Advancing knowledge in stepped-wedge cluster randomised controlled trials

By

James Thomas Martin

A thesis submitted to the University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

Institute of Applied Health Research
University of Birmingham
June 2017

Abstract

The stepped-wedge cluster randomised trial (SW-RT) is an increasingly used alternative to the standard parallel cluster randomised trial (P-CRT). This thesis extends the existing knowledge to enhance the methodological quality of future SW-CRTs.

A methodological review of sample size calculations found a poor standard of reporting and substandard methodological rigor. Two key points emanating from the review were the lack of consideration of the decay of correlation over time, and little evidence of varying cluster size being included in the sample size calculation.

Since SW-CRTs are longitudinal in design, they can be split into numerous time-periods – which may impact the correlation structure of the observations within a cluster. The intra-cluster correlation coefficient is usually used in sample size calculations, but may not be sufficient in longitudinal CRTs. Instead the decay of correlation over time may need to be described using the inter-period correlation and the within-period correlation. However, there is currently a dearth in the literature on likely values of these – for which a set of estimates are reported for outcomes associated with type-2 diabetes.

Though CRTs are likely to contain varying cluster size, there is a paucity of research into the effect of varying cluster size in a SW-CRT and a lack of appropriate methodology to adjust power calculations when clusters vary in size. A simulation study provides evidence that the SW-CRT is affected less, on average, than a P-CRT by varying cluster size, but there is a much larger degree of variability in the power of a SW-CRT. A practical method for estimating power in a SW-CRT with varying cluster size is then established through a Stata function.

Acknowledgements

Firstly, I would like to thank Richard Riley. Without his inspiring lectures, I may never have found the world of medical statistics, and so the opportunity to carry out this research would never have arisen.

I would like to thank my supervisors, Karla Hemming, Alan Girling and Jon Deeks, for their guidance and support throughout the last 5 years. I have been extremely lucky to have worked with you. This hasn’t been an easy journey for me, and your continuous help and advice has guided me towards completing this thesis. You have given me opportunities to further myself personally and academically, and have helped to shape my future. I will be eternally grateful.

To my family, thank you for all of your support. From school to university, it has been a long journey, and I could not have managed it without you. Mom and Dad, thank you for your everlasting belief and for supporting me through this. Ryan, thank you for keeping me grounded throughout my life as a student.

Lastly to Kelsey – the most amazing partner I could ask for – I dedicate this work to you. Our relationship began at the same time as this PhD started, and you have been the biggest source of support throughout it all. You have helped me through all of the highs and lows I have had, and have made these 5 years an unforgettable experience. I am so thankful to have you by my side and would not have managed this without you.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF BOXES

Dissemination of research

Whilst completing this thesis, aspects of the methods and results have been disseminated into the research community. For this, the author is the primary contributor of the design, methods, analysis, and authorship of these pieces of research. Supervisors and other researchers have collaborated for publication purposes. The results of chapter 3 and 5 have been presented as poster presentations at two international conferences.

The results of chapter 3 have been published as a peer-reviewed article. The methods and results of chapter 4 have been published as a peer-reviewed article and presented as a poster presentation and as an oral presentation at two international conferences.

List of publications:

Martin J, Girling A, Nirantharakumar K, Ryan R, Marshall T, Hemming K. Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. Trials. 2016;17:402.

Martin J, Taljaard M, Girling A, Hemming K. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. BMJ open. 2016;6(2):e010166.

Related manuscripts:

Hemming K, Girling A, Martin J, Bond SJ. Stepped wedge cluster randomized trials are efficient and provide a method of evaluation without which some interventions would not be evaluated. Journal of clinical epidemiology. 2013;66(9):1058-9.

Oral presentations:

Martin J, Girling A, Hemming K. How are stepped-wedge cluster trials impacted by unequal cluster sizes? Oral presentation at the Young Statisticians Meeting; 2016 Aug 16-19; University College London, UK.

Poster presentations:

Martin J, Girling A, Nirantharakumar K, Ryan R, Marshall T, Hemming K. Intra-cluster correlation coefficients for clustered randomised controlled trials in type II diabetes. Poster presentation at the Young Statisticians Meeting; 2013 July 4-5; Imperial College London, UK

Martin J, Girling A, Hemming K. The effect of varying cluster size on the precision of stepped wedge cluster randomised trials. Poster presentation at the International Society for Clinical Biostatistics; 2014 Aug 24-28; University of Vienna, Austria.

Martin J, Taljaard M, Girling A, Hemming K. Reporting and methodological quality of sample size calculations in stepped-wedge cluster randomised trials: a methodological review. Poster presentation at the International Society for Clinical Biostatistics; 2015 Aug 23-27; University of Utrecht, Netherlands.

CHAPTER 1: INTRODUCTION

1.1 Cluster randomised trials

Clustered randomised trials (CRTs) are often a more feasible trial type than a conventional individually randomised controlled trial (RCT) (1). Situations may arise in which the intervention can be more naturally implemented at the cluster level, such as an obesity prevention programme targeting primary schools (2). In these situations, the randomisation of clusters, rather than the individuals within them, may reduce the risk of contamination (1). Another reason that a CRT may be more appropriate than an individually randomised trial is financial limitations - whereby the cost of implementing an intervention at the cluster level may be significantly less than implementation at the individual level (3). Additionally, the randomisation of clusters may alleviate concerns with ethics, or logistics, which may otherwise be present with an individually randomised design.

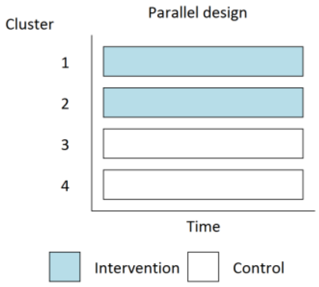
1.2 Different forms of a cluster randomised trial

When considering CRTs, there are a variety of ways in which they may be setup, some of which are highlighted below.

1.2.1 Parallel cluster randomised trials

The most common design type is the parallel cluster randomised trial (P-CRT). A P-CRT involves randomising half of the clusters to the intervention, and the remaining half to the control. Figure 1.1 shows that all participants in the cluster either receive the intervention (shaded region) or the control (non-shaded region), and the intervention is not removed once implemented.

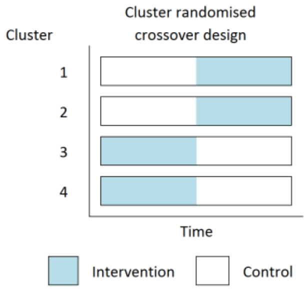
Figure 1.1: Schematic representation of a parallel design



1.2.2 Cluster randomised crossover trials

A cluster randomised crossover (CRXO) design implements the intervention in all clusters in the study, but removes the intervention from some of the clusters during the study. A CRXO trial typically incorporates two time-periods, with half of the clusters randomised to receive the intervention during time-period one, before switching to the control during time-period two, whilst the other half receive the control during time-period one before switching to the intervention during time-period two (Figure 1.2).

Figure 1.2: Schematic representation of a cluster randomised crossover design



Some CRXO trials require a washout period to remove the effects of the treatment from the participants (or clusters) (4). This washout period may extend the study duration by a significant amount and require knowledge of the dynamics of the treatment which may be unknown during the planning stage of a trial. In trials in which the intervention is an educational package, it may be impossible to withdraw the intervention, since it would be impossible to unlearn a skill – see for example, the trial by Bashour et al. (5).

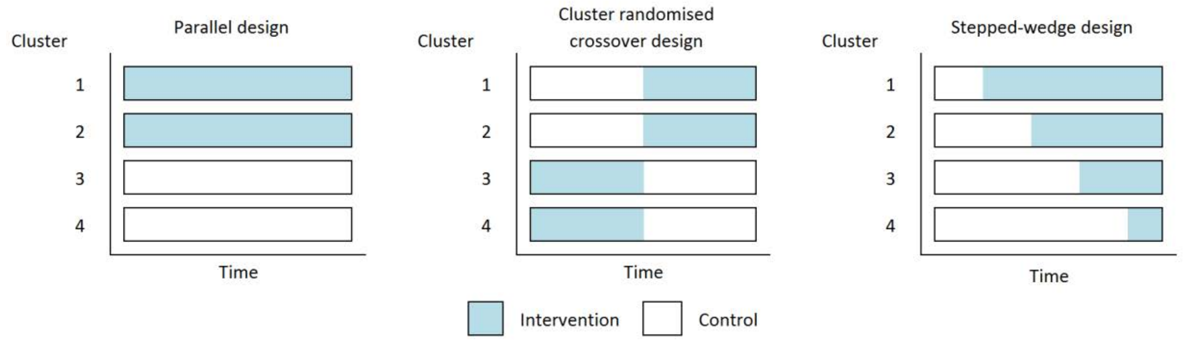
1.3 Stepped-wedge cluster randomised trials

A stepped-wedge cluster randomised trial (SW-CRT) involves the sequential, but random, rollout of the intervention to clusters over multiple time-periods (6-8) and so can be viewed as a unidirectional crossover trial. In a SW-CRT, rather than randomising clusters to an intervention or control at the study beginning, the clusters are randomised to a time-period in which they will receive the intervention. A SW-CRT typically involves collecting observations at a baseline period in which no clusters are exposed to the intervention, and then at regular intervals (time-periods) in which the intervention is rolled out to a cluster (or group of clusters) – who then cross from the control condition to the intervention condition. This process continues at each time-period (step) until all clusters have crossed over to receive the intervention. A SW-CRT usually has one time-period in which observations are made whilst all clusters are exposed to the intervention.

Observations are typically made in every cluster during each time-period in which a new cluster is exposed to the intervention (9). However, there is some thoughts as to whether a SW-CRT can contain designs in which observations are not collected at each time-period or do not contribute to the analysis (9-12). By collecting data during each time-period, each cluster contributes observations to the control and intervention conditions. A key aspect of the SW-CRT is that the intervention is not removed once implemented. Nevertheless, whilst here we are concerned with clusters crossing from unexposed to exposed, there have been instances of individual randomised stepped-wedge trials (13-17) and studies that cross from exposed to unexposed (18).

Figure 1.3 illustrates a SW-CRT, alongside a CRXO trial and a P-CRT for a trial with 4 clusters (or four groups of clusters).

Figure 1.3: Schematic illustration of three different cluster trial designs



In a P-CRT, clusters are randomised to receive the control or intervention only; whilst in a CRXO trial, clusters are randomised as to whether they will receive the control or intervention first. In a SW-CRT the time-period in which the cluster will crossover to the intervention is randomised.

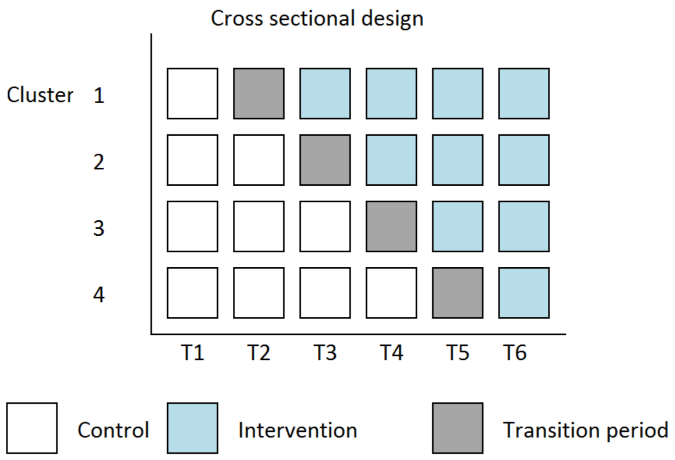
Unlike a P-CRT and CRXO trial, in a SW-CRT, as time increases, the number of clusters exposed to the intervention increases. This means that more clusters are exposed to the intervention at later time-periods than earlier ones. As such, an underlying temporal trend may confound the intervention effect, and so the effect of time must be accounted for in pre-trial power calculations and post-trial analysis.

1.3.1 A typical stepped-wedge cluster randomised trial

In a traditional SW-CRT, one cluster crosses from unexposed to exposed at each step. As such, in a traditional SW-CRT with 5 clusters, the study would contain 5 crossover points. There would be one additional time-period in which data is collected whilst all clusters are unexposed, resulting in 6 time-periods. However, the SW-CRT could also contain a group of clusters crossing over at each step.

There have been many examples in the methodological literature and in trial reports of SW-CRTs following a non-typical design. A SW-CRT may look like the design given in Figure 1.3, but it is not a necessity. All clusters switch from the control to the intervention during a SW-CRT, but this transition may not be instantaneous (19). In these circumstances, clusters may not be fully exposed to the intervention, and so the cluster may be considered as being in a transition period. This transition period can easily be displayed on an illustration of the design (Figure 1.4). Alternatively, the concept of an incomplete design can be used, which indicates that at some steps, and for some clusters, data is not collected, or does not contribute towards the analysis (9, 10). This can be achieved in a trial by not taking measurements, or by not including observations in the analysis, and would allow other designs to be considered as incomplete SW-CRTs (20).

Figure 1.4: Schematic representation of study including a transition period



Since a SW-CRT contains multiple time-periods, it is often suggested that repeated observations from a participant is required (11, 12, 21). However, this may not be the case, and whilst participants may be followed up for the entire study, it is also possible to recruit new participants at each time period. This leads us to define three different design types for a SW-CRT: a cross-sectional design, a cohort design, or an open cohort design, which we discuss below.

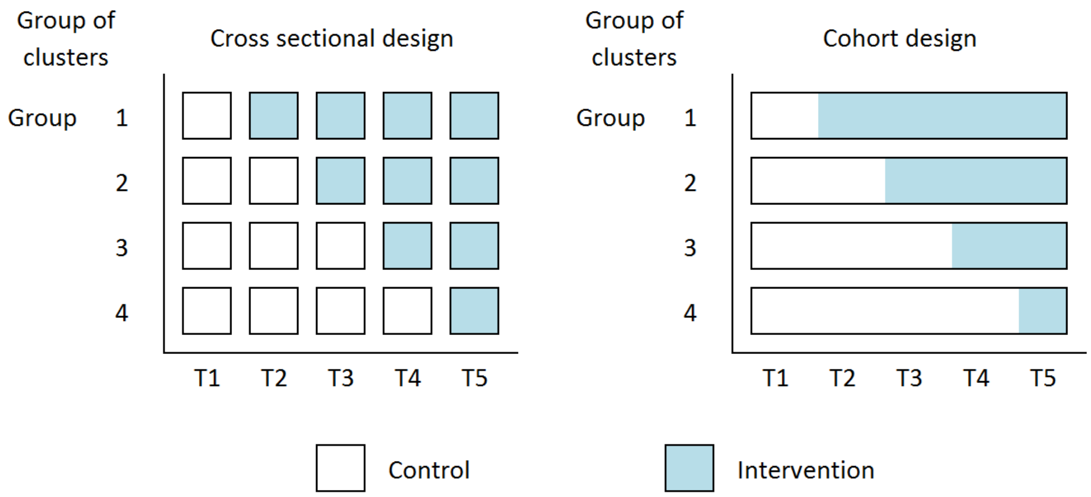
1.4 Cross-sectional, cohort and open cohort designs

As with other CRTs, observations within a SW-CRT may stem from repeat measurements of the same cohort of participants, who are recruited at the start of the study and followed-up for the study length (cohort design); from single measurements taken from individual participants, with new participants recruited at each step (cross-sectional design); or a mixture of the two designs, in which some participants are followed-up for multiple steps, but participants are free to join and leave the study (open cohort design) (6, 22, 23).

Although the term is not used here, a cross-sectional design may instead be described as a continuous recruitment with short-exposure design (19).

Both a cohort design and a cross-sectional study design can be depicted by a schematic representation, which easily allows the difference between the designs types to be seen (Figure 1.5). When illustrating a study design in this manner, each block indicates a distinct group of participants. In the cross-sectional designs, there is a separate block at each time-period, indicating a new group of participants at each time-period. In the cohort design, one block covers the entire study length, signifying that the same participants are followed-up for the study duration.

Figure 1.5: Schematic representation of a cross-sectional design and a cohort design



Each time period (T1 - T5) represents a data collection point. Each block represents a new group of individuals.In a cross-sectional design, data are collected from different samples of individuals, whilst in a cohort design, observations are made on the same individuals over time.

In a SW-CRT that follows a cross-sectional design, the correlation between observations within a cluster is assumed to be independent of the timing in which the observations are made. However, in a cohort design (or an open cohort design) participants are followed up throughout the study, and so there will be a within-participant correlation over time.

