemming, K. & Taljaard, M. 2016 Sample size calculations for stepped wedge and cluster randomised trials: A unified approach. *Journal of Clinical Epidemiology* **69**, 137–146.

Hemming, K. & Taljaard, M. 2020 Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? *International journal of epidemiology* **49** (3), 1043–1052.

Hogancamp, W. E., Rodriguez, M. & Weinshenker, B. G. 1997 The Epidemiology of Multiple Sclerosis. *Mayo Clinic Proceedings.* **72** (9), 871–878.

Hooper, R., Teerenstra, S., de Hoop, E. & Eldridge, S. 2016 Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine* **35** (26), 4718–4728.

Huang, X. 2016 A Simulation Study to Decide the Timing of an Interim Analysis in a Bayesian Adaptive Dose-Finding Studies with Delayed Responses. *Biometrics & Biostatistics International Journal* **3** (6), 206–212.

Huang, X., Ning, J., Li, Y., Estey, E., Issa, J. P. & Berry, D. A. 2009 Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Statistics in Medicine* **28** (12), 1680–1689.

Hussey, M. A. & Hughes, J. P. 2007 Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials* **28** (2), 182–191.

Hwang, T. J., Carpenter, D., Lauffenburger, J. C., Wang, B., Franklin, J. M. & Kesselheim, A. S. 2016 Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA Internal Medicine* **176** (12), 1826–1833.

Ijsselmuiden, C., Marais, D. L., Becerra-Posada, F. & Ghannem, H. 2012 Africa's neglected area of human resources for health research - the way forward. *South African Medical Journal* **102** (4), 228–233.

International Chronic Granulomatous Disease Cooperative Study Group 1991 A controlled trial of interferon gamma to prevent infection in chronic granulomatous disease. *New England Journal of Medicine* **324** (8), 509–516.

Ip, E. H., Wasserman, R. & Barkin, S. 2011 Comparison of intraclass correlation coefficient estimates and standard errors between using cross-sectional and repeated measurement data: The Safety Check cluster randomized trial. *Contemporary Clinical Trials* **32** (2), 225–232.

Isaakidis, P. & Ioannidis, J. P. 2003 Evaluation of Cluster Randomized Controlled Trials in Sub-Saharan Africa. *American Journal of Epidemiology* **158** (9), 921–926.

Ivers, N. M., Taljaard, M., Dixon, S., Bennett, C., McRae, A., Taleban, J., Skea, Z., Brehaut, J. C., Boruch, R. F., Eccles, M. P., Grimshaw, J. M., Weijer, C., Zwarenstein, M. & Donner, A. 2011 Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: Review of random sample of 300 trials, 2000-8. *BMJ (Online)* **343** (7827), 1–14.

Jaki, T. & Magirr, D. 2013 Considerations on covariates and endpoints in multi-arm multi-stage clinical trials selecting all promisingtreatments. *Statistics in Medicine* **32** (7), 1150–1163.

Jenkins, M., Stone, A. & Jennison, C. 2011 An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* **10** (4), 347–356.

Jennison, C. & Turnbull, B. W. 2006 Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: Opportunities and limitations. *Biometrical Journal* **48** (4), 650–655.

Jones, B. G., Streeter, A. J., Baker, A., Moyeed, R. & Creanor, S. 2021 Bayesian statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review. *Systematic Reviews* **10** (1), 1–14.

Jones, N., Bailey, M. & Lyytikainen, M. 2007 Research Capacity Strengthening in Africa: Trends, Gaps and Opportunities A scoping study commissioned by DFID on behalf of IFORD. *Tech. Rep.*. Oversears Development Institute.

Joseph, P. D., Caldwell, P. H., Tong, A., Hanson, C. S. & Craig, J. C. 2016 Stakeholder views of clinical trials in low- and middle-income countries: A systematic review. *Pediatrics* **137** (2), 1–19.