Although the cohort and open cohort designs are commonly used, the methodology surrounding the estimation of sample size and post-trial analysis differs from the cross-sectional design. The assumption of a cross-sectional design for a SW-CRT allows for simpler statistical modelling, and so the cross-sectional design has been studied in much more depth than the cohort and open cohort design. In this thesis, we focus on the cross-sectional design.

An alternative view of the designs has been presented based upon the exposure and measurement of observations in which a SW-CRT can be broken down into three types – continuous recruitment with short exposure, closed cohort, and open cohort (19).

1.5 Introduction of terminology and notations

To create consistency throughout this work, some terminology and basic notation will be introduced here. Whilst notations will be explained again, where appropriate, some clarification is provided here.

1.5.1 General notation

When discussing the total number of clusters that a study involves, we shall denote this as C. This is the total number of unique clusters, and not the number of groups of clusters. The number of time-periods in a study is denoted as T. Although in a conventional SW-CRT, this is simply the number of clusters plus one, this is not always the case.

1.5.2 Describing the study design

When considering the design of a SW-CRT, whilst the schematic representation is easy to understand, in some situations it is better represented in matrix form. By denoting the design of a study in matrix form, it can be transformed easily for use in power calculations (see chapter 5). Throughout this body of work, we denote the matrix that represents the study design as the design pattern matrix (DPM). The design pattern matrix typically comprises of 0's and 1's, where a "0" indicates that the cluster is not exposed to the intervention and a "1" indicates that the cluster is exposed to the intervention. If we consider the cross-sectional study design given in Figure 1.5, then the corresponding design pattern matrix can be given by Figure 1.6.

Figure 1.6: Design pattern matrix of a complete design

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

1.5.3 Describing cluster sizes

Since the SW-CRT is comprised of multiple time-periods, the size of a cluster can relate to one of two concepts: the size of a cluster at one time-period (the cluster-period size, m); or the size of a cluster over the whole study period (the total cluster size, M). The differentiation between the cluster-period size (m) and the total cluster size (M) is important when considering whether a SW-CRT is a cross-sectional or cohort design.

Consider a SW-CRT that may follow either a cohort or cross-sectional design, in which observations of participants are made during 5 time-periods and the cluster-period size (m) is 50. In a cohort design, participants are followed-up for the study duration, with repeated measurements for each participant. Here, each of the 50 participants have 5 observations (one per time-period), but the total cluster size will still be 50. Here, $m = 50$ and $M = 50$. In a cross-sectional design, new participants are recruited at each time-period, so that at each time-period, there are 50 new participants that each have one observation made. Now, the total cluster size will be 250 (= 50 x 5). Here $m = 50$ and $M = 250$.

1.5.4 Summary

A summary of the terminology and notations that are used throughout this work is given below in Table 1.1.

Table 1.1: Summary of terminology and notation used

Terminology	Notation	Definition
Cluster-period size	m	The size of a cluster during one measurement period
Total cluster size	M	The size of a cluster over the study duration
Design pattern matrix	DPM	A matrix that can represent the study design
The number of clusters	C	The total number of clusters in the study
Number of time-periods	T	The total number of time-periods in the study

1.6 Examples of a SW-CRT

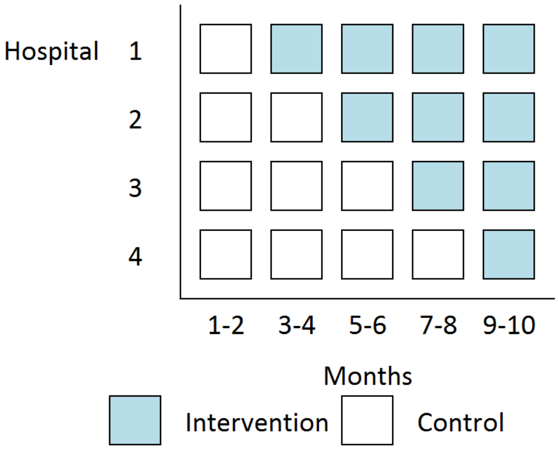
To provide further understanding, a selection of SW-CRTs are presented below, which are reflected on throughout this body of work. For each SW-CRT, we introduce the study aims, alongside some basic details of the study, and a schematic representation of the study design.

1.6.1 Example 1 – The WOSLAD trial

Effect of training doctors in communication skills on women's satisfaction with doctor-woman relationship during labour and delivery: a stepped wedge cluster randomised trial in Damascus

The first study we highlight is a SW-CRT to determine the effect of training hospital staff in interpersonal and communications skills to improve women's satisfaction with doctor-participant relations in labour and delivery room, and we refer to this as the WOSLAD trial (5). This cross-sectional SW-CRT involved the randomisation of 4 hospitals (clusters) over 4 randomisation steps with one additional period following the roll-out (Figure 1.7). Each step was 2 months in duration, resulting in a 10 month study. The study contained 5 time-periods and a fixed cluster size of 100 participants per cluster per time-period were recruited, resulting in 2000 (= 5 x 100 x 4) participants in the study. The included study population were women delivering a baby at each of the included hospitals. The intervention was educational based, with a training package in communication skills given to all resident healthcare professionals at the included hospitals, whilst usual care was conducted in unexposed clusters. The primary outcome was women's satisfaction with interpersonal and communication skills of doctors – measured using a Likert scale questionnaire, with higher scores indicating higher satisfaction. The cluster sizes were fixed at 100, so that only the first 100 participants would be included in the study. As such, there was no variation in the cluster size. The sample size was carried out using the Hussey and Hughes methods – which we discuss in section 2.4.1 – but they did not report a measure of correlation between observations within a cluster.

Figure 1.7: Schematic representation of WOSLAD study



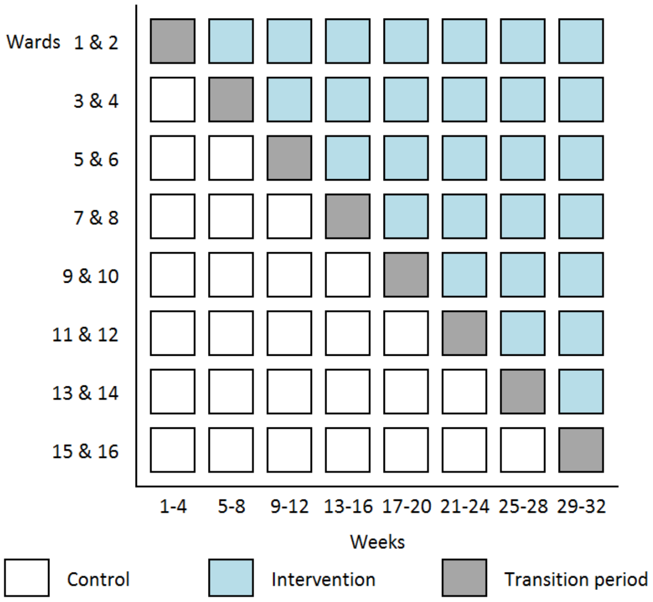
1.6.2 Example 2 – The ICCOT trial

Introducing Critical Care Outreach: a ward randomised trial of phased introduction in a general hospital

The second study aims to investigate whether critical care outreach teams (CCOTs) impact in-hospital mortality and length of stay, and we refer to this as the ICCOT study (24). This incomplete cross-sectional SW-CRT involved the randomisation of 16 acute wards (clusters) within a hospital over 8 randomisation steps, which included a transition phase (Figure 1.8).

Each step was 4 weeks in duration, resulting in a 32 week study. All participants admitted to a ward were included, with 2,903 participants contributing to the primary analysis. The intervention involved educational support and practical help to existing staff, as well as the addition of a team of specialist critical care nurses to provide support. Unexposed clusters received usual care. The primary outcome was in-hospital mortality. The study included all participants admitted to each of 16 wards within the hospital, but did not describe the average ward size or whether the wards were likely to vary in terms of the number of included participants. There is no mention of the correlation between observations within the power calculation. Noticeably, in this SW-CRT, wards 1 and 2 do not contribute any observations to the control period, and wards 15 and 16 do not contribute any observations to the intervention period.

Figure 1.8: Schematic representation of Example ICCOT study

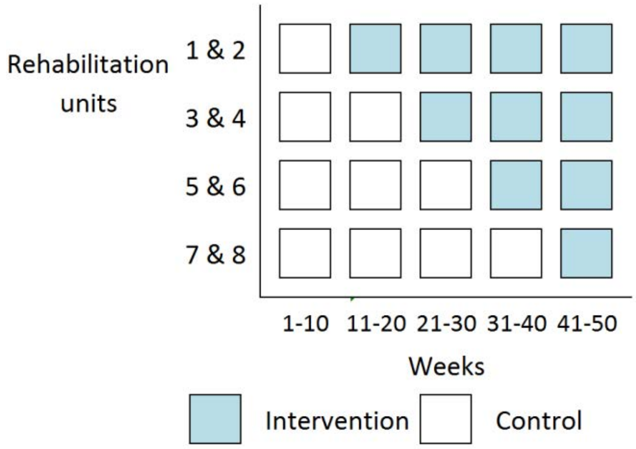


1.6.3 Example 3 – The ERFIC trial

A stepped-wedge cluster randomised controlled trial for evaluating rates of falls among inpatients in aged care rehabilitation units receiving tailored multimedia education in addition to usual care: a trial protocol

The third study aims to evaluate the effectiveness of patient education on the number of falls occurring in a rehabilitation unit, which we refer to as the ERFIC trial (25). This cross-sectional SW-CRT involves the randomisation of eight rehabilitation units (clusters) over four randomisation steps (Figure 1.9), with five time-periods in total. Each step lasted 10-weeks, resulting in a 50-week study. The clusters vary in size, and the sizes are known prior to the study, as they are the number of beds per unit, and are given in ascending order as: 14, 17, 20, 20, 24, 30, 36, and 90. This does not correspond to the order the clusters were randomised in. Only patients who were over 60 years old, cognitively intact, and likely to benefit from the intervention were included in the study. The intervention included patient education, delivered by a physiotherapist, alongside a series of follow-up sessions. Clusters unexposed to the intervention received usual care. The primary outcome was the number of falls made by the patient whilst at the rehabilitation unit. In the power calculation, the average cluster size was used, with no discussion of the variation in cluster sizes. A design effect appropriate for a P-CRT was used to calculate the sample size. An intra-cluster correlation coefficient (ICC) is used to describe the correlation between participants within a cluster.

Figure 1.9: Schematic representation of ERFIC study



Rehabilitation units	1-10	11-20	21-30	31-40	41-50
1 & 2	□	🟦	🟦	🟦	🟦
3 & 4	□	□	🟦	🟦	🟦
5 & 6	□	□	□	🟦	🟦
7 & 8	□	□	□	□	🟦

1.7 Importance of improving methodology in SW-CRTs

The SW-CRT is still a relatively novel design and there is limited methodological research other than into sample size and efficiency, in which a restrictive model framework has been used. As such, there are still many gaps in the literature. For example, it is often assumed when estimating the power of a SW-CRT that a) the correlation between observations within a cluster is independent of the timing of the observations, and b) that the clusters are of equal size. As such, it is likely that pre-trial sample size calculations will not report any deviation to these two assumptions. This is emphasised by the sample size calculations for each of the three example SW-CRTs given in section 1.6. The correlation between observations within a cluster was not reported in the sample size calculation in two of the examples, whilst the third used the ICC – which assumes that the correlation is independent of the timing of the observations. In terms of cluster size in the sample size calculation, of the examples given, one uses a fixed sampling rate, and so the size will not vary between clusters; one did not report the average cluster size or whether any variation would be present; and one reported the known (and varying) cluster sizes – but did not account for this in the power calculation. Recently, the literature has made calls for research into the decay of correlation over time in a SW-CRT, with emphasis on the publication of values of the correlation for future use (26, 27). Additionally, there has been little evidence of research into varying cluster size, and whether the methodology exists to deal with it in a SW-CRT.

1.8 Aims and overview of thesis

The overarching aim of this thesis is to develop the understanding of some key design features of a SW-CRT. In particular, there are four key aims that this work seeks to address:

1. Review the sample size calculations for SW-CRTs to assess whether they are sufficiently reported, and whether appropriate methodology has been used.

2. Evaluate the validity of the assumption that the correlation between observations within a cluster is independent of time.

3. Demonstrate the impact of varying cluster size in a SW-CRT, and make comparisons to the impact of varying cluster size in a P-CRT.

4. Propose a method for estimating power in a SW-CRT when clusters vary in size.

1.9 Summary

The thesis has seven chapters. Chapter 2 reviews the current methodological literature to motivate the research aims. Chapter 3 focuses on reviewing published SW-CRTs to assess the quality of reporting. Chapter 4 focuses on the correlation structure in CRTs. Chapters 5 and 6 focus on varying cluster size. A chapter outline is given below.

Chapter 2 appraises the current methodological literature in SW-CRTs. The main aim is to establish the literature already published, and the current knowledge gaps.

Chapter 3 is a methodological review of SW-CRTs. The CONSORT statement for RCTs and the CONSORT extension for CRTs are used to assess the quality of reporting of sample size calculations in published SW-CRTs. A selection of SW-CRT specific items is also included. An assessment of the methodology is presented to highlight whether previous SW-CRTs have adjusted for varying cluster size or used an extended correlation structure.

Chapter 4 presents a framework for estimating time-dependent correlation in longitudinal CRTs, and defines three types of correlation for settings in which the correlation between observations is dependent on the timing of them. Estimates of the correlation are illustrated for relevant outcomes associated with type-2 diabetes. Also presented is a comparison of methods to be used to estimate the correlation between observations within a cluster for binary outcomes.

Chapter 5 uses simulation to determine the impact of varying cluster size in a SW-CRT. Particularly, to assess how a SW-CRT with varying cluster size compares to a SW-CRT with equal cluster size in terms of precision of a treatment effect estimate. A variety of scenarios are used to fully evaluate the impact. The methodology is extended to P-CRTs, to compare how varying cluster size affects a SW-CRT in relation to a P-CRT.

Chapter 6 proposes a Stata function for estimating power in a SW-CRT, by applying the methodology used in Chapter 5. A variety of examples are presented to illustrate how this function can be used in practice.

Chapter 7 includes a discussion of the overall findings from this thesis. This includes the implications on current research and recommendations for future work.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The stepped-wedge cluster randomised trial (SW-CRT) is a relatively new study design, and so there is a dearth of literature compared to other trial designs. A SW-CRT generally involves an intervention related to education or training, within a health care environment (28). The increasing amount of trial publications has led to a growth in the methodological literature. Early systematic reviews in this area found few trials using this design (23, 29), but this has increased in recent years (30). Below, we discuss the systematic reviews of SW-CRTs that have been published.

2.2 Results from previous systematic reviews

There have been numerous reviews of the SW-CRT literature in order to evaluate the design. Early systematic reviews aimed to give an overview of the SW-CRT (23, 29), whereas recent reviews intended to address more specific problems – such as the design rationale (28) and the statistical methodology used (31).

2.2.1 The Brown and Lilford systematic review

The first systematic review of both randomised and non-randomised stepped-wedge studies in healthcare was carried out in 2006 by Brown and Lilford (23). The review identified 12 SW trials that used either individual level randomisation (4 studies) or cluster level randomisation (8 studies) – of which 3 were cohort designs, and 5 were cross-sectional. Only studies designed with multiple baselines were excluded. It was identified that there was an inconsistency in the motivation for using a SW design, but that a SW-CRT could be useful for the evaluation of an intervention in developing countries. However, there was an inconsistent approach to the reporting of the study, and of the data analysis. None of the included studies would have filled the CONSORT requirements for the reporting of items related to randomisation.

2.2.2 The Mdege systematic review

A 2010 systematic review investigated the motivation for carrying out a SW-CRT (29). As such, they excluded studies in which the randomisation was carried out at the participant level. They also excluded studies that were retrospectively analysed as a SW-CRT when they were not originally designed as one. The 25 included studies highlighted that a SW-CRT may be useful the evaluating an intervention that is believed to do more good than harm, but in which there is lack of evidence of effectiveness. Additionally, a SW-CRT may be more suitable than other trial designs for the evaluation of the routine implementation of an intervention with proven efficacy. However, the included studies were generally of low reporting quality – particularly whether the trial was randomised or whether a power calculation had been used. There was also little reporting of sequence generation, allocation concealment, or blinding, and few studies published a flow diagram of participants through the study. The review also found a large heterogeneity in the analysis methods used and so called for standardised data analysis and reporting. There was no indication of whether trials followed a cross-section or cohort design, but many of the studies contained only two randomisation steps.

2.2.3 The Beard systematic review

A 2015 systematic review of articles published between 2010 and 2014 identified 37 relevant publications (28). This review focused on the types of intervention used, what the studies primary outcome was (and data type), the rationale for conducting a SW-CRT, and whether a sample size calculation was reported. It is shown that SW-CRTs are increasing in frequency – particularly in high-income countries. The majority of the included SW-CRTs involved a training/education based intervention which made a cross-over design impossible as the intervention could not be removed once implemented. It was highlighted that a SW-CRT was often chosen over a P-CRT to prevent participants randomised to the control arm from dropping out, and that a potentially effective intervention should not be withheld from participants. The review highlighted a lack of reporting of the trial design and of a sample size calculation, and echoed calls for guidance for the reporting of a SW-CRT.

2.2.4 The Barker systematic review

A 2016 review included 102 articles, identifying the statistical methods used in the analysis and whether the sample size methodology was appropriate (31). Upon highlighting the primary outcome, they emphasise the statistical method used in the analysis, and whether an adjustment for time was made. There was a large heterogeneity in the analysis methods used. Although the majority acknowledged the longitudinal nature of a SW-CRT, there was evidence of many studies not including time in the analysis. They also report whether a sample size justification was made, and if so, what methodology was used. They also state the design type of the included studies. They found that the sample size calculation was often done using either a design effect appropriate for a P-CRT, or by the methods recommended by Hussey and Hughes (32) – which we later discuss. However, the methods are often only appropriate for cross-sectional designs, and have often been misapplied to cohort designs. Nevertheless, the majority of included studies followed a cross-sectional design. A large proportion of the included studies contained fewer than 20 clusters, and many had fewer than 10 clusters. Due to the small number of clusters in the study, some SW-CRTs are using inappropriate methods of analysis, since many methods require more clusters to be unbiased. It is highlighted that there was a lack of methodological literature at this point in time into cohort designs, and there is often a lack of clarity between a cohort design and a cross-sectional design, and so trialists are confusing sample size methods and methods of analysis that are appropriate for cross-sectional designs and applying them in cohort designs.

2.2.5 Summary of previous systematic reviews

Early systematic reviews of SW-CRTs found a large degree of variation in the reporting quality and their description of the study design (23, 29). The reporting of sample size calculations have changed over time – though are still inadequately reported. Brown and Lilford found only 5 of 12 studies (42%) reported a sample size calculation (23). This had improved to 77 of 102 (77%) in the most recent review (31). However, whilst the reviews assessed whether a sample size was justified, they did not assess individual elements of the sample size calculation in detail.

The reviews have shown there is a lack of consistency in the analysis methods used across published trials (23, 29, 31), highlighting the need for a more consistent approach to data analysis. A substandard degree of reporting may stem from the absence of standardised CONSORT guidelines for SW-CRTs (28).

2.3 Advantages and disadvantages of a SW-CRT

As a relatively novel design type, some potential advantages and disadvantages of the SW-CRT have been explored. Many of these relate to other design types, and whether the SW-CRT is more (or less) appropriate in some settings (11, 12, 21, 22, 33-35). Below, we highlight some of the potential advantages and disadvantages.

2.3.1 Design advantages

There are two reasons often cited for conducting a SW-CRT over a P-CRT: the phased implementation of the intervention; and the providing of the intervention to all clusters (28, 36). The logistical argument is often the strongest motivation for performing a SW-CRT (36) – particularly in implementation research (37). Randomising half of the clusters to the intervention at the beginning of the study is often logistically impossible, and so the randomisation of the intervention over multiple steps may be the only feasible approach to conducting a randomised trial (6, 23, 29). A key feature of the SW-CRT is that all clusters receive the intervention during the study (38) – meaning that in a cohort SW-CRT all participants receive the intervention. However, in a cross-sectional SW-CRT, whilst all clusters will receive the intervention, not all participants will. The SW-CRT may influence study acceptance from the clusters and the participants within them (39). In cohort studies, this advantage may provide motivation for participants (40), and clusters (38) that may be more inclined to participate under the guarantee of the intervention.

2.3.2 Clinical equipoise

It is commonly suggested that a SW-CRT is beneficial for studies in which the intervention is believed to be more beneficial than harmful (23, 29, 41) (i.e. it is believed to be effective) or else has been established as effective in a more controlled settings (39) (i.e. efficacy has been shown). Whilst closely related, these are two distinct points. A SW-CRT may be useful in evaluating the effect of an intervention on a population level, if evidence from an individually randomised trial has shown benefits (1) or if a level of efficacy has been shown in a controlled research setting (39).

In all trials, a central tenet is that there is equipoise between the interventions being compared. Noticeably, in the context of the SW-CRT, it is often said that the SW-CRT should be used when there is a prior belief that the intervention will be superior to the control, and so would provide an ethical paradox to withhold participants from a beneficial treatment (23, 29, 42). Controversially, it has been stated that the SW-CRT can be used in settings in which clinical equipoise is not met, and so would be unethical to withhold the intervention (41). However, in the absence of clinical equipoise, others have argued there is no ethical justification in conducting a trial or delaying the implementation of the intervention across all clusters (37, 43).

The issues of clinical equipoise may provide a convincing argument in some settings to conduct a SW-CRT. However, whilst in cohort designed SW-CRTs, all clusters and participants will receive the intervention, in a cross-sectional SW-CRT, only half of the participants will receive the intervention (39) – the same as a P-CRT.

2.3.3 Number of clusters and observations

Under the caveat of most methodology for SW-CRTs following the Hussey & Hughes mixed-effect model framework (32) (see section 2.4.1), there has been much research surrounding which, of a P-CRT and a SW-CRT, is more efficient, in terms of requiring fewer observations and/or clusters. Early research claimed that a SW-CRT was always more efficient than a P-CRT in terms of sample size (21), which has created much debate (33, 35). Commonly, it is concluded that a SW-CT is more efficient than a P-CRT for larger ICC (22, 44, 45) – due to a SW-CRT using both between-cluster and within-cluster comparisons to estimate the intervention effect (22, 46-48). This may allow studies with a limited number of clusters to obtain sufficient power, rather than conduct a P-CRT that would require an unobtainable number of participants (22, 35). In a P-CRT, increasing the cluster size has a plateauing effect on the power, whereas this is not the case for a SW-CRT (45).

Nevertheless, trials with few clusters may contain problems at the analysis stage (27). The recommended model for SW-CRTs proposed by Hussey and Hughes (32) requires a minimum of ten clusters to adequately estimate the random effects (27). This issue might be amplified for cohort SW-CRTs, since the analysis would require a three level model (as participants would have repeated observations over time) (31). Furthermore, in the case of a small number of clusters, there is an increased risk of confounders being dissimilar between the control exposure and the intervention exposure (27).

2.3.4 Study length

It is often cited that a SW-CRT will take longer to perform than a P-CRT (12, 21, 23, 29, 32, 38, 46, 49). This has led to a SW-CRT being considered less feasible and desirable than a P-

CRT (49). It is likely that this is caused by the inherent requirement to have multiple observation periods. It is often argued that the length of one randomisation step in a SW-CRT is equivalent to the full P-CRT study (11, 38). However, it is clear that this is not always the case and often the P-CRT may need to be of longer length than a SW-CRT since it may require more participants to achieve the same power (35).

2.3.5 Data collection

A common misunderstanding in the literature is the difference between a cohort and cross-sectional SW-CRT (see section 1.4). If a SW-CRT follows a cohort design (participants are followed-up for the duration of the study), then a SW-CRT may create a burden for participants by collecting data at each time-period in which a new cluster receives the intervention (11, 29). If the outcome is not routinely collected then it could lead to a higher trial cost than other designs (29, 46). All SW-CRTs are usually unblinded at some level – since the clusters will be aware of the crossover from unexposed to exposed (38). This is also true at the participant level in cohort designs. However, blinding may still be able to be done at the participant level in a cross-sectional design.

2.4 Statistical considerations

In P-CRTs, the methodology required for power calculations and for the study analysis is well established – but is still developing for SW-CRTs (45). Currently, the focus of statistical and methodological research is on the cross-sectional design (23) despite this not being the most common design type (28). As such, many trials are using a sample size methodology that does not match the method of analysis (28). Below, we highlight the methodology used to estimate power in P-CRTs and SW-CRTs.

2.4.1 Analysis of a stepped-wedge cluster randomised trial

Whilst there are publications providing a framework for the analysis of SW-CRTs (9, 32) there has been limited research into alternative methods of analysis (31). In a SW-CRT, observations in the intervention arm are, on average, made at a later date than observations in the control arm and so the design could be biased by any secular changes (37), which should be adjusted for in any analysis (6, 23, 31, 39, 40). If there truly is a temporal trend, then failing to model the effect of time will produce a biased estimate of the treatment effect (6). An appropriate method of analysis has been presented by Hussey and Hughes (32) which is discussed further in Chapter 5, but an outline is presented here. Essentially, a generalised linear mixed model is recommended, with fixed time effects and random cluster effects, which can be given as:

$$Y_{ijk} = A + \beta_j + X_{ij}\delta + \alpha_i + \varepsilon_{ijk} \qquad 2.1$$

Where, Y_{ijk} is the outcome for participant *k* in cluster *i* at time *j*, *A* is the mean outcome in the unexposed group in the first time-period, β_j is a time effect, fixed for time-periods j = 2,…, T ($\beta_1 = 0$ for identifiability), δ is the treatment effect, α_i is a random effect for cluster *i*, ε_{ijk} is the residual error and X_{ij} is an indicator of treatment, where:

$$X_i \begin{cases} 1 & \text{if cluster } i \text{ is exposed to the intervention at time } j \\ 0 & \text{if cluster } i \text{ is not exposed to the intervention at time } j \end{cases}$$

Adaptations to this model for studies with incomplete designs or multiple layers of clustering have been presented (9). In cohort designs, short-term and long-term intervention effects can be examined by adjusting for the duration a participant has spent in the intervention arm (50).

2.4.2 Estimating sample size in a cluster randomised trial

2.4.2.1 *Sample size in a P-CRT with equal sized clusters*

When conducting a CRT, the sample size must be inflated over that required for an individually randomised controlled trial (RCT) to account for the non-independence of observations within a cluster (51-53), using a design effect (DE) (54, 55). For a P-CRT with equal cluster sizes, the DE is (56):

$$DE = 1 + (M - 1)\rho \qquad \textbf{2.2}$$

Where M is the cluster size and ρ is the ICC.

2.4.2.2 *Sample size in a P-CRT with unequal sized clusters*

However, CRTs often contain clusters of varying size, and this should be acknowledged within the sample size calculation (57). Using the above DE (equation 2.2) would provide an inaccurate estimate of the sample size. A number of DEs, appropriate under certain conditions, have been derived for P-CRTs with varying cluster size. The DEs to allow for varying cluster size can be dichotomised into two groups – those that require the size of each cluster to be known; and those that require an estimate of the mean and standard deviation of the cluster sizes.

Design effects that require size of each cluster to be known

Several DEs have been published that rely on the size of each cluster to be known prior to the study starting. These DEs weight the information that is contributed from each cluster, and the weights can be chosen in different ways.

Equal weight

When the DE weights the information that is contributed from each cluster equally, the DE is given as (58) :

$$DE = \rho\bar{M} + \frac{\bar{M}}{\overline{M_h}}(1-\rho) \qquad \textbf{2.3}$$

Where $\overline{M_h} = \frac{1}{\sum M_i}$ is the harmonic mean of the number of participants per cluster, and $\bar{M}$ is the mean cluster size. This DE is appropriate for a cluster level analysis with equal weights. The above DE is classified here as a DE that requires the cluster sizes since the harmonic mean may be difficult to estimate without the original cluster sizes, and since equation 2.3 is derived in the same manner as the below DEs.

Cluster size weight

When weighting by the cluster size: the clusters with more participants will be given more weight than smaller sized clusters. When weighting by cluster size, the DE is (58):

$$DE = 1 + \left(\frac{\sum {M_i}^2}{\sum M_i} - 1\right)\rho \qquad \textbf{2.4}$$

This DE is appropriate for a cluster level analysis weighted by cluster size.

Minimum variance weight

Alternatively, it is possible to use a weighting that is proportional to the inverse of the variance of the cluster mean, called the minimum variance weight. Here, the DE becomes:

$$DE = \frac{C\bar{M}}{\left(\sum_{i=1}^{C} \frac{M_i}{1 + (M_i - 1)\rho}\right)} \qquad \textbf{2.5}$$

This DE is appropriate for a cluster level analysis using a weighting proportional to the inverse of the variance of the cluster mean (58). It is also suitable for an approach using generalised estimating equation (GEE) with an exchangeable correlation structure (59).

Design effects that require mean cluster size and a measure of dispersion

Whilst useful, the above design effects (equations 2.3, 2.4, and 2.5) require knowledge of the exact cluster sizes. In many trials, it is unlikely that the exact cluster sizes will be known pre-trial. Instead, an a priori estimate of the coefficient of variation of cluster sizes (the ratio of the standard deviation of the cluster sizes to the mean cluster size) can be used to estimate the sample size without the need for exact cluster sizes. Below we highlight two DEs that allow for varying cluster sizes and only require the pre-specification of the mean cluster size and the coefficient of variation.

Design effect using cluster level analysis

By assuming a cluster level analysis, weighted by cluster size, Eldridge et al. (57) provide a design effect, given as:

$$DE = 1 + \left(\left[\frac{C-1}{C} cv^2 + 1\right] \bar{M} - 1\right) \rho \qquad \textbf{2.6}$$

Where C represents the number of clusters and cv is the coefficient of variation of cluster sizes.

For a trial with a large number of clusters, ignoring $\frac{C-1}{C}$, allows this design effect to be simplified to:

$$DE = 1 + ([cv^2 + 1]\bar{M} - 1)\rho \qquad \textbf{2.7}$$

The design effect given by equation 2.7 is for a cluster level analysis, providing a conservative approximation of the true design effect (57) since a mixed effect model at the individual level is more efficient (52). The true design effect for a P-CRT with unequal cluster size will be between equation 2.2 and 2.7 (57).

Taylor approximated design effect

A design effect has been proposed that is appropriate for an analysis using maximum likelihood estimation, as (60):

$$DE = \frac{1 + (M - 1)\rho}{1 - cv^2 \times R(1 - R)} \qquad \textbf{2.8}$$

Here R is the cluster mean correlation – which is the correlation between the cluster means of two repeated sets of observations taken from the same cluster (61) – which is discussed further in section 5.2.3.7.

Two other approximated design effects have been reported that use an equal weighting, and a cluster size weighting that both rely on the coefficient of variation (62). However, these are not reported here, since it has been established that the above DE (equation 2.8) is a more accurate representation of the true inflation factor. Equation 2.8 is also more appropriate than the commonly used DE presented by Eldridge et al. (equation 2.7) since equation 2.8

refers to an individual level analysis, rather than a cluster level analysis, which may be inefficient when the cluster sizes vary.

Overview of design effects for a P-CRT

When designing a P-CRT, many studies assume equal cluster size. When there is evidence of variability in the cluster sizes, failure to acknowledge this variability would lead to an over-estimate of the power. Studies should use a DE that includes an adjustment for varying cluster size. The selection of the DE will crucially depend on whether the size of each cluster is known in advance, or only a measure of the dispersion is known (52). Each DE is directly linked to a post-trial analysis method, and so only appropriate in some circumstances. CRTs are commonly analysed at the individual level through mixed effect models, and so the use of a DE appropriate for a cluster level analysis (such as equation 2.7) may produce a conservative estimate of the power. The most appropriate DEs are given by equations 2.5 and 2.8 as they are appropriate for individual level analysis, rather than cluster level analysis.

2.4.2.3 *Sample size in a SW-CRT*

Multiple methods exist to estimate the sample size in a SW-CRT (31) that stem from the Hussey and Hughes (32) framework described in section 2.4.1 – which is appropriate for cross-sectional designs. This framework involves a multilevel mixed-effect model with a random cluster effect and fixed time effect. An additional random effect for participants can be included for cohort designs (6). Since the estimation of sample size varies depending on study design, below we discuss the methods separately for cross-sectional and cohort designs.

A coherent unified framework has been set out for cross-sectional SW-CRTs for complete and incomplete designs, including multiple layers of clustering (9, 45). This has been implemented in a Stata function (63). However, there are also a number of DEs that are applicable for SW-CRTs based upon the design type – cross-sectional or cohort.