Jukes, M. C., Turner, E. L., Dubeck, M. M., Halliday, K. E., Inyega, H. N., Wolf, S., Zuilkowski, S. S. & Brooker, S. J. 2017 Improving Literacy Instruction in Kenya Through Teacher Professional Development and Text Messages Support: A Cluster Randomized Trial. *Journal of Research on Educational Effectiveness* **10** (3), 449–481.

Julious, S. A. & Owen, R. J. 2011 A comparison of methods for sample size estimation for non-inferiority studies with binary outcomes. *Statistical Methods in Medical Research* **20** (6), 595–612.

Kairalla, J. A., Coffey, C. S., Thomann, M. A. & Muller, K. E. 2012 Adaptive trial designs: A review of barriers and opportunities. *Trials* **13**, 1–9.

Kasza, J., Hemming, K., Hooper, R., Matthews, J. N. & Forbes, A. B. 2019 Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Statistical Methods in Medical Research* **28** (3), 703–716.

Kerry, S. M. & Bland, M. J. 1998 Trials that randomise practices. How should they be analysed. *Family Practice* **15**, 80–83.

Khan, M. I., Soofi, S. B., Ochiai, R. L., Habib, M. A., Sahito, S. M., Nizami, S. Q., Acosta, C. J., Clemens, J. D. & Bhutta, Z. A. 2012 Effectiveness of Vi capsular polysaccharide typhoid vaccine among children: A cluster randomized trial in Karachi, Pakistan. *Vaccine* **30** (36), 5389–5395.

Khoja, A., Kazim, F. & Ali, N. 2019 Barriers to conducting clinical trials in developing countries. *Ochsner Journal* **19** (4), 294.

Kieser, M. & Friede, T. 2000 Blinded Sample Size Reestimation in Multiarmed Clinical Trials. *Therapeutic Innovation & Regulatory Science* **34** (2), 455–460.

Kieser, M. & Friede, T. 2003 Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine* **22** (23), 3571–3581.

Killip, S., Mahfoud, Z. & Pearce, K. 2004 What Is an Intracluster Correlation Coeffi cient? Crucial Concepts for Primary Care Researchers. *Annals Of Family Medicine* **2** (3), 204–208.

Koenig, F., Brannath, W., Bretz, F. & Posch, M. 2008 Adaptive Dunnett tests for treatment selection. *Statistics in medicine* **27**, 1612 – 1625.

Koepsell, T. D., Wagner, E. H., Cheadle, A. C., Patrick, D. L., Martin, D. C., Diehr, P. H., Perrin, E. B., Kristal, A. R., Allan-Andrilla, C. H. &

Dey, L. J. 1992 Selected methodological issues in evaluating community-based health promotion and disease prevention programs. *Annual review of public health* **13** (1), 31–57.

Koffman, J., Yorganci, E., Murtagh, F., Yi, D., Gao, W., Barclay, S., Pickles, A., Higginson, I., Johnson, H., Wilson, R., Bailey, S., Ewart, C. & Evans, C. 2019 The AMBER care bundle for hospital inpatients with uncertain recovery nearing the end of life: The improvecare feasibility cluster RCT. *Health Technology Assessment* **23** (55).

Korendijk, E. J., Moerbeek, M. & Maas, C. J. 2010 The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics* **35** (5), 566–585.

Kumakech, E., Cantor-Graae, E., Maling, S. & Bajunirwe, F. 2009 Peer-group support intervention improves the psychosocial well-being of AIDS orphans: Cluster randomized trial. *Social Science and Medicine* **68** (6), 1038–1043.

Kunz, C. U., Wason, J. M. & Kieser, M. 2017 Two-stage phase II oncology designs using short-term endpoints for early stopping. *Statistical Methods in Medical Research* **26** (4), 1671–1683.

Kunzmann, K., Grayling, M. J., Lee, K. M., Robertson, D. S., Rufibach, K. & Wason, J. M. S. 2021 A Review of Bayesian Perspectives on Sample Size Derivation for Confirmatory Trials. *American Statistician* **75** (4), 424–432.

L. Gómez, H., A. Pinto, J., Castañeda, C. & S. Vallejos, C. 2015 Current barriers for developing clinical research in Latin America: A cross-sectional survey of medical oncologists. *Clinical Research and Trials* **1** (2), 22–28.