There has been much confusion surrounding the implementation of a design effect for cross-sectional SW-CRTs (45). For designs in which participants are observed once per step, and one baseline measurement period is made, a simplified version of the design effect can be given as a function of the ICC (ρ), the number of steps (s), the number of time-periods (T) and the cluster-period size (size of a cluster at one time-period) (m) as (21):

$$DE_{SW} = \frac{1+\rho(sm+m-1)}{1+\rho\left(\frac{1}{2}sm+m-1\right)} \times \frac{3(1-\rho)}{2\left(s-\frac{1}{s}\right)} \times T \qquad \mathbf{2.9}$$

To estimate the number of patients needed for a SW-CRT (SS_{SW}), the sample size should then be modified as follows:

$$SS_{SW} = SS_{RCT} \times DE_{SW}$$

Where SS_{RCT} is the sample size needed for an individually randomised trial.

In this design effect (equation 2.9), the correlation between observations within a cluster is independent of the time-period in which the observations are made, which is often assumed when following the Hussey & Hughes methodology for cross-sectional designs (32). However, recent research suggests a cluster by time interaction should be included (49), to allow for a different correlation structure over time (26) – see Box 2.1.

Box 2.1: Extended statistical model for stepped-wedge cluster randomised trials

$$\boldsymbol{Y_{ijk} = \mu + t_j + \alpha_i + \omega_{ij} + \varepsilon_{ijk}}$$

Where Y_{ijk} is the outcome for patient *k* in cluster *i* at time *j*, μ is the mean outcome, t_j is a fixed effect for time point t, $\alpha_i$is the cluster effect, ω_{ij} is a random effect for cluster *i* at time *j*, and $\varepsilon_{ijk}$is the residual error.

It is assumed that the following distributions apply:

$\alpha_i \sim N(0, \tau^2)$ $\omega_{ij} \sim N(0, {\sigma_t}^2)$ $\varepsilon_{ijk} \sim N\left(0, {\sigma_p}^2\right)$

This methodology allows for two different measures of clustering – within cluster, and between cross-sections within a cluster, and leads to an adapted design effect (26):

$$DE_{SW} = \frac{3s(1-\pi)(1+s\pi)}{(s-1)(2+s\pi)} \times (1+(m-1)\rho) \times T \qquad 2.10$$

Where $= \frac{m \times IPC}{1+(m-1)WPC}$, so that π is a function of the inter-period correlation (IPC) and the within-period correlation (WPC). Equation 2.10 simplifies to equation 2.9 if there is no time by cluster interaction.

A design effect for cohort SW-CRTs has been reported (26, 64) using the cluster (ρ_c) and subject (ρ_s) autocorrelation:

$$DE_{SW} = \frac{1+(m-1)\rho(s)(\rho_s + (\rho_c - \rho_s)\rho)}{1+(m-1)\rho\left(\frac{1}{2}s\right)(\rho_s + (\rho_c - \rho_s)\rho)} \times \frac{3\big(1+(m-1)\rho - (\rho_s + (\rho_c - \rho_s)\rho)\big)}{2\left(s - \frac{1}{s}\right)}$$

The DEs can be used to determine the sample size for a fixed power, and the DEs appropriate for a SW-CRT all follow a similar model framework, and so follow the same set of assumptions. Each design effect is only appropriate for studies with clusters of equal size (44) and it is assumed that the SW-CRT will have an identical number of clusters crossover at each step, since designs in which unequal number of clusters crossover at each step will lead to variations in the design effect (65). However, the Hussey and Hughes approach using matrices does not make this assumption, and can be used to estimate the power for a fixed sample size (32).

Irrespective of the methodology used to estimate the sample size, any misspecification of the time effect in the model (when a time effect is present) would falsely lead to an inflated power (44). A recommended alternative method is the simulation of expected trial data under the same underlying model used in the analysis, which can allow the power to be estimated for more complex designs – such as cohort and open cohort designs, count and binary outcomes, and studies with cluster specific variations in the intervention effect (44).

2.4.3 Design features that impact the power

2.4.3.1 *Effect of the correlation between observations on the power*

A defining feature in CRTs is the correlation between participants in a cluster – usually described by the intra-cluster correlation coefficient (ICC) – which must be incorporated into pre-trial power calculations. However, in longitudinal CRTs (CRTs with repeated cross-sections), the inter-period correlation (IPC) and within-period correlation (WPC) are also important.

The ICC is known to impact the power in SW-CRTs (6, 9, 21, 38, 66) and the decision of whether a P-CRT or a SW-CRT is more efficient, in terms of the number of participants or clusters, typically hinges on the ICC and the cluster sizes (45, 61). The power of a SW-CRT is mainly determined by the within-cluster variation (38, 66), and so is more efficient than a P-CRT for larger values of the ICC (37, 61). A set of cut off values of the ICC for which a SW-CRT would have greater power than a P-CRT has previously been presented (47) – in which the value is dependent on the size of clusters (45). The ICC follows a monotonic relationship with power for P-CRTs (power decreases as ICC increases), but exhibits a non-linear relationship with power for SW-CRTs (9).

Recently, it has been suggested that the power in a SW-CRT is influenced by values of the IPC and WPC (48) – which are describe in more depth in Chapter 4. If the correlation between participants from the same cluster at different periods is less than the correlation between participants from the same cluster at the same time-period, then ignoring the IPC and WPC would lead to an underpowered study (48). There has been no previous reporting of the IPC and WPC in the literature – which has led to a call for values to be published (26).

2.4.3.2 *Effect of the design structure on the power*

The power of a SW-CRT is influenced by the number of clusters, the cluster size, and the number of steps in the design (6) – which are often derived from logistical and financial considerations. A SW-CRT may require fewer participants than a P-CRT as the cluster sizes or number of steps increases (21), with the optimal power achieved in SW-CRTs in which each cluster is assigned a unique randomisation step (32). It has been claimed that a SW-CRT will always require fewer clusters than a P-CRT (47). In some settings, only a SW-CRT would meet the minimum power requirements for a study with few clusters (22).

2.4.3.3 *Varying cluster size in stepped-wedge cluster randomised trials*

The majority of methodological literature for the SW-CRT has focussed on the analytical framework given in section 2.4.1 – which is also used in this thesis. Whilst there has been slight variations to the framework reported in the literature (6, 26), there is a distinct lack of coverage of varying cluster sizes in all topical papers. Currently, the most readily accessed material for sample size calculations for SW-CRTs (21, 26, 63), do not allow for variability in cluster sizes. As such, authors often assume that the number of participants per time-period does not vary (6). A more comprehensive discussion of varying cluster size in SW-CRTs can be found in Chapter 5.

2.5 Summary

The methodological literature for the SW-CRT is slowly increasing, with additional design features being considered and examined. However, the literature is much sparser than for P-CRTs, which often provides difficulties when designing a trial. The lack of reporting guidelines for SW-CRTs has led to a poor quality of reporting in published SW-CRTs.

An increasing number of SW-CRTs are being conducted, typically in healthcare, and involving an education or training based intervention (28). Generally, the methodological focus has been on cross-sectional designs – which were originally the most common design type (23). More recently, it seems that studies follow a cohort or open cohort design (28). Whilst there has been some recent developments surrounding the methods for cohort SW-CRTs (26) many studies have incorrectly applied cross-sectional methods to cohort designs. The cross-sectional framework was originally proposed by Hussey and Hughes (32), and has been adapted for variations in the design (9). However, it is often assumed that there is no cluster by time interaction, and that the clusters are equal in size. The absence of a cluster by time interaction would indicate that the correlation within a cluster does not decay over time, but there has been little research into whether this is expected to be true. Within the literature, it is often assumed that the clusters in a study are equal size (44), and it is not known what impact cluster size variation would have on the power of a SW-CRT.

Previous systematic review have highlighted that the reporting of a sample size justification is often poor. However, it not been assessed as to whether the justified sample size calculation were well reported in terms of reproducibility or whether the sample size methodology included any additions to the Hussey and Hughes model given by equation 2.1 – such as varying cluster size, or the inclusion of the IPC and WPC.

In the following chapter, a methodological review of published SW-CRTs is performed, to assess the quality of reporting of sample size calculations in published SW-CRTs. This will include an assessment of the methodological quality of the sample size calculations. This will give an indication of whether the sample size calculation of previous SW-CRTs have reported an adjustment for varying cluster size or a decay in correlation over time (through the IPC and WPC).

CHAPTER 3: METHODOLOGICAL REVIEW OF THE QUALITY OF REPORTING OF STEPPED-WEDGE CLUSTER RANDOMISED TRIALS

A paper based on the work from this chapter has been published.
Citation: *Martin J, Taljaard M, Girling A, Hemming K. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. BMJ open. 2016;6(2):e010166.*

3.1 Introduction

3.1.1 Background

Previous systematic reviews have been conducted to examine past stepped-wedge cluster randomised trials (SW-CRTs) (23, 28, 29, 31). These reviews have considered the types of intervention used, the motivation for employing the stepped-wedge design, and the statistical methods used in the analysis (see Chapter 2). It has been highlighted that there is a large heterogeneity between methods used in the analysis of the SW-CRT (23, 29). Since recommendations regarding the analytical approach for SW-CRTs have been published (6, 32), it is likely that this variation will have reduced over time. However, further recommendations may need to be made to fully address the issue. Additionally, it has been noted that the sample size calculations differ greatly between studies (23, 28, 29). Often, the sample size methodology employed in the design stage is inappropriate, and frequently under reported (28, 29). The first systematic review of SW-CRTs found only 5 of 12 studies reported a sample size calculation (23). A second systematic review of SW-CRTs found that the sample size calculation was reported in 8 of 15 studies, with only 3 accounting for clustering (29). A more recent review highlighted that a greater number of SW-CRTs are reporting a sample size calculation but the methodology used differs (28). Within these reviews, some consideration has been made to the quality of reporting, with the lack of clarity and consistency often highlighted. However, no review has considered the quality of reporting in relation to published recommendations. Indeed, the reviews have only considered whether a sample size calculation was reported, and whether it accounted for clustering. Here, we will consider in more depth, the quality of reporting in published SW-CRTs and whether the methodology used is appropriate.

3.1.2 Reporting of sample size calculations

To allow for the critical appraisal of a study, the reporting of the study design, including the derivation of the sample size, and the results, should be transparent. Study reports that are difficult to comprehend or impossible to recreate may add less to the literature than they may otherwise have. To ensure that sufficient information was reported in trial reports, the consolidated standard for reporting (CONSORT) statement for individually randomised controlled trials was created (67). Clear reporting of clinical trials can allow for a critical appraisal and ensures that the results can be assessed for robustness (67).

In relation to the sample size calculation of trials, recommendations are given for the minimum number of items that should be reported to allow the sample size calculation to be reproducible (68, 69). It recommends that authors report: the level of significance; the power of the study; the estimated treatment effect; and whether an allowance was made for attrition or non-compliance. Following this, an extension to the CONSORT statement for cluster trials was published, which made additional recommendations for cluster trials (70). It recommends that authors of a cluster trial additionally report: the cluster size; a measure of its variation; an estimate of the clustering (e.g. ICC); and a measure of the uncertainty of the estimate of the clustering (70). In addition to this, the reporting of the method used to determine the sample size should be made. Without a clear specification of the methodology used, it may otherwise be difficult to comprehend whether an allowance has been made for clustering and for variation of cluster sizes. By fully explaining the methods used, the methodological rigor of the study can be assessed. Although there are currently no published reporting guidelines for a SW-CRT, an extension to the CONSORT statement for SW-CRTs is in development (71). Prior to this extension, several items have been recommended for reporting (6). However, even excluding these extension items, since the SW-CRT is a form of cluster randomised trials, the authors should report, as a minimum, items from the CONSORT statement and the CONSORT extension for cluster trials.

Perhaps somewhat understandably, early CRTs were often underpowered and analysed incorrectly (72). However, the literature surrounding these trial types has now advanced dramatically. Now, sample size methodology, and reporting guidelines, is well established for parallel cluster randomised trials (P-CRTs). However, recent systematic reviews relating to the reporting of items from the CONSORT statement in P-CRTs highlight that the quality of reporting is still inadequate (72-74).

Whilst it is expected that early SW-CRTs will not conform to the recommendations made in the proposed CONSORT extension for SW-CRTs, it should be expected that publications made post 2001 should adhere to the recommendations made in the original CONSORT statement (69). Similarly, studies published post 2010 should follow recommendations made in the CONSORT extension for CRTs (68). Previously little evidence existed on the quality of reporting, or the methodological rigor, of published SW-CRTs. Previous reviews of SW-CRTs have either been small (23, 29), or have considered other areas of interest (28, 31). Whilst they have identified whether studies have reported a sample size calculation and briefly considered the methods used, no previous review has considered the adherence of reporting of the sample size calculations, or indeed the methodological rigor of these calculations. Here, we seek to identify mistakes regularly made in the reporting of SW-CRTs in the literature, to prevent these errors from becoming common practice.

3.1.3 Methodological requirements in sample size calculations

In P-CRTs, it is well established that clustering should be accounted for in both the sample size calculations and the analysis (53, 75). It is recognised that failing to account for clustering in the sample size calculation can lead to an underestimate of the sample size required. Likewise, an analysis conducted without adjusting for clustering would produce an over precise estimate of the treatment effect, that is, the confidence interval would be too narrow. When conducting a P-CRT, it is conventional to utilise a design effect to obtain the sample required for a cluster trial from a sample size required under the assumption of individual randomisation (53, 57, 58, 60, 75, 76).

When conducting a SW-CRT, an integral part of the design is the randomisation of clusters over multiple time-periods (or steps). Since time is a potential confounder in the analysis of a SW-CRT, sample size calculations should account for time, in addition to the effects of clustering. Time should also be adjusted for in the analysis of a SW-CRT. In the topical paper by Hussey and Hughes, one method for estimating the power in a SW-CRT is discussed, in which allowances are made for both the effects of time and clustering (32).

3.1.4 Analytical methods used in cluster randomised trials

Upon conducting a CRT, the analytical approach used to evaluate the intervention should acknowledge the clustered nature of the data. Failing to do so will lead to an over-precise treatment effect estimate. A simple method to achieve this is to estimate an appropriate summary statistic for each cluster individually, before performing a statistical test on the pooled cluster summary measures. For example, the average BMI may be calculated for each cluster, before a linear regression is fitted to analyse the cluster level means. However, this cluster level analysis approach is often not appropriate in trials, since individual level covariates cannot be adjusted for. One potential approach for analysis at the individual level is a generalised linear mixed model (GLMM). This methodology allows for adjustment of individual level covariates, in addition to the effect of clustering.

Since in a SW-CRT, calendar time may impact the treatment effect, the analytical approaches used for a P-CRT may not be appropriate. In the seminal paper by Hussey and Hughes, they propose a methodology appropriate for the analysis of a SW-CRT (32). They recommend a mixed effects model with a random cluster effect and a fixed effect for time (see section 2.4.1). Conducting an analysis of a SW-CRT via this approach allows for both between cluster and within cluster information to be used to estimate the treatment effect. By following this approach, authors can allow for the effect of clustering and the confounding effect of time. In fact, failing to model the effect of time will result in a biased treatment effect estimate, unless there is truly no underlying temporal trend (6).

3.1.5 Chapter aim

The main aim of this chapter is to review the sample size calculation for published SW-CRTs to assess whether they are sufficiently reported and whether appropriate methodology has been used. As part of this, we seek to:

1. Determine adherence to reporting of nine sample size items recommended in the extension to the CONSORT statement for cluster trials.
2. Determine the reporting standard of additional items relevant to SW-CRTs.
3. Identify methodology used in the estimation of power or sample size and determine whether appropriate methodology is being used – such as whether an allowance has been made for clustering and time effects.
4. Identify methodology used in the analysis and determine whether appropriate methodology is being used – such as whether clustering and time effects have been accounted for.

To achieve these aims, a methodological review of SW-CRTs is presented, with an evaluation of the adherence to the CONSORT statement and the cluster extension. Additionally, an assessment was made of the methodological rigor of sample size methodology used. Finally, the methods used to analyse full studies was critically evaluated, noting whether studies had correctly accounted for clustering and time effects.

3.2 Methods

3.2.1 Search strategy and inclusion and exclusion criteria

The search strategy was an adaptation of the search strategy used in two previously published systematic reviews in SW-CRTs (23, 29). The phrases used to generate the search were:

- Stepped wedge
- Step wedge
- Experimentally staged introduction
- Delayed intervention
- One directional cross over (& crossover) design
- All possible permutations of the following terms:
    - Incremental, phased, staggered stepwise, step wise, delayed and recruitment, introduction, implementation

This search strategy is given in full in Appendix A and was conducted in October 2014.

Early systematic reviews of SW-CRTs considered randomised and non-randomised designs, or were limited to healthcare studies only (23, 29). Here, we included randomised studies only, but considered both healthcare and non-healthcare studies. To inflate our sample size, we consider both full trials and protocols. However, the analytical methods were extracted from full trials only, since the methodology reported in a protocol may be subject to change for the final analysis. As per our definition of a SW-CRT, the study had to be randomised, with the randomisation at the cluster level, and all clusters had to be randomised to receive the intervention over two or more steps.

The search strategy was conducted using Medline, Embase (including Embase classic) and PsycINFO and all identified studies were extracted into a database. The titles and abstracts were screened by two authors to ensure an agreement subject to the inclusion criteria. From this, full text articles were obtained for all studies that seemed eligible for the review. Studies that were subsequently deemed to be ineligible were excluded, with the reasons tabulated, with any disagreements between authors resolved via discussion. To ensure all possible studies were included, the reference list of studies that met the inclusion criteria were screened, as well as the reference list of any previous systematic reviews in SW-CRTs. Since our primary motive was to assess the quality of reporting of studies, no contact was made to the authors of any study to request any additional information.

Only fully published trials and protocols were included in the review. Any unpublished trials or protocols that were cited or related to an included study were not included or assessed for information. Since this review aims to assess the quality of reporting in relation to the CONSORT statements, it is expected that these items should be reported in the full trial reports, and so additional information should not need to be extracted from related trial protocols. Study protocols, for which a full report was not available or the study was not yet completed were included in some aspects of the review. Table 3.1 illustrates the categories that were used to group the excluded studies.

Table 3.1: An elaboration of the exclusion criteria groupings

Exclusion category	Explanation
Duplicate	If multiple copies of a study had been included, then one copy of it was included in the review, and all other versions were excluded.
Individually randomised design	All studies that the randomisation was conducted at the individual level, regardless of design type were excluded.
Not a SW-CRT	This included all cluster trials that were not a SW-CRT. This could have included studies in which the intervention was given to some clusters after the trial, or trials in which the intervention was removed during the study (e.g. Cross-over designs).
Not a trial	This included any articles that were not a trial report or a study protocol. This included conference abstracts, discussion of trials, and book chapters discussing trials.
Non-randomised designs	This included any studies that used the stepped-wedge design, but for whom the intervention was not given out in a random manner.
Secondary analysis:	Any secondary analysis of a SW-CRT that had already been included in the review were excluded.
Unable to access:	Any paper that could not be accessed was excluded.
Protocols of an included full trial report:	Full trial reports were included in the review, so the related protocols were not included.
Other:	This includes any article that did not fall into one of the above categories. This included process evaluations, reviews of a trial, and studies not published in English.

3.2.2 Data abstraction

A data abstraction form was used to extract data in a fair and unbiased manner. Following testing on a small number of studies, refinements were made, with the final form given in Appendix B. To ensure a high level of consistency across the review, data was extracted for each study by two independent reviewers. Any differences were listed, and a discussion between all authors allowed a consensus decision to be made. The order in which data was extracted from studies was produced by a random number generator. If the order is not randomised, there is the potential for bias in the reporting. The reviewer may adapt how the extraction form is seen for latter papers, altering their response in light of the experience they have gained in completing earlier extraction forms. Additionally, the reviewer may spot the reporting better at the beginning, but fail to spot them as fatigue sets in. Randomising the order should minimise bias. All data were abstracted and saved in Microsoft Access. Table 3.2 highlights the categories that have been used to group the extracted data.

Table 3.2: Categories used to group extracted data

Data extraction categories	Description
Trial demographics	This included basic trial demographics and characteristics to describe the SW-CRT, alongside realised design characteristics from full trial reports.
Design justification	This included the justification for using a cluster trial, alongside the justification for conducting a SW-CRT.
Quality of reporting of sample size elements	This includes the reporting of sample size elements from the 2010 CONSORT statement for individually randomised controlled trials; reporting of cluster sample size items from the 2012 CONSORT extension for CRTs; reporting of sample size elements that relate to a SW-CRT.
Methodological rigor of sample size calculations	This includes the methodology used to determine sample size.
Methodological rigor of analysis methods used	For full trial reports, data was also abstracted on the methods of analysis used.

3.2.3 Trial demographics

When extracting data on trial demographics, we consider both general trial demographics and realised design characteristics. General trial demographics were obtained for both complete trial reports and protocols. Realised design characteristics were extracted for full trial reports only.

General trial demographics included year of publication, country, journal impact factor, cluster type, healthcare or non-healthcare setting, number of interventions, randomisation process and primary outcome data type (Table 3.3). The journal impact factor was taken from the Web of Science, JCR Science Edition 2013.

Table 3.3: Brief description of extracted items on general trial demographics

Item	Explanation
Journal impact factor	Taken from Web of Science, JCR Science Edition 2013
Year	What was the year of publication of the report?
Country	What country did the study take place in?
Cluster	What was the cluster (i.e. hospital, ward, general practice, residential area, health professional)?
Study setting	Was the study conducted in a health care or non-health care setting?
Number of arms	How many interventions were being compared?
Method of randomisation	What was the method of randomisation used (i.e. unrestricted, paired, stratified)?
Primary outcome	What data type was the primary outcome? (i.e. continuous, binary, count, categorical)

For complete trials only, realised design characteristics were extracted. By this, general items related to CRTs were extracted, such as the duration of the study, number of clusters and average cluster size. Details of design features specific to the SW-CRT were also extracted, such as the number of randomisation points (or steps), the number of clusters randomised per step, the average step length, design type (cross-sectional, cohort, or open cohort) and whether any variations of a traditional SW-CRT were used (such as a transition period). The average step length was calculated as the median duration between two successive randomisation points, and so did not account for extra time-periods that may have been included pre or post intervention.

All values obtained related to an intention to treat analysis. As such, if a study began with 10 clusters, but one dropped out post randomisation, then it was noted that the trial had 10 clusters. Details were not obtained from trial protocols since the outcomes may have been subject to change, and so may not truly reflect the trial to be conducted. Whilst the protocols may indicate a planned number of steps, or cluster-period size, for example, factors may influence this during the trial and so these values may be adapted during the study. Because of this, only details pertaining to full trial reports were included when summarising design characteristics. A list of the items in which data was abstracted from, along with a brief description of them can be found in Table 3.4.

Table 3.4: Brief description of extracted items on realised design characteristics

Item	Explanation
No. of steps	If reported, how many randomisation steps were there?
Measurement periods	If reported how many measurement periods were there?
Clusters	If reported, how many clusters were there?
Clusters per step	If reported, how many clusters were randomised at each step?
Cluster size	If reported, what is the median cluster size across all measurement points?
Duration of study	If reported, how long did the study run for (not including outcome follow-up times or retrospective data collection)?
Duration of each step	If reported, how long what was the median duration of the time between each step?
Design type	If reported, did the study use repeated cross-sectional sampling, cohort design or open cohort design?
Transition periods	Did the design include transition periods during which the intervention is embedded into practice and the cluster considered neither exposed nor unexposed?
Variation on design	Did the design include any variation on the conventional SW-CRT, such as extended pre and post periods?

**Information was extracted on the realised design characteristics of full trial reports only and this might in some cases be different to that which was actually planned.*

3.2.4 Design justification

When conducting a clinical trial, it is important that the most appropriate trial type is used. As such, item 2a of the 2012 extension for cluster trial highlights that authors should detail why they conducted a cluster trial (70). For SW-CRTs, this is also essential. By establishing the rationale for a design choice, an assessment can be made as to whether any increase to the sample size is warranted. We extracted justification of both the sample size calculation, and the study design.

Firstly, information was gathered on the justification for the sample size of the study. This justification may stem from a statistical approach, in which a clinically important effect size was used to determine the sample size, or based on a pragmatic approach of using all available participants. To prevent confusion, a study was said to use a pragmatic rationale of the sample size if the authors clearly stated that the sample size calculation was limited by logistical or pragmatic constraints.

Now, information pertaining to the motivation of the trial type choice was extracted. Firstly, the rationale for choosing a cluster trial over an individually randomised trial was noted, and then the motivation for conducting a SW-CRT (Table 3.5). Since studies may provide generic statements of the advantages of a SW-CRT, only motivation relative to the study was included.

Table 3.5: Brief description of extracted items on design justification

Item	Explanation
Justification of sample size	Did the sample size stem from a statistical calculation or a pragmatic approach based on including all available participants.
Cluster trial rationale	What rationale was reported for using a cluster randomised trial over an individually randomised design? This could include: the possibility of contamination; practical reasons; or a cluster level intervention.
Stepped-wedge trial rationale	What rationale was reported for using a stepped-wedge cluster randomised trial over a parallel cluster randomised trial? This could include: desire for all clusters to receive the intervention; need for sequential implementation; prior evidence of effectiveness; or ethical issues.

3.2.5 Quality of reporting of sample size calculations

The reporting of sample size calculations allows other trialists or reviewers to understand the statistical methodology behind a study, and helps in the critical appraisal of a study. Transparent reporting ensures that the calculation is easily replicable. Data was extracted identically for both full trial reports and trial protocols. Primarily, the interest is surrounding the adherence to reporting of items relating to the CONSORT statement, and the CONSORT extension for cluster randomised trials. This is then extended to consider items recommended for reporting in SW-CRTs – though the absence of a CONSORT extension for SW-CRTs may lead to poor reporting.

In relation to the CONSORT statement, the following items should be reported:

i. Power of the study.

ii. Significance level used.

iii. Treatment effect sufficiently reported.

iv. Consistency between outcome in sample size calculation and primary outcome.

v. Whether attrition was allowed for.

A treatment effect was deemed to be sufficiently reported if any of the following criterions were reported:

a) Mean in both arms and standard deviation.

b) A mean difference and standard deviation.

c) Proportion in both arms.

d) Proportion in one arm and a relative (or absolute) difference.

e) Standardised effect size

For studies in which there was a lack of clarity surrounding the primary outcome, or else it was not clear what outcome had been used for the sample size calculation, then it was reported as unclear. If a full trial report utilised more clusters or participants than they had planned in the sample size calculation, but there was no explanation that this was due to attrition or non-compliance, they it was not reported as having allowed for attrition (Table 3.6).

Table 3.6: Sample size reporting explanation – assessment of quality of reporting of basic sample size elements

Item	Explanation
Sample size justification	Was a sample size justification reported?
Level of significance	Was the level of significance reported?
Power	Was the level of power reported?
Treatment effect	Was the treatment effect used in the sample size calculation sufficiently reported? Sufficient reporting of the treatment effect consists of either a standardised effect size; a mean difference and standard deviation; means in both arms and standard deviation; proportions in both arms; proportion in one arm and a relative or absolute difference.
Primary outcome	Was the outcome used in the sample size calculation consistent with the primary outcome of the trial?
Attrition	Was attrition allowed for in the sample size calculation?

In addition to these basic sample size elements, CRTs should be reporting items related to the CONSORT extension for cluster trials. These items pertain specifically to cluster trials, and include the reporting of:

i. The number of clusters

ii. The cluster size

iii. The variation of cluster sizes

iv. The ICC (or equivalent)

v. Measure of uncertainty surrounding estimate of ICC (or equivalent)

The number of clusters and cluster size was extracted and reported separately. However, when considering the number of items reported by a trial, they are combined, since for a given sample size, one item is deducible from the other (Table 3.7).

Table 3.7: Sample size reporting explanation – assessment of quality of reporting cluster sample size elements

Item	Explanation
Measure of variation of outcomes across clusters (i.e. ICC)	Was a measure of variation in outcomes across clusters reported? Measures of variation include the ICC, coefficient of variation (of the outcome) or a design effect.
Measure of uncertainty of measure of variation (i.e. ICC)	Was a measure of uncertainty of the ICC (or equivalent) reported; or was sensitivity to power considered under alternative ICCs?
Number of clusters	Was the number of clusters explicitly reported or deducible?
Cluster size	Was the cluster size explicitly reported or deducible?
Measure of variation of cluster sizes	Was a measure of variation in cluster sizes reported (or it clear that there was not variation), such as a coefficient of variation (of cluster size) or standard deviation of cluster sizes?

For SW-CRTs, it has been recommended that the following items are reported (6):

i. The number of steps

ii. The number of clusters randomised per step

iii. Design type

iv. Clarity between total cluster size and cluster size per measurement period

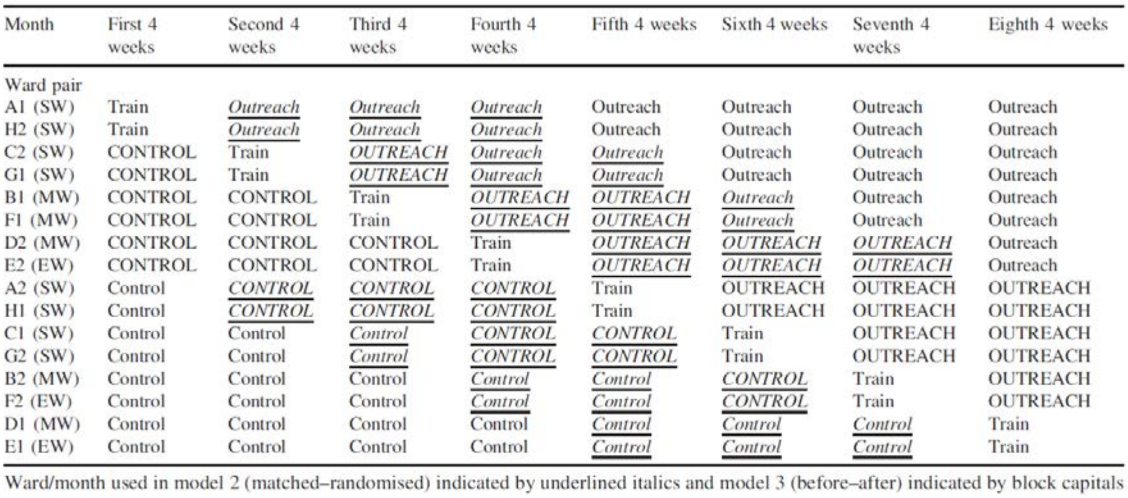
v. Schematic representation

Since the number of clusters randomised per step can vary, the authors should state the number randomised per step, to allow the sample size calculation to be replicable.

Authors should report the design type, that is, whether the design was cross-sectional, cohort or open cohort (section 1.3.1). The reporting of this item allows for clarity as to whether patients were followed-up for the study duration or for only part of it. This follows on to item iv. When reporting the sample size calculation, it is vital that there is clarity between the total cluster size and the cluster size per measurement period. Failure to clearly report this would increase the difficulty in replicating any sample size calculation.

The inclusion of a schematic representation may seem unintuitive in relation to a sample size calculation. However, it is a simple way to display some of the reporting items for a SW-CRT. For example, a diagram such as Figure 3.1, taken from the ICCOT study (section 1.6.2), clearly shows that the study has 16 clusters, with 2 clusters randomised per step and a total of 8 time-periods (24). This may help to decipher whether any additional methodological features should have been included in the sample size calculation, such as transition periods. As such, we include here whether a schematic representation was included in the trial report. Furthermore, data items that were not reported in the text but could be clearly identified from a schematic diagram of the study were classified as reported.