Lagakos, S. W. & Hoth, D. F. 1992 Surrogate markers in AIDS: where are we? Where are we going? *Annals of Internal Medicine* **116** (7), 599–601.

Lake, S., Kammann, E., Klar, N. & Betensky, R. 2002 Sample size re-estimation in cluster randomization trials. *Statistics in Medicine* **21** (10), 1337–1350.

Lan, K. K. & Wittes, J. T. 2012 Some thoughts on sample size: A Bayesian-frequentist hybrid approach. *Clinical Trials* **9** (5), 561–569.

Lancaster, G. A., Dodd, S. & Williamson, P. R. 2004 Design and analysis of pilot studies: Recommendations for good practice. *Journal of Evaluation in Clinical Practice* **10** (2), 307–312.

Lang, T. 2011 Adaptive trial design: Could we use this approach to improve clinical trials in the field of global health? *American Journal of Tropical Medicine and Hygiene* **85** (6), 967–970.

Law, M., Grayling, M. J. & Mander, A. P. 2020 Multi-outcome trials with a generalised number of efficacious outcomes.

Lawrie, J., Carlin, J. B. & Forbes, A. B. 2015 Optimal stepped wedge designs. *Statistics and Probability Letters* **99**, 210–214.

Lewis, J. & Julious, S. A. 2021 Sample sizes for cluster-randomised trials with continuous outcomes: Accounting for uncertainty in a single intra-cluster correlation estimate. *Statistical Methods in Medical Research* **30** (11), 2459–2470.

Leyrat, C., Morgan, K. E., Leurent, B. & Kahan, B. C. 2018 Cluster randomized trials with a small number of clusters: Which analyses should be used? *International Journal of Epidemiology* **47** (1), 321–331.

Lim, C. Y. & In, J. 2019 Randomization in clinical studies. *Korean Journal of Anesthesiology* **72** (3), 221–232.

Lindquist, E. F. 1940 *Statistical analysis in educational research.*. Boston: Houghton Mifflin.

Liu, H., Lin, X. & Huang, X. 2019 An oncology clinical trial design with randomization adaptive to both short- and long-term responses. *Statistical Methods in Medical Research* **28** (7), 2015–2031.

Liu, J. L., Wyatt, J. C., Deeks, J. J., Clamp, S., Keen, J., Verde, P., Ohmann, C., Wellwood, J., Dawes, M. & Altman, D. G. 2006 Systematic reviews of clinical decision tools for acute abdominal pain. *Health Technology Assessment* **10** (47).

Maca, J. 2006 Opportunities and challenges in seamless development. *Drug Information Journal* **40**, 463–473.

Magirr, D., Jaki, T. & Whitehead, J. 2012 A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* **99** (2), 494–501.

Marshall, S. J. 2004 In focus Developing countries face. *Bulletin of the World Health Organization* **82** (7).

Martin, L., Hutchens, M., Hawkins, C. & Radnov, A. 2017 How much do clinical trials cost? *Nature Reviews Drug Discovery* **16** (6), 381–382.

Matsoso, P., Auton, M., Banoo, S., Fomundam, H., Leng, H. & Noazin, S. 2005 How does the regulatory framework affect incentives for research and development. *Geneva: CIPIH Study Paper* .

Mayeux, R., Marder, K., Cote, L. J., Denaro, J., Hemengegildo, N., Mejia, H., Tang, M.-X., Lantigua, R., Wilder, D., Gurland, B. & Hauser, A. 1995 The frequency of idiopathic Parkinson's disease by age, ethnic group, and sex in northern Manhattan, 1988-1993. *American journal of epidemiology* **142** (8), 820 – 827.

Mbuagbaw, L., Thabane, L., Ongolo-Zogo, P. & Lang, T. 2011 The challenges and opportunities of conducting a clinical trial in a low resource setting: the case of the Cameroon mobile phone SMS (CAMPS) trial, an investigator initiated trial. *Trials* **12** (1), 145.

McMichael, C., Waters, E. & Volmink, J. 2005 Evidence-based public health: What does it offer developing countries? *Journal of Public Health* **27** (2), 215–221.

Menya, D., Platt, A., Manji, I., Sang, E., Wafula, R., Ren, J., Cheruiyot, O., Armstrong, J., Neelon, B. & O'Meara, W. P. 2015 Using pay for performance incentives (P4P) to improve management of suspected malaria fevers in rural Kenya: A cluster randomized controlled trial. *BMC Medicine* **13** (1), 1–13.

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

Moll, E., Bossuyt, P. M., Korevaar, J. C., Lambalk, C. B. & Van Der Veen, F. 2006 Effect of clomifene citrate plus metformin and clomifene citrate plus placebo on induction of ovulation in women with newly diagnosed polycystic ovary syndrome: Randomised double blind clinical trial. *British Medical Journal* **332** (7556), 1485–1488.

Mor, V., Volandes, A. E., Gutman, R., Gatsonis, C. & Mitchell, S. L. 2017 PRagmatic trial of Video Education in Nursing homes: The design and rationale for a pragmatic cluster randomized trial in the nursing home setting. *Clinical Trials* **14** (2), 140–151.

Mosha, J. F., Kulkarni, M. A., Lukole, E., Matowo, N. S., Pitt, C., Messenger, L. A., Mallya, E., Jumanne, M., Aziz, T., Kaaya, R., Shirima, B. A., Isaya, G., Taljaard, M., Martin, J., Hashim, R., Thickstun, C., Manjurano, A., Kleinschmidt, I., Mosha, F. W., Rowland, M. & Protopopoff, N. 2022 Effectiveness and cost-effectiveness against malaria of three types of dual-active-ingredient long-lasting insecticidal nets (LLINs) compared with pyrethroid-only LLINs in Tanzania: a four-arm, cluster-randomised trial. *The Lancet* **399** (10331), 1227–1241.

Mullard, A. 2014 New drugs cost US$2.6 billion to develop. *Nature Reviews Drug Discovery* **13** (12), 877–877.

Müller, H.-H. & Schäfer, H. 2001 Adaptive Group Sequential Designs for Clinical Trials : Combining the Advantages of Adaptive and of Classical Group Sequential Approaches. *Biometrics* **57** (3), 886–891.

Munos, B. 2009 Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug Discovery* **8** (12), 959–968.

Murray, D. M., Varnell, S. P. & Blitstein, J. L. 2004 Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments. *American Journal of Public Health* **94** (3), 423–432.

National Institute for Health and Care Reseearch 2023 About the HTA Journal.

Nevens, H., Harrison, J., Vrijens, F., Verleye, L., Stocquart, N., Marynen, E. & Hulstaert, F. 2019 Budgeting of non-commercial clinical trials: Development of a budget tool by a public funding agency. *Trials* **20** (1), 1–10.

NHS 2019 Why vaccination is safe and important.

Nor Aripin, K. N., Sammons, H. M. & Choonara, I. 2010 Published pediatric randomized drug trials in developing countries, 1996-2002. *Pediatric Drugs* **12** (2), 99–103.

O'Brien, K. S., Byanju, R., Kandel, R. P., Poudyal, B., Gautam, M., Gonzales, J. A., Porco, T. C., Whitcher, J. P., Srinivasan, M., Upadhyay, M., Lietman, T. M. & Keenan, J. D. 2018 Village-Integrated Eye Worker trial (VIEW): Rationale and design of a cluster-randomised trial to prevent corneal ulcers in resource-limited settings. *BMJ Open* **8** (8), 1–9.

Offorha, B. C., Walters, S. J. & Jacques, R. M. 2022 Statistical analysis of publicly funded cluster randomised controlled trials: a review of the National Institute for Health Research Journals Library. *Trials* **23** (1), 1–16.

O'Hagan, A. & Stevens, J. W. 2001 Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making* **21** (3), 219–230.