Figure 3.1: Example schematic representation of a SW-CRT. Original article by Priestley et al. (24)



Month	First 4 weeks	Second 4 weeks	Third 4 weeks	Fourth 4 weeks	Fifth 4 weeks	Sixth 4 weeks	Seventh 4 weeks	Eighth 4 weeks
Ward pair								
A1 (SW)	Train	*Outreach*	*Outreach*	*Outreach*	Outreach	Outreach	Outreach	Outreach
H2 (SW)	Train	*Outreach*	*Outreach*	*Outreach*	Outreach	Outreach	Outreach	Outreach
C2 (SW)	CONTROL	Train	*OUTREACH*	*Outreach*	*Outreach*	Outreach	Outreach	Outreach
G1 (SW)	CONTROL	Train	*OUTREACH*	*Outreach*	*Outreach*	Outreach	Outreach	Outreach
B1 (MW)	CONTROL	CONTROL	Train	*OUTREACH*	*OUTREACH*	*Outreach*	Outreach	Outreach
F1 (MW)	CONTROL	CONTROL	Train	*OUTREACH*	*OUTREACH*	*Outreach*	Outreach	Outreach
D2 (MW)	CONTROL	CONTROL	CONTROL	Train	*OUTREACH*	*OUTREACH*	*OUTREACH*	Outreach
E2 (EW)	CONTROL	CONTROL	CONTROL	Train	*OUTREACH*	*OUTREACH*	*OUTREACH*	Outreach
A2 (SW)	Control	*CONTROL*	*CONTROL*	*CONTROL*	Train	OUTREACH	OUTREACH	OUTREACH
H1 (SW)	Control	*CONTROL*	*CONTROL*	*CONTROL*	Train	OUTREACH	OUTREACH	OUTREACH
C1 (SW)	Control	Control	*Control*	*CONTROL*	*CONTROL*	Train	OUTREACH	OUTREACH
G2 (SW)	Control	Control	*Control*	*CONTROL*	*CONTROL*	Train	OUTREACH	OUTREACH
B2 (MW)	Control	Control	Control	*Control*	*Control*	*CONTROL*	Train	OUTREACH
F2 (EW)	Control	Control	Control	*Control*	*Control*	*CONTROL*	Train	OUTREACH
D1 (MW)	Control	Control	Control	Control	*Control*	*Control*	*Control*	Train
E1 (EW)	Control	Control	Control	Control	*Control*	*Control*	*Control*	Train

Ward/month used in model 2 (matched–randomised) indicated by underlined italics and model 3 (before–after) indicated by block capitals

A summary of the reporting items related to sample size elements for the SW-CRT are given in Table 3.8.

Table 3.8: Sample size reporting explanation – assessment of quality of reporting of stepped-wedge sample size elements

Item	Explanation
Number of steps	Was the number of randomisation steps reported?
Number of clusters per step	Was the number of clusters randomised per step reported?
Measurement points	Was the number of measurement points reported?
Clarity of cluster size	Was there clarity between cluster size per measurement point or period and total cluster size?
Design type	Was it clear whether the design was cross-sectional, cohort or open (i.e. mixture)?
Schematic illustration	Was the design represented using a schematic illustration?

To classify the studies using the number of items reported, only the adherence to the CONSORT statement and the cluster extension were used, so that there were 9 items relating to the quality of reporting. These were (i) the power, (ii) the significance level, (iii) treatment effect sufficiently reported, (iv) consistency between outcome in sample size calculation and primary outcome, (v) whether attrition was allowed for, (vi) the number of clusters (or cluster size), (vii) the variation of cluster sizes, (viii) the ICC (or equivalent), (ix) a measure of uncertainty surrounding the ICC.

All items were categorised as reported, not reported, or unclear. Items were reported as unclear if there was a lack of distinctness in the reporting of said item, or if reviewers could not fully agree. For some items, reporting was sub-branched into explicitly reported or deducible. Items were classified as explicitly reported if they were stated clearly in the text, whilst classified as deducible if they could be obtained from calculations via other reporting items or via intuition from other items. For some reporting items, it may be enough that they can be deduced, such as cluster size, whereas for other items, it is more important that they are explicitly reported, such as the design type.

3.2.6 Methodological rigor of sample size calculations

There are published guidelines for an appropriate method to estimate the sample size for a SW-CRT (9, 32, 63). Previous reviews have shown that the methodology used in previous SW-CRTs varies greatly. As such, it is vital to highlight the number of studies who are using correct methods, and highlight the areas in which improvements should be made.

To establish whether the methodology used in a sample size calculation had allowed for cluster and time effects, the underpinning information was extracted. Adjustments for time should be made to allow for any underlying secular trend, which may influence the outcome, irrespective of whether the observations is from the exposed or unexposed period. Adjustments for time effects are likely to use the methodology set by Hussey and Hughes (32), else perhaps the design effect provided by Woertman et al. (21). Since the design effect proposed by Woertman et al. requires an inflation factor to account for the nature of the design, it is possible that studies may have been powered with or without this correction. The recent DE for cohort SW-CRTs (26, 64) was published after this review, and so no studies would have used it.

Additional design features may be present in a SW-CRT. Depending on the intervention in question, it is possible that a transition period may be required, which should be included in the power calculation. Also, there is potential for an interaction between the intervention effect and time, which is noted as a time by treatment interaction. By this, it is possible for the intervention effect to be influenced by the length of time since it was implemented. Most importantly, it is possible for the within-cluster correlation to differ between measurement periods – that is, the correlation can decay over time. This is referred to as an inter-period correlation. It was noted as to whether any additional design features had been allowed for in the power calculation. Allowing for these items would be a deviance from the norm, so it is expected that allowing for such items would be clearly reported. As such, if they were not mentioned, it is assumed that an allowance for them did not take place.

Information was extracted as to whether any allowance was made for varying cluster size. However, it was not recorded as to what methods were used to account for this variability. For both cohort and open cohort study designs, it was reported as to whether an allowance had been made to account for the repeated measurements of the same participant. A full list of the methodological assessment elements, accompanied by a short narrative describing the item can be found in Table 3.9.

Table 3.9: Sample size reporting information - methodological assessment elements

Item	Explanation
Method used to compute power	Was the methodology used to determine the power reported? If so, was this using a parallel cluster method; Hussey and Hughes; Woertman; simulation methods (and did this allow for time effects); or was no allowance for clustering or time effects made?
Powered design variations	If the trial included a variation on the typical design, such as transition periods, or extra pre and post measurement periods, or repeated measures on the same individual, or varying cluster sizes, were these allowed for in the power calculation?
Powered interactions	Was the trial powered for any interactions, such as for example and interaction between calendar time and the treatment effect; or an interaction between the time since exposure and the intervention (e.g. a treatment lag)?
Extended correlations	Was the trial powered for an allowance for any extended correlation structures, such as a different correlation between observations within a cluster over time?
Method used to describe between cluster heterogeneity?	What method was used to describe the between cluster heterogeneity (i.e. ICC, coefficient of variation or design effect)?

3.2.7 Methodological rigor of analysis methods used

For trialists employing a SW-CRT, there are only a limited number of published articles that offer advice for the analytical methods that should be used. Whilst the seminal paper by Hussey and Hughes (32) and a recent paper by Hemming et al. (6) offer some statistical guidance, there has been limited work comparing potential methods of analysis. As such, it is likely that the methodology utilised in past trials will vary greatly.

In order for the methodology to be assessed, it is first vital to highlight the necessary items that should be reported in the analysis. As a minimum, all analysis of a SW-CRT should allow for clustering. This may be via mixed effect model, robust standard errors, or by including fixed cluster effects, for example. It is therefore of interest to indicate the framework that was utilised. It is also necessary for the analysis of a SW-CRT to include the additional effects of time (32). Failing to account for time effects could produce misleading results and lead to an over-estimate of the precision of the treatment effect. Due to the complex design of the SW-CRT, it is possible that other additional features may have been included in the model. This may include a time by treatment effect interaction term. Since a large proportion of the included studies were cohort or open-cohort, the analysis should include an allowance for repeated measurements. Failing to adjust for this would produce an over-precise treatment effect estimate.

An overview of the items relating to the methods of analysis and a short descriptive passage can be found in Table 3.10.

Table 3.10: Analytical methods - item description

Item	Explanation
Allowance for clustering	Was an allowance made for clustering?
Model framework	What model choice did they use for the analysis?
Allowance for time effects	Did the analysis make allowance for time effects?
Framework for time effects	What framework was used to allow for time effects
Interaction terms	Were any interaction terms included in the primary analysis model?
Extended clustering	Were any extended correlations added to allow for correlation within a time-period and between time-periods?
Allowances for repeated measurements	Were allowances made for repeated measurements of individuals?

3.2.8 Impact of CONSORT statement and CONSORT cluster extension

The primary analysis consists of the reporting of items only. However, a secondary interest is the impact of the CONSORT 2012 extension for cluster trials on the quality of reporting in SW-CRTs. Since the CONSORT extension to cluster trials was published in 2012, it is expected that reporting of items should have improved after this. To allow for the bedding in of this publication, and to allow for publications that may have been made in the same year prior to the publication of the cluster extension, studies will be dichotomised using 2013 as the cut-off point, and we compare publications made prior to 2013 to those published in 2013 or later.

3.2.9 Methods used for analysis of included studies

When presenting results, each item of interest with respect to the quality of reporting is given first as a number and percentage. This will simply indicate the number of studies that have reported the item, along with the accompanying percentage of the total studies. For continuous outcomes (such as journal impact factor), a median and inter quartile range is given.

When considering the difference in reporting between subgroups, the differences were described using absolute difference and a corresponding 95% confidence interval. Although convention may dictate that only a p-value is chronicled to highlight whether there is a significant difference, the CONSORT statement recommends that a more clinically meaningful interpretation is included (68, 69). For dichotomous outcomes, we report a confidence interval for the mean difference, which is given as:

$$P_1 - P_2 \pm Z_{1-\alpha/2} \times \sqrt{\frac{P_1(1-P_1)}{N_1} + \frac{P_2(1-P_2)}{N_2}}$$

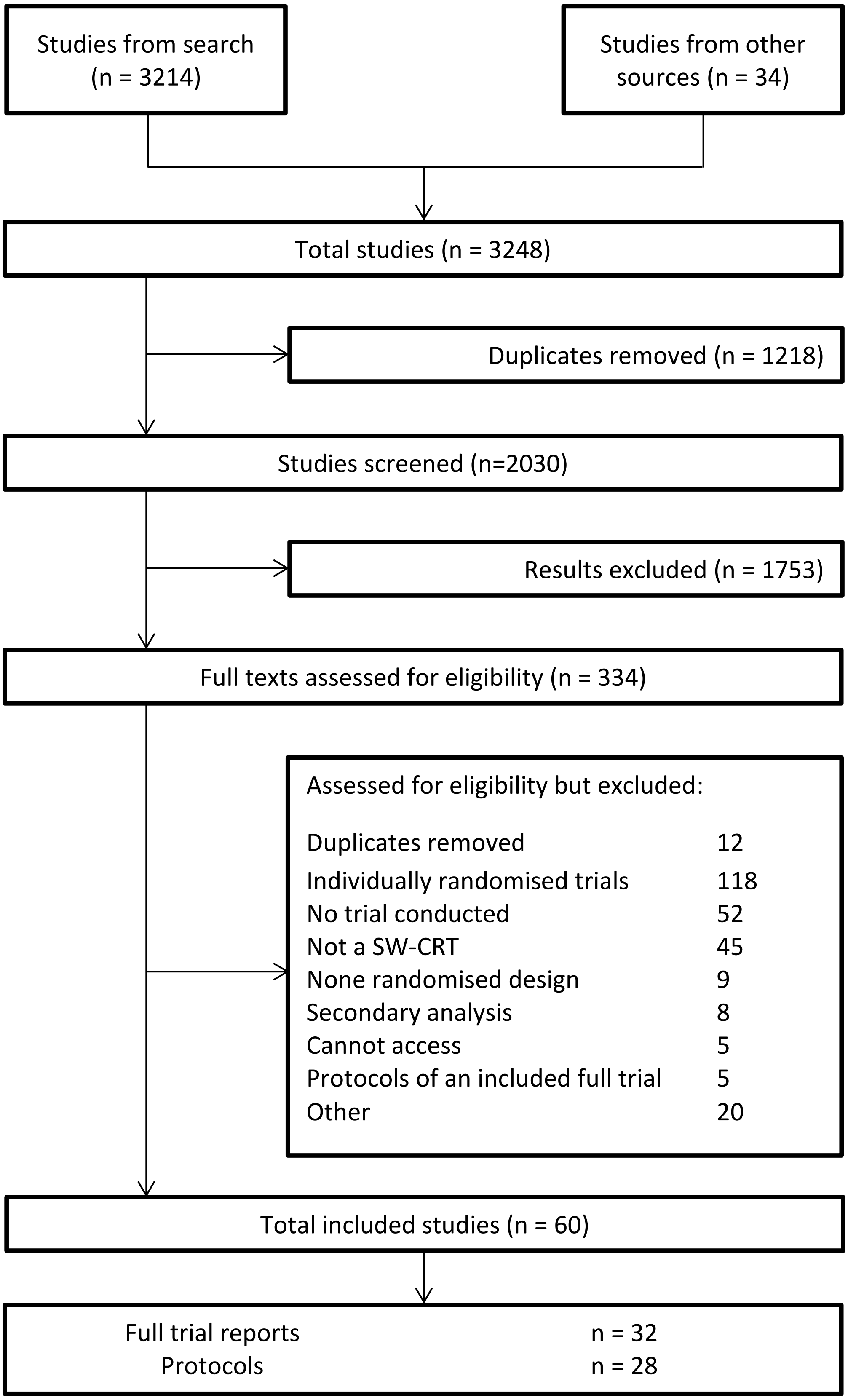
Where P_1 and P_2 are the probabilities of the outcome in groups 1 and 2, and N_1 and N_2 represent the number of observations within groups 1 and 2. For dichotomous outcomes with sufficient number of observations in each group, a chi-squared test was used to produce a p-value, whilst Fishers exact test was used for outcomes with a small number of observations. A small number of observations were deemed as the expected number of outcomes less than 5. For all dichotomous outcomes with a small number of observations, exact confidence intervals were calculated and reported.

Continuous outcomes were compared using a t-test and confidence intervals were formed assuming normality. The differences were also tested using a Mann-Whitney U test for continuous data in which the data seemed to be non-normal. In order to test whether there has been an increase in the quality of reporting over time, for some items, a linear test for trend was conducted. This is simply the fitting of a logistic model, in which the outcome variable is the reporting item in question. The year of publication is then added as an independent variable. A positive parameter value that corresponds to the year of publication will typically indicate that the reporting is, on average, increasing over time. Likewise, a negative parameter will correspond to a decrease in reporting over time.

3.3 Results

The search strategy (section 3.2.1) found 3214 studies. After de-duplication, 1996 studies were screened leaving 300 items for assessment. During this assessment, an additional 34 items were included. These papers were identified from other sources, such as the reference list of relevant papers. As such, there were 334 full text articles assessed for eligibility (Figure 3.2). The full texts were screened by two independent reviewers, to identify whether they were suitable for inclusion. Any disagreements were discussed, before an agreement made. 274 studies were excluded, with the reasons categorised into the following types: duplicate, individually randomised study, not a SW-CRT, not a trial, non-randomised designs, secondary analysis, unable to access, protocols of included full trial reports, and other. There were 60 studies left for inclusion in the review. A flowchart of the studies from the search strategy is given in Figure 3.2.

Figure 3.2: Flow chart of studies


Studies from search
(n = 3214)
Studies from other sources (n = 34)
Total studies (n = 3248)
Duplicates removed (n = 1218)
Studies screened (n=2030)
Results excluded (n = 1753)
Full texts assessed for eligibility (n = 334)
Assessed for eligibility but excluded:
Duplicates removed 12
Individually randomised trials 118
No trial conducted 52
Not a SW-CRT 45
None randomised design 9
Secondary analysis 8
Cannot access 5
Protocols of an included full trial 5
Other 20
Total included studies (n = 60)
Full trial reports n = 32
Protocols n = 28

3.3.1 Trial demographics

3.3.1.1 *General trial demographics*

The trial demographics of all included studies are summarised in Table 3.11. These have been dichotomised via report status into full trial reports and trial protocols. Of the 60 studies included in the review, 32 were full trial reports and 28 were study protocols. Over half of the included studies were published during or after 2013. The majority of trials were conducted in high income countries, with Australia, the British Isles, and North America the most common countries conducting a SW-CRT. Whilst only one full trial had been conducted in Australia, there were numerous protocols detailing planned trials, with over 20% of published protocols relating to planned SW-CRTs in Australia. Most of the included studies (83.3%) were carried out in a healthcare environment. Non-healthcare settings included an intervention to reduce absenteeism in the workplace (77), an evaluation of the effect of free school meals on academic achievement (78) and an intervention designed to reduce risk of dyslexia (79). Only 7% of the studies compared 3 or more trial arms, indicating that the majority of studies compared only two trial arms. Typically this included the testing of an intervention versus control or standard care. Over half of the studies utilised a simple, unrestricted form of randomisation. Stratification was carried out in 14 (23%) of studies, but only in 4 (13%) of the included full trial reports. 15 (25%) of studies used a continuous outcome, with binary outcomes being much more common (57%). Only 5 (16%) of the full trial reports had an accompanying trial protocol.

Table 3.11: Trial demographics of included SW-CRTs

	Total N = 60	Protocols N = 28	Full reports N = 32
Year of publication			
1987-2012	28 (47)	12 (43)	16 (50)
2013-2014	32 (53)	16 (57)	16 (50)
Journal Impact Factor			
Median [IQR]	2.6 [2.0 - 3.5]	2.3 [2.1 - 4.8]	3.3 [2.0 - 4.8]
Country of study			
Australia	7 (12)	6 (21)	1 (3)
Canada or US	15 (25)	4 (14)	11 (34)
UK or Ireland	11 (18)	3 (11)	8 (25)
Other higher income country	15 (25)	9 (32)	6 (19)
Middle income country	9 (15)	4 (14)	5 (16)
Low income country	3 (5)	2 (7)	1 (3)
Type of setting			
Health-care	50 (83)	25 (89)	25 (78)
Non health-care	10 (17)	3 (11)	7 (22)
Cluster			
General practice	7 (12)	6 (21)	1 (3)
Hospital/Ward/Specialities	12 (20)	5 (18)	7 (22)
Other health cluster	20 (33)	9 (32)	11 (34)
Geographical unit	11 (18)	5 (18)	6 (19)
Other / Unclear	10 (17)	3 (11)	7 (22)
Number of study arms			
Two	56 (93)	25 (89)	31 (97)
Three or more	4 (7)	3 (11)	1 (3)
Randomisation type			
Simple	35 (58)	15 (54)	20 (63)
Paired	4 (7)	0 (0)	4 (13)
Stratified	14 (23)	10 (36)	4 (13)
Other / Unclear	7 (12)	3 (11)	4 (13)
Primary Outcome type			
Continuous	15 (25)	10 (36)	5 (16)
Binary	34 (57)	13 (46)	21 (66)
Other	5 (8)	2 (7)	3 (9)
Unclear/not reported	6 (10)	3 (11)	3 (9)
Published protocol		N/A	5 (16)

Values are numbers (percentages) unless stated otherwise. IQR: Inter Quartile Range.

3.3.1.2 *Realised design characteristics*

The realised design characteristics of the full trial reports are summarised in Table 3.12. Clusters were randomised over two time-periods in 9 (28.1%) of the included studies. The median number of randomisation points was 4 (IQR 2 – 6). Over a quarter of the studies (28.1%) contained less than ten clusters. The median number of clusters included was 17 (IQR 8 – 38). The median cluster size was 21.5 (IQR 15 – 30) for cohort studies, 288.5 (IQR 43.8 – 493) for open cohort studies, and 326 (IQR 182 – 500) for cross-sectional studies. The median duration of the included studies was 17 months (IQR 8 – 24), with a median step length of 2 months (IQR 1 – 4). Only 5 (15.6%) of the completed trial reported contained a design that was cross-sectional in nature. In contrast, 12 (37.5%) and 10 (31.3%) of the designs were cohort or open cohort. Whilst not reported here, the design type was recorded for protocols also. Of the protocols, 11 (39.3%) were planned to be cross-sectional in nature, 5 (17.9%) cohort studies, 7 (25%) open cohort designs and 5 (17.9%) did not report the planned design type. Of the full trial reports, 14 (43.8%) included a variation on a traditional SW-CRT. The most common variation was an extension to the pre or post periods of the study whilst all clusters remained unexposed or exposed (10 studies). Of the 6 studies in the “other” category, there were 3 trial reports in which not all clusters were exposed to the intervention at the end of study (though the intention was for them to receive the intervention). An additional 3 studies contained no period in which all clusters were unexposed to the intervention.

Table 3.12: Summary of the realised design characteristics of included full trial reports

	Full trial report N = 32
Number of steps[1]	
Two	9 (28.1)
Three or four	8 (25.0)
More than four	14 (43.8)
Not reported	1 (3.1)
Median [IQR]	4.0 [2.0 - 6.0]
Number of clusters	
Less than ten	9 (28.1)
Ten or more	22 (68.8)
Not reported	1 (3.1)
Median [IQR]	17.0 [8.0 - 38.0]
Total cluster size[2]	
Median [IQR]	55.0 [24.0 - 326.0]
Number of clusters randomised per step	
Median [IQR]	3.0 [1.0 - 8.0]
Number of measurement points[3]	
Median [IQR]	5.0 [3.0 - 7.5]
Study duration (months)	
Median [IQR]	17.0 [8.0 - 24.0]
Step duration (months)	
Median [IQR]	2.0 [1.0 - 4.0]
Design type[4]	
Cross-sectional	5 (15.6)
Cohort	12 (37.5)
Open cohort	10 (31.3)
Unclear	5 (15.6)
Variations on design	
Transition periods	1 (3.1)
Extended pre or post periods	11 (34.4)
Other	6 (18.8)

Values are numbers (percentages) unless stated otherwise. [1]:steps are points at which clusters are randomised; [2]: for cohort studies this is the total number of observations made within the cluster, it includes the size of clusters in which there was lack of clarity of cluster size and cluster size per measurement period but for which a judgement was made; [3]:measurement points are the number of separate periods or points in time in which outcome data are collected; [4]:Design type includes those for which there was lack of clarity but for which a judgement was made; IQR: Inter Quartile Range

3.3.2 Design justification

Over 75% of the included studies incorporated a justification of their sample size. Only a small number of studies (5 trials, 8%) were identified that used a pragmatic argument to justify sample size, with the majority of studies (41, 68%) using statistical reasoning (Table 3.13). The majority of studies (72%) did not divulge the rationale for randomisation at the cluster level over the individual level. Of the 17 studies that reported the reasons for using a cluster randomised design, the most common motives were to avoid contamination and because of a cluster level intervention. The reporting of rationale for choosing a SW-CRT was better than simply the rationale for a cluster design. Only 18 (30%) studies provided no justification of design choice. The most common justification provided was the need for staggered implementation (37%). Other common justifications included the intention (or desire) for all clusters to receive the intervention (28%), and to mitigate ethical concerns that may arise from withholding an intervention (18%). Examples of rationale contained in the other category include the evaluation of a routine roll-out, operational simplicity and the requirement of fewer clusters than a P-CRT.

Table 3.13: Summary of the design justification for included SW-CRTs

Values are numbers (percentages) unless stated otherwise. P-value is for the comparison of full trials to protocols using a chi-squared test for proportions or (*) using

	All reports N = 60	Protocols N = 28	Full reports N = 32	Absolute difference (95% confidence interval	P-value
Sample size justification					
Pragmatic*	5 (8)	3 (11)	2 (6)	4.5 (-9.7 to 18.7)	0.657
Statistical	41 (68)	24 (86)	17 (53)	32.6 (11.0 to 54.2)	0.007
No justification*	14 (23)	1 (4)	13 (41)	-37.1 (-55.4 to -18.7)	0.001
Motivation for cluster randomisation					
Contamination*	9 (15)	6 (21)	3 (9)	12.1 (-6.2 to 30.3)	0.281
Cluster level intervention*	9 (15)	5 (18)	4 (13)	5.4 (-12.9 to 23.6)	0.721
Practical reasons*	3 (5)	3 (11)	0 (0)	10.7 (-0.7 to 22.2)	0.096
Other *	1 (2)	0 (0)	1 (3)	-3.1 (-9.2 to 2.9)	1.000
None reported	43 (72)	17 (61)	26 (81)	-20.5 (-43.1 to 2.1)	0.078
Motivation for stepped-wedge design					
Ethical concerns*	11 (18)	5 (18)	6 (19)	-0.9 (-20.5 to 18.7)	1.000
Sequential implementation	22 (37)	12 (43)	10 (31)	11.6 (-12.8 to 36.0)	0.352
Social acceptability*	5 (8)	4 (14)	1 (3)	11.2 (-3.1 to 25.5)	0.175
Resource Constraints*	9 (15)	6 (21)	3 (9)	12.1 (-6.2 to 30.3)	0.281
Methodological reasons*	6 (10)	6 (21)	0 (0)	21.4 (6.2 to 36.6)	0.008
Clusters act as own control*	7 (12)	4 (14)	3 (9)	4.9 (-11.5 to 21.3)	0.695
Adjust for temporal trends*	5 (8)	3 (11)	2 (6)	4.5 (-9.7 to 18.7)	0.657
Desire for all clusters to receive intervention	17 (28)	10 (36)	7 (22)	13.8 (-9.0 to 36.6)	0.235
Expect to do more good than harm*	10 (17)	6 (21)	4 (13)	8.9 (-10.1 to 28.0)	0.491
Other*	11 (18)	4 (14)	7 (22)	-7.6 (-26.9 to 11.7)	0.519
None reported	18 (30)	6 (21)	12 (38)	-16.1 (-38.7 to 6.6)	0.259

Fisher's exact test.

3.3.3 Quality of reporting of sample size calculation

3.3.3.1 *Quality of reporting of items recommended by CONSORT*

A comparison of the adherence to the reporting of the 9 CONSORT items for trials published in 2013 and 2014 compared to those published prior to 2013 is given in Table 3.14. Of the 60 studies included in the review, 45 (75%) reported a sample size justification.

Generally, the reporting has improved over time, with an increase in the sample size justification (p = 0.073) and recent studies reporting on average 1.22 items more (95%CI: 0.07 to 2.36) than those published pre-2013. On average, studies were reporting 5 of the 9 CONSORT items [IQR: 2 to 6], though no study reported all 9. Assuming normality, a t-test of group means would correspond to a significant difference (p=0.037) but a Mann-Whitney U-test assuming non-normality leads to a non-significant result at the 5% level (p=0.067). As expected, almost all studies (97%) reported the number of clusters in the study. The expected power of the trial was reported in 45 (75%) of the studies, but the treatment effect was only sufficiently reported in 33 (55%). Almost half (55%) described the variation in outcomes across clusters, but few reported the uncertainty surrounding this value. An allowance for attrition was reported in only 30% of studies. Notably, the reporting of clustering (often by the ICC) has improved from 29% pre 2012 to 69% post 2013 (p = 0.022).

A linear test for trend, to examine whether the reporting of each item has improved over time, was fitted to the data. This highlighted that as time has progressed, the likelihood of reporting the variation in outcomes across cluster has increased (p = 0.025). For all other reporting items, whilst it showed a positive trend, there were no other statistically significant results at the 5% level.

Table 3.14: Sample size reporting quality from CONSORT statements

Values are numbers (percentages) unless stated otherwise. 1: a sufficient reporting of the treatment effect consists of either a standardised effect size; a mean difference and SD; means in both arms and SD; proportions in both arms; proportion in one arm and a difference. IQR: Inter-quartile range. ICC: Intra Cluster Correlation. P-value is for the comparison of 1987-2012 publications and 2013-2014 publications using a λ^2 test for proportion or Mann-Whitney U test, or (*) using Fisher's exact test.

	All studies	1987-2012	2013-2014	Absolute difference (95%	P-value
	N = 60	N = 28	N = 32	confidence interval)	
Sample size justification					
Reported	45 (75)	18 (64)	27 (84)	20.1 (-1.7 - 41.8)	0.073
ITEM 1:					
Level of significance	39 (65)	16 (57)	23 (72)	14.7 (-9.3 - 38.8)	0.233
ITEM 2:					
Power	45 (75)	18 (64)	27 (84)	20.1 (-1.7 - 41.8)	0.073
ITEM 3:					
Treatment effect [1]	33 (55)	15 (54)	18 (56)	2.7 (-22.6 - 27.9)	0.835
ITEM 4:					
Consistency with primary outcome	38 (63)	14 (50)	24 (75)	25.0 (1.2 - 48.8)	0.045
ITEM 5:					
Allowance for attrition	18 (30)	7 (25)	11 (34)	9.4 (-13.6 - 32.4)	0.429
ITEM 6:					
Number of clusters	58 (97)	27 (96)	31 (97)	3.4 (-67.0 - 73.9)	0.923
Median cluster size	39 (65)	15 (54)	24 (75)	23.4 (-2.3 - 49.2)	0.083
ITEM 7:					
Variation in cluster size*	6 (10)	1 (4)	5 (16)	12.1 (-2.3 - 26.4)	0.201
ITEM 8					
Variation in outcomes across clusters (i.e. ICC)	33 (55)	11 (39)	22 (69)	29.5 (5.3 - 53.7)	0.022
ITEM 9:					
Uncertainty of ICC (or equivalent)*	8 (13)	3 (11)	5 (16)	4.9 (-12.1 - 21.9)	0.712
All ITEMS					
Number items reported Median [IQR]	5 [2 – 6]	4 [1 – 6]	6 [5 – 6]	1.22 (0.07 - 2.36)	0.067
Reporting all 9 items	0 (0)	0 (0)	0 (0)		

3.3.3.2 *Quality of reporting of items relevant to a SW-CRT*

The reporting of items relevant to the SW-CRT by year of publication is given below (Table 3.15). The number of steps was well reported, either explicitly (90%) or deducible (98%), and almost all trials reported the number of clusters randomised per step (93%). Over 75% of studies included a schematic representation of the design. Although it was possible to deduce the design type in 72% of studies, only 27% explicitly reported the design. In almost 50% of studies, it was unclear whether the cluster size reported in the sample size calculation indicated the total cluster size or the cluster size per measurement period.

Although there is some indication of an increase in the quality of reporting over time, it is not substantially different between the groups. When fitting a linear test for trend, though on average it seems that there is an increase in reporting over time, there were no statistically significant results.

There is some evidence of a difference in sample size reporting of the 9 CONSORT items between full trials and protocols (Appendix C), with protocols generally reporting a greater number of items. However, there is no evidence of a difference in the reporting of SW-CRT items.

Table 3.15: Reporting of SW-CRT sample size elements

	All studies N = 60	1987-2012 N = 28	2013-2014 N = 32	Absolute difference (95% confidence interval)	P-value
Number of steps					
Explicitly reported	54 (90)	23 (82)	31 (97)	14.7 (-0.7 - 30.1)	0.058
Reported or deducible	59 (98)	27 (96)	32 (100)	3.6 (-3.3 - 10.4)	0.281
Number clusters randomised per step					
Reported	56 (93)	25 (89)	31 (97)	7.6 (-5.4 - 20.5)	0.240
Schematic representation					
Reported	46 (77)	20 (71)	26 (81)	9.8 (-11.7 - 31.3)	0.370
Design type (i.e. cross-sectional/cohort)					
Explicitly reported	16 (27)	6 (21)	10 (31)	9.8 (-12.3 - 31.9)	0.391
Reported or deducible	43 (72)	19 (68)	24 (75)	7.1 (-15.8 - 30.0)	0.540
Clarity of cluster size[1]					
Total cluster size	17(28)	8 (29)	9 (28)	-0.4 (-23.3 - 22.4)	0.969
Cluster size per measurement period	25 (42)	10 (36)	15 (47)	11.2 (-13.6 - 35.9)	0.382
Unclear/not reported	28 (47)	15 (54)	14 (44)	-9.8 (-35.1 - 15.4)	0.448

Values are numbers (percentages) unless stated otherwise. [1]*: some studies reported both total cluster size and cluster size per measurement period; P-value is for the comparison of 1987-2012 publications and 2013-2014 publications using a chi-squared test for proportions.*

3.3.4 Methodological rigor of sample size calculations

When considering the studies published pre 2012 to those published in 2013 & 2014, it seems that the methodology is improving (Table 3.16). A large majority of studies (73%) allowed for clustering, though only one-third allowed for time effects. However, the proportion seems to have increased over time, with a linear test for trend showing some evidence of this. Of the 45 included studies, 24 should have allowed for repeated measures, though only 3 (13%) acknowledged this in the sample size calculation. Whilst this seems to have decreased over time, there are too few studies to make a conclusive judgement.

The reporting of the methodology used is quite poor, with one-third of studies not reporting the method used to estimate the sample size. There has been in an increase over time in the number of studies using the Hussey and Hughes methodology (17% to 41%). When treating time as continuous, this corresponds to a p-value of 0.074. Pre 2012, the majority (83%) of studies did not account for time effects. Whilst this proportion has decreased over time (p=0.004), there are still a large proportion (56%) who have not accounted for time effects in the sample size calculation, with methodology appropriate for a P-CRT or a before and after CRT often used.

Whilst it is not surprising that such few studies acknowledger additional design features in the sample size calculation, it should be noted that several studies included a transition period in the schematic representation of the study, but did not account for this in the sample size calculation.

The variation in the outcome across clusters was commonly described using the ICC (20/33) and CV (10/33), with the proportion using the CV increasing in recent studies.