O'Hagan, A., Stevens, J. W. & Campbell, M. J. 2005 Assurance in clinical trial design. *Pharmaceutical Statistics* **4** (3), 187–201.

O'Grady, M. A., Lincourt, P., Greenfield, B., Manseau, M. W., Hussain, S., Genece, K. G. & Neighbors, C. J. 2021 A facilitation model for implementing quality improvement practices to enhance outpatient substance use disorder treatment outcomes: a stepped-wedge randomized controlled trial study protocol. *Implementation Science* **16** (1), 1–12.

Pagel, C., Prost, A., Lewycka, S., Das, S., Colbourn, T., Mahapatra, R., Azad, K., Costello, A. & Osrin, D. 2011 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**, 1–12.

Pai, H. 2011 Barriers to conducting clinical research in reproductive medicine: India. *Fertility and Sterility* **96** (4), 809–810.

Pallan, M., Griffin, T., Hurley, K. L., Lancashire, E., Blissett, J., Frew, E., Griffith, L., Hemming, K., Jolly, K., McGee, E., Thompson, J. L., Jackson, L., Gill, P., Parry, J. & Adab, P. 2019 Cultural adaptation of an existing children's weight management programme: The CHANGE intervention and feasibility RCT. *Health Technology Assessment* **23** (33), vii–165.

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

Parkinson Study Group 2000 Pramipexole vs levodopa as initial treatment for Parkinson Disease: A 4-year randomized controlled trial. *Archives of Neurology* **284** (15), 1044–1053.

Perriat, D., Balzer, L., Hayes, R., Lockman, S., Walsh, F., Ayles, H., Floyd, S., Havlir, D., Kamya, M., Lebelonyane, R., Mills, L. A., Okello, V., Petersen, M., Pillay, D., Sabapathy, K., Wirth, K., Orne-Gliemann, J. & Dabis, F. 2018 Comparative assessment of five trials of universal HIV testing and treatment in sub-Saharan Africa. *Journal of the International AIDS Society* **21** (1).

Plummer, M. 2021 rjags: Bayesian Graphical Models using MCMC.

Pocock, S. J. 2013 *Clinical trials: a practical approach*. John Wiley & Sons.

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C. & Bauer, P. 2005 Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* **24** (24), 3697–3714.

Prentice, R., Rossouw, J., Furberg, C., Johnson, S., Henderson, M., Cummings, S., Manson, J., Freedman, L., Oberman, A., Kuller, L. & Anderson, G. 1998 Design of the WHI Clinical Trial and Observational Study. *Control Clin Trials* **19**, 61 – 109.

Prentice, R. L. 1989 Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine* **8** (4), 431–440.

Prost, A., Binik, A., Abubakar, I., Roy, A., De Allegri, M., Mouchoux, C., Dreischulte, T., Ayles, H., Lewis, J. J. & Osrin, D. 2015 Logistic, ethical, and political dimensions of stepped wedge trials: Critical review and case studies. *Trials* **16** (1).

Prudhomme O'Meara, W., Menya, D., Laktabai, J., Platt, A., Saran, I., Maffioli, E., Kipkoech, J., Mohanan, M. & Turner, E. L. 2018 Improving rational use of ACTs through diagnosis-dependent subsidies: Evidence from a cluster-randomized controlled trial in western Kenya. *PLoS Medicine* **15** (7), 1–24.

Quan, H., Luo, X., Zhou, T. & Zhao, P. L. 2020 Seamless phase II/III/IIIb clinical trial designs with different endpoints for different phases. *Communications in Statistics - Theory and Methods* **49** (22), 5436–5454.

Rahman, A., Malik, A., Sikander, S., Roberts, C. & Creed, F. 2008 Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: a cluster-randomised controlled trial. *The Lancet* **372** (9642), 902–909.

Ridout, M. S., Demetrio, C. G. B. & Firth, D. 1999 Estimating intraclass correlation for binary data. *Biometrics* **55** (1), 137–148.

Ring, H., Howlett, J., Pennington, M., Smith, C., Redley, M., Murphy, C., Hook, R., Platt, A., Gilbert, N., Jones, E., Kelly, J., Pullen, A., Mander, A., Donaldson, C., Rowe, S., Wason, J. & Irvine, F. 2018 Training nurses in a competency framework to support adults with epilepsy and intellectual disability: The EpAID cluster RCT. *Health Technology Assessment* **22** (10), 1–104.

Roca, A., Hill, P. C., Townend, J., Egere, U., Antonio, M., Bojang, A., Akisanya, A., Litchfield, T., Nsekpong, D. E., Oluwalana, C., Howie, S. R., Greenwood, B. & Adegbola, R. A. 2011 Effects of community-wide vaccination

with PCV-7 on pneumococcal nasopharyngeal carriage in the Gambia: A cluster-randomized trial. *PLoS Medicine* **8** (10).

Rosala-Hallas, A., Bhangu, A., Blazeby, J., Bowman, L., Clarke, M., Lang, T., Nasser, M., Siegfried, N., Soares-Weiser, K., Sydes, M. R., Wang, D., Zhang, J. & Williamson, P. R. 2018 Global health trials methodological research agenda: Results from a priority setting exercise. *Trials* **19** (1), 1–8.

Rothwell, J. C., Julious, S. A. & Cooper, C. L. 2018 A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal Suzie Cro. *Trials* **19** (1), 1–13.

Rothwell, J. C., Julious, S. A. & Cooper, C. L. 2022 Adjusting for bias in the mean for primary and secondary outcomes when trials are in sequence. *Pharmaceutical Statistics* **21** (2), 460–475.

Rubagumya, F., Hopman, W. M., Gyawali, B., Mukherji, D., Hammad, N., Pramesh, C. S., Zubaryev, M., Eniu, A., Tsunoda, A. T., Kutluk, T., Aggarwal, A., Sullivan, R. & Booth, C. M. 2022 Participation of Lower and Upper Middle-Income Countries in Clinical Trials Led by High-Income Countries. *JAMA Network Open* **5** (8).

Rutterford, C., Copas, A. & Eldridge, S. 2015 Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology* **44** (3), 1051–1067.

Sampson, A. R. & Sill, M. W. 2005 Drop-the-Losers Design: Normal Case. *Biometrical Journal* **47** (3), 257–268.

Schie, S. & Moerbeek, M. 2014 Re-estimating sample size in cluster randomised trials with active recruitment within clusters.

Schlaff, W. D. 2011 Barriers to conducting clinical research in reproductive medicine around the world. *Fertility and Sterility* **96** (4), 801.

Searle, S. 1971 *Linear Models*. New York: Wiley.

Seruga, B., Sadikov, A., Cazap, E. L., Delgado, L. B., Digumarti, R., Leighl, N. B., Meshref, M. M., Minami, H., Robinson, E., Yamaguchi, N. H., Pyle, D. & Cufer, T. 2014 Barriers and Challenges to Global Clinical Cancer Research. *The Oncologist* **19** (1), 61–67.

Shanyinde, M., Pickering, R. M. & Weatherall, M. 2011 Questions asked and answered in pilot and feasibility randomized controlled trials. *BMC Medical Research Methodology* **11**.

Siderowf, A. D. 2004 Evidence from Clinical Trials: Can We Do Better? *NeuroRx* **1** (3), 363–371.

Speed, C., Heaven, B., Adamson, A., Bond, J., Corbett, S., Lake, A. A., May, C., Vanoli, A., McMeekin, P., Moynihan, P., Rubin, G., Steen, I. N. & McColl, E. 2010 LIFELAX - diet and LIFEstyle versus LAXatives in the management of chronic constipation in older people: Randomised controlled trial. *Health Technology Assessment* **14** (52), 1–105.

Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. 2004 *Bayesian approaches to clinical trials and health-care evaluation*, , vol. 13. John Wiley & Sons.

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

Spiegelhalter, D. J., Freedman, L. S. & Parmar, M. K. B. 1994 Bayesian Approaches to Randomized Trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **157** (3), 357–387.

Stallard, N. 2010 A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine* **29** (9), 959–971.

Stallard, N. & Friede, T. 2008 A group-sequential design for clinical trials with treatment selection. *Statistics in medicine* **27**, 6209–6277.

Stallard, N., Hampson, L., Benda, N., Brannath, W., Burnett, T., Friede, T., Kimani, P. K., Koenig, F., Krisam, J., Mozgunov, P., Posch, M., Wason, J., Wassmer, G., Whitehead, J., Williamson, S. F., Zohar, S. & Jaki, T. 2020 Efficient Adaptive Designs for Clinical Trials of Interventions for COVID-19. *Statistics in Biopharmaceutical Research* **12** (4), 483–497.

Stallard, N. & Todd, S. 2003 Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* **22** (5), 689–703.

Stallard, N. & Todd, S. 2011 Seamless phase II/III designs. *Statistical Methods in Medical Research* **20** (6), 623–634.

Sulthan, N. 2015 Perception of clinical research among clinical investigators in Saudi Arabia. *Asian Journal of Pharmaceutical and Clinical Research* **8** (3), 243–246.

Surr, C. A., Holloway, I., Walwyn, R. E., Griffiths, A. W., Meads, D., Kelley, R., Martin, A., McLellan, V., Ballard, C., Fossey, J., Burnley, N., Chenoweth, L., Creese, B., Downs, M., Garrod, L., Graham, E. H., Amanda Lilley-Kelley, J. M., Millard, H., Perfect, D., Robinson, L., Robinson, O., Shoesmith, E., Siddiqi, N., Stokes, G., Wallace, D. & Farrin, A. J. 2020 Dementia Care Mapping ™ to reduce agitation in care home residents with dementia : the EPIC cluster RCT. *Health Economics Review* **24** (16).

Swiger, L. A., Harvey, W. R., Everson, D. O. & Gregory, K. E. 1964 The variance of intraclass correlation involving groups with one observation. *Biometrics* **20** (4), 818–826.

Taljaard, M., Weijer, C., Grimshaw, J. M., Eccles, M. P. & Ottawa Ethics of Cluster Randomised Trials Consensus Group 2013 The Ottawa Statement on the ethical design and conduct of cluster randomised trials: precis for researchers and research ethics committees. *BMJ (Clinical research ed.)* **346**, 1–7.

Teare, M. D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A. & Walters, S. J. 2014 Sample size requirements to estimate key design parameters from external pilot randomised controlled trials : a simulation study. *Trials* **15**, 1–13.

Temple, R. 1999 Are surrogate markers adequate to assess cardiovascular disease drugs? *Jama* **282** (8), 790–795.

Thall, P. F., Simon, R. & Ellenberg, S. S. 1988 Two-Stage Selection and Testing Designs for Comparative Clinical Trials. *Biometrika* **75** (2), 303–310.

Thall, P. F., Siomn, R. & Ellenberg, S. S. 1989 A Two-Stage Design for Choosing among Several Experimental Treatments and a Control in Clinical Trials. *Biometrics* **45**, 537–547.

The Action to Control Cardiovascular Risk in Diabetes Study Group 2017 ACCORD: Action to Control Cardiovascular Risk in Diabetes. *Diabetes and Primary Care* **19** (3), 106–107.

The World Bank 2022 The World Bank in Middle Income Countries.

Theocharidou, L. & Mulvey, M. R. 2018 The effect of deprivation on coronary heart disease mortality rate. *Bioscience Horizons* **11**, 1–6.

Thompson, S. G., Pyke, S. D. M. & Hardy, R. J. 1997 The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. *Statistics in medicine* **16** (18), 2063–2079.

Thorlund, K., Haggstrom, J., Park, J. J. & Mills, E. J. 2018 Key design considerations for adaptive clinical trials: A primer for clinicians. *BMJ (Online)* **360**, 1–5.

Tishkovskaya, S. V., Sutton, C. J., Thomas, L. H. & Watkins, C. L. 2023 Determining the sample size for a cluster-randomised trial using knowledge elicitation: Bayesian hierarchical modelling of the intracluster correlation coefficient. *Clinical Trials* **20** (3), 293–306.

Torgerson, D. J. 2001 Contamination in trials: Is cluster randomisation the answer? *BMJ(Education and Debate)* **322**, 355–357.

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

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

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

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

UNCTAD 2022 Now 8 billion and counting: Where the world's population has grown most and why that matters.

Van Breukelen, G. J. & Candel, M. J. 2012 Calculating sample sizes for cluster randomized trials: We can keep it simple and efficient! *Journal of Clinical Epidemiology* **65** (11), 1212–1218.

Walter, S. D., Han, H., Guyatt, G. H., Bassler, D., Bhatnagar, N., Gloy, V., Schandelmaier, S. & Briel, M. 2020 A systematic survey of randomised trials that stopped early for reasons of futility. *BMC Medical Research Methodology* **20** (1), 1–11.

Wang, M. D. 2007 Sample size reestimation by Bayesian prediction. *Biometrical Journal* **49** (3), 365–377.

Wason, J., Marshall, A., Dunn, J., Stein, R. C. & Stallard, N. 2014 Adaptive designs for clinical trials assessing biomarker-guided treatment strategies. *British journal of cancer* **110** (8), 1950–1957.

Wason, J., Stallard, N., Bowden, J. & Jennison, C. 2017 A multi-stage drop-the-losers design for multi-arm clinical trials. *Statistical Methods in Medical Research* **26** (1), 508–524.

Wason, J. M., Brocklehurst, P. & Yap, C. 2019 When to keep it simple - Adaptive designs are not always useful. *BMC Medicine* **17** (1), 1–7.

WHO 2013 Research for universal health coverage: World health report 2013.

Witham, M. D., Anderson, E., Carroll, C., Dark, P. M., Down, K., Hall, A. S., Knee, J., Maier, R. H., Mountain, G. A., Nestor, G., Oliva, L., Prowse, S. R., Tortice, A., Wason, J. & Rochester, L. 2020 Developing a roadmap to improve trial delivery for under-served groups: Results from a UK multi-stakeholder process. *Trials* **21** (1), 1–9.

Wittes, J. & Brittain, E. 1990 The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in medicine* **9** (1-2), 65–72.

Wittes, J., Lakatos, E. & Probstfield, J. 1989 Surrogate endpoints in clinical trials: cardiovascular diseases. *Statistics in Medicine* **8** (4), 415 – 425.

Woertman, W., De Hoop, E., Moerbeek, M., Zuidema, S. U., Gerritsen, D. L. & Teerenstra, S. 2013 Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology* **66** (7), 752–758.

World Health Organisation 2022 Vaccines and immunization.

Wright, B., Marshall, D., Adamson, J., Ainsworth, H., Ali, S., Allgar, V., Moore, D. C., Cook, E., Dempster, P., Hackney, L., McMillan, D., Trepél, D. & Williams, C. 2016 Social stories™ to alleviate challenging behaviour and social difficulties exhibited by children with autism spectrum disorder in mainstream schools: Design of a manualised training toolkit and feasibility study for a cluster randomised controlled trial w. *Health Technology Assessment* **20** (6).

Wu, S., Crespi, C. M. & Wong, W. K. 2012 Comparison of Methods for Estimating the Intraclass Correlation Coefficient for Binary Responses in Cancer Prevention Cluster Randomized Trials. *Contemp Clin Trials.* **33** (5), 869 – 880.

Zegers-Hochschild, F. 2011 Barriers to conducting clinical research in reproductive medicine: Latin America. *Fertility and Sterility* **96** (4), 802–804.

Zhong, W., Koopmeiners, J. S. & Carlin, B. P. 2013 A two-stage Bayesian design with sample size reestimation and subgroup analysis for phase II binary response trials. *Contemp Clin Trials.* **36** (2), 1–18.

Zhuang, R. & Chen, Y. Q. 2020 Measuring Surrogacy in Clinical Research: With an application to studying surrogate markers for HIV Treatment-as-Prevention. *Stat Biosci.* **12** (3), 295–323.