Table 3.16: Methodological assessment of sample size calculations

	All reports N = 45	1987-2012 N = 18	2013-2014 N = 27	Absolute difference (95% confidence interval)	P-value
Allowance for clustering					
Number (%)	33 (73)	11 (61)	22 (81)	20.4 (-6.5 to 47.2)	0.130
Allowance for time effects					
Number (%)*	15 (33)	3 (17)	12 (44)	27.8 (2.3 to 53.2)	0.063
Allowance for repeated measurements[1]					
Number (%)*	3/24 (13)	2/11 (18)	1/13 (8)	-10.5 (-37.5 to 16.5)	0.576
Power methodology					
Hussey and Hughes*	14 (31)	3 (17)	11 (41)	24.1 (-1.2 to 49.4)	0.111
Other, allowing for time effects *	2 (4)	0 (0)	2 (7)	7.4 (-2.5 to 17.3)	0.509
Other, not allowing for time effects*	14 (31)	10 (55)	4 (15)	-40.7 (-67.3 to -14.2)	0.007
Not stated*	15 (33)	5 (28)	10 (37)	9.3 (-18.3 to 36.8)	0.748
Power methodology for additional features					
Transition periods	0 (0)	0 (0)	0 (0)		
Interactions (e.g. lag effects)	0 (0)	0 (0)	0 (0)		
Extended correlations*	2 (4)	2 (11)	0 (0)	-11.1 (-25.6 to 3.4)	0.155
Varying cluster size *	3 (7)	1 (6)	2 (7)	1.9 (-12.6 to 16.3)	1.000
Variation in outcomes across clusters[2]					
Reported using ICC*	20/33 (61)	8/11 (73)	12/22 (55)	-18.2 (-51.7 to 15.4)	0.456
Reported using CV*	10/33 (30)	2/11 (18)	8/22 (36)	18.2 (-12.2 to 48.6)	0.430
Reported using DE*	1/33 (3)	1/11 (9)	0/22 (0)	-9.1 (-26.1 to 7.9)	0.333
Reported using between cluster variation*	2/33 (6)	0/11 (0)	2/22 (9)	9.1 (-2.9 to 21.1)	0.542

*Values are numbers (percentages) unless stated otherwise. [1]among those with a cohort design. P-value is for the comparison of full reports and protocols using a chi-squared test for proportions (categorical outcomes) or Mann-Whitney U test (where medians are reported). [2]As a percentage of studies for which this was appropriate. *using fishers exact test.*

3.3.5 Methodological rigor of the analysis methods used

Of the 32 full trial reports, the majority of them (75%) made allowances for clustering in the data analysis (Table 3.17). Typically this was done using a generalised linear mixed model (GLMM), though there were instances of studies using generalised estimating equations (GEE) and generalised linear model (GLM) with fixed cluster effects. However, only 53% of the studies acknowledged the confounding effect of time in the analysis. Of the 17 studies that accounted for time effects, this was most often done using a fixed effect, whilst 6 studies acknowledged that they adjusted for time, but did not specify clearly how it had been done. One study recognised that time should be adjusted for, but argued that it was not significant in the analysis. The majority of studies were not clear as to whether allowances had been made on repeated individuals, with only 9/24 studies correctly allowing for repeated measures. None of the included studies reported an allowance for the within-period and inter-period correlations by including a cluster by time interaction.

Table 3.17: Methods of analysis used in full trial reports

	Full trial report N = 32
Allowance for clustering	
Yes	24 (75)
No	7 (22)
Unclear	1 (3)
Model framework with clustering	
GLM with robust SEs*	0 (0)
GLM with fixed cluster effects	2 (6)
GLMM	16 (50)
GEE	3 (9)
Other	3 (9)
Unclear	2 (6)
Allowance for time effects	
Yes	17 (53)
No	15 (47)
Framework for time effects	
Fixed effects	7 (22)
Other	4 (13)
Not specified	6 (19)
Linear time effect	0 (0)
Unclear	4 (13)
No, but argued not significant	1 (3)
No allowance	10 (31)
Interactions	
Time by treatment	0 (0)
Extended clustering	
Allowance for within-period and inter-period correlations	0 (0)
Allowance for repeated measures on same individuals[1]	
Yes	9 (28)
No	3 (9)
Unclear	12 (38)
N/A	8 (25)

Values are numbers (percentages) unless stated otherwise. [1]*: Study designs that were cross-sectional in nature were categorised as n/a. GLM: generalised linear model. SE: standard errors. GLMM: generalised linear mixed model. GEE: generalised estimating equation.*

3.4 Discussion

Here, we have described a methodological review that assessed the quality of reporting of sample size calculations and of the methods of analysis used in full trial reports. The adherence to the CONSORT statement for randomised trials and the extension for cluster trials were reported for each study. Following this, the methodological rigor of both the sample size calculation and the analytical method used were assessed.

It is clear that the quality of reporting in SW-CRTs in sub-optimal. Only 33% of the studies allowed for time effects in the sample size calculation, despite it being an item that distinguishes it very clearly from a P-CRT. Failing to acknowledge the presence of time effects could lead to an under-estimate of the sample size needed to detect a relevant treatment effect. Similarly, in the analysis stage of a study, failing to adjust for time effects could lead to an over precise estimate of the treatment effect.

Whilst it seems apparent that a SW-CRT is used mainly by studies that are following a cohort or open cohort design, that is that some (or all) patients are followed-up for multiple time-periods, there is a distinct lack of clarification for the sample size for these studies. In fact, of the 24 studies who should have accounted for repeated measurements, only 3 (13%) made an allowance. Similarly, in the analysis stage, only 9/24 studies allowed for repeated measures.

In previous methodological papers considering estimating the power or sample size of a SW-CRT, it has often been assumed that the outcome is continuous, or else can be approximated via a normal distribution. However, it has not been tested how accurate the methods are if the outcome variable does not exhibit normality. It is therefore interesting that only 25% of studies used a continuous outcome, highlighting the fact that the majority used methodology that has been developed for a different outcome type. There is therefore an urgent need to develop sample size methodology for use in studies employing a binary outcome, as well as for other outcome types.

Noticeably, very few studies reported an extended correlation. One cohort SW-CRT protocol reported a correlation coefficient of repeated measurements (80) and a cross-sectional SW-CRT protocol reported a within-patient correlation over time (81) and powered as a before and after study. As such, the IPC and WPC have not previously been reported in a sample size calculation for a SW-CRT. However, since there is a dearth power methodology with reference to the IPC and WPC, it was expected that few studies would allow for any enhancement of previous sample size methodology.

Several studies reported a variation in cluster size within the report. However, some of these studies either did not report a sample size calculation (82) or did not report the method used to obtain the sample size (83). Two studies used DEs appropriate for a P-CRT, with one using the average cluster size (25) and the other using an adjusted DE acknowledging the coefficient of variation of cluster sizes (84). Only one study that reported a variation in cluster size used the Hussey and Hughes method to obtain the sample size (85) – but they did not report how they adjusted for the varying cluster size.

Many of the studies published most recently were protocols, and so some of the improvement in quality of reporting may be attributable to the report type (full trial or protocol) (86).

3.4.1 How does this review differ to those already carried out

Although there exists other systematic reviews that have considered the SW-CRT, and indeed the sample size calculation used, there are numerous differences between them and this review (23, 28, 29). Indeed, whilst previous reviews have broadly considered the methodology used to determine the sample size, none have systematically assessed the quality of reporting of SW-CRTs against the existing CONSORT statements. Furthermore, no previous review has assessed the methodological rigor of the sample size calculations, and investigated whether current standards are sufficient. Additionally, of the two earliest reviews in SW-CRTs, one review focused solely on health care interventions, whilst the other included non-randomised studies (23, 29). Since randomised studies differ greatly from non-randomised studies, the objective of this review was to assess the quality of randomised studies only. Moreover, since the CONSORT statements are related to randomised studies, studies published employing a non-randomised design may not adhere to the CONSORT guidelines.

In the area of P-CRTs, there exist numerous methodological reviews that assess the quality of reporting of the sample size calculations in relation to a smaller number of items recommended in the 2004 extension to the CONSORT statement (73, 74). In this review, assessments are made versus the 2012 extension for cluster trials. Although many trials may have been designed, or published, before the reporting guidelines were published, it is vital that we considered the most recent recommendations. Also, since the primary outcome is the quality of reporting, and not the adherence to guidelines, the most recent CONSORT extension was chosen.

3.4.2 Implications

In P-CRTs, it is well known that failing to acknowledge clustering in the sample size calculation will lead to an underpowered trial. Similarly, in the analysis stage, failing to adjust for clustering will lead to a treatment effect estimate that is overly precise. Since a SW-CRT is a type of cluster trial, similar consequences will occur if clustering is not accounted for in the sample size calculation and in the post-trial analysis. Unlike a P-CRT, the design and analysis of a SW-CRT should take temporal trends into account.

Whilst analytical methods for P-CRTs have been explored in more detail than the SW-CRT, it is clear that some methods cannot be easily transposed. For example, if a P-CRT has been conducted with equal cluster sizes, then one possible method of inference is to conduct an analysis of cluster means. Whilst for a P-CRT, this would lead to valid inferences, in a SW-CRT; an analysis of cluster level means would not take into account any time effects, and so would lead to invalid inferences.

In CRTs that follow a cohort design, participants have multiple observations over time. Methods for estimating the sample size in a P-CRT with a cohort design, such as a pre-post design, have previously been presented (52). Repeated measurements on the same participant can be acknowledged at the analysis stage through an additional random effect. This results in an additional correlation term that depicts the correlation between observations made on the same participant over time (87). As shown in previous reviews, SW-CRTs often follow a cohort design, but use methods of design and analysis appropriate for a cross-sectional design – so do not include this additional random effect. The implications of a cohort design on the sample size calculation for a SW-CRT are now being considered (26). Nevertheless, the additional complexities of a cohort design need to be stressed further to make sure correct methods are used in the design and analysis of cohort CRTs, ensuring that methods for cross-sectional and cohort designs are not conflated. In this piece of research, the cohort design is not considered further. Instead, the focus is solely on the cross-sectional SW-CRT.

3.4.3 Study limitations

This review assesses the reporting of items from both the CONSORT statement for individually randomised trials and the extension for cluster randomised trials. However, whilst relevant, some items from these statements are not naturally extendable to a SW-CRT. For example, it is recommended in the CONSORT statement for individually randomised trials that authors report whether attrition has been accounted for in the sample size calculation. However, it is often reported that SW-CRTs are useful study designs for studies that are using routinely collected data, in which case, attrition would not be relevant. In this review, 20% of studies were using routinely collected data, and so attrition would not have been an issue. However, rather than record this item as not applicable, it was still recorded as reported or not. As such, trials that are utilising routinely collected data may not report attrition as they feel it is irrelevant for their study design, rather than be a lack of reporting.

Whilst the most appropriate methodology currently derived for SW-CRTs is that proposed by Hussey and Hughes – which was the most used methodology – it has some limitations. There are a few key design features that are not immediately powered for. For example, it is assumed that the correlation between two observations within a cluster is the same regardless of measurement point. In reality, it is likely that there will be some variation. In this review, two studies allowed for a variation in the ICC across time, whilst no studies included a time by treatment effect interaction in their power calculation.

3.5 Conclusions

Whilst it has been acknowledged for P-CRTs that failing to account for clustering will lead to under estimates of the power, it should be highlighted that sample size calculations in SW-CRTs that fail to account for time effects may affect the power of the trial, and so lead to too few or too many participants being included. Many recent methodological developments mean it is now possible to determine the cluster size for specified design constraints, or determine the number of clusters or steps needed.

Although the SW-CRT is a very flexible design that allows for a pragmatic evaluation determined by the number of available clusters or participants, properly designed evaluations should include a robust justification of the sample size.

As expected, the quality of reporting of sample size calculations is sub-optimal. Although there is evidence to suggest that the quality is increasing over time, there is still need for further improvement. Furthermore, there is a distinct difference between the quality of reporting of full trials compared to protocols. Whilst it is known that methodological developments are needed to address areas of this complicated design, it is vital that studies currently using the design are using appropriate methodology. Of particular note, less than half of the included studies accounted for temporal trends in the sample size calculation, very few allowed for the repeated measurements of participants and studies were not clearly identifying whether the design was cross-sectional or cohort. As such, the majority of studies may be using sample size methodology that does not match the study design. This is also highlighted when considering the study outcome types. The primary outcome for the majority of studies was binary, whereas the most common methodology used assumes a continuous outcome. This emphasises the need for methodological development in these areas.

There are two key issues that emanate from this review. Firstly, despite the knowledge that clusters are likely to vary in size, there is little evidence that sample size calculations are accounting for varying cluster sizes. Indeed, those that mention varying cluster size are not then adjusting for it in the sample size calculation. Secondly, despite the SW-CRT being a longitudinal design, there has been little indication of SW-CRTs considering an extended correlation structure within the sample size calculation – and no indication of the IPC and WPC being included. However, there is little evidence in the methodological literature of the likely impact of these two issues on the sample size calculations for a SW-CRT.

The next chapter considers an extended correlation structure and establishes the methods required to estimate the inter-period correlation (IPC) and the within-period correlation (WPC) and creates a resource of IPC and WPC values.

CHAPTER 4: CORRELATION IN CLUSTER TRIALS

A paper based on the work from this chapter has been published.
Citation: *Martin J, Girling A, Nirantharakumar K, Ryan R, Marshall T, Hemming K. Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. Trials. 2016;17:402.*

4.1 Introduction

Typically in healthcare research, observations used in the evaluation and analysis of an intervention may stem from individuals, but the intervention is often aimed at the cluster level – such as hospitals, clinicians, or GP surgeries (88-92). In an individually randomised controlled trial (RCT), the patients themselves are randomised to the intervention or control arm. Increasingly, interventions are being evaluated using cluster randomised trials (CRT) (90, 91, 93). In a CRT, the clusters are the unit of randomisation, rather than the individuals within them. In an RCT, it is assumed that all participants are independent (92). This assumption may not be valid in a CRT (56). For example, in the WOSLAD study (section 1.6.1) (5), participants within a hospital share common staff and facilities, in addition to likely geographical and socio-economic similarities. As such, they are likely to be more similar than participants from different hospitals (89, 90). The similarities between participants may be relevant or irrelevant to the trial, but leads to a lack of independence between observations within a cluster (90, 92). The design and analysis of CRTs should recognise this within pre-trial sample size calculations and post-trial analysis (92, 94-97).

4.1.1 Design effects

In a RCT, pre-trial sample size calculations are straightforward, but without adjustment would produce an inaccurate sample size for a CRT (98). Sample sizes required under cluster randomisation (SS_{CRT}) are inflated over that required under individual randomisation (SS_{RCT}) to achieve the same power, through the use of a design effect (DE) (53), so that:

$$SS_{CRT} = SS_{RCT} \text{ x } DE$$

For a CRT with equal cluster sizes, the design effect required is well established (53), and is given as:

$$1 + (m - 1)\rho \qquad 4.1$$

Here m represents the cluster size and ρ is the correlation between two randomly chosen observations within a cluster. The value of ρ has important implications on the sample size requirements in a CRT (56, 99). This design effect should only be used for studies with equal sized clusters. Studies with unequal cluster sizes using the above design effect may obtain an underestimate of the necessary sample size (57). Alternative design effects for studies with varying cluster size are discussed further in Chapter 2.

4.1.2 Correlation between observations within a cluster

The most common type of CRT is a parallel CRT (P-CRT). In this study design, clusters are randomised to either intervention or control and this allocation is maintained over the study duration. However, alternative study designs are being increasingly used – such as the stepped-wedge design (6, 23), the cluster cross-over design (100, 101), and the dog-leg design (20, 48). These designs are longitudinal in the sense that observations are made over multiple time-periods.

Traditionally, in the analysis of a CRT, the models fitted to the data assume an exchangeable correlation structure. That is, the observations belong to one study-period, and the correlation between any two observations within a cluster is constant over time (time independent). However, in longitudinal designs, observations may be taken from different time-periods, and so the correlation may not be constant over time (time dependent), and an exchangeable correlation structure may not be appropriate. That is to say, ρ, which is necessary for a DE, is reliant on whether a time-dependent model or a time-independent model is fitted to the data.

Throughout this work, a cross-sectional design is assumed, so that participants contribute only one observation to the analysis. In a setting with correlation treated as exchangeable (time independent model), we define the intra-cluster correlation coefficient (ICC) as the correlation between two randomly selected participants within a cluster. In a setting with a time dependent model fitted to the data, we define the within-cluster correlation coefficient (WCC) as the correlation between two randomly selected participants within a cluster. As such, both the ICC and WCC are equivalent to ρ, but are only defined when an appropriate model is fitted.

4.1.2.1 *Intra-cluster correlation coefficients*

The ICC is defined as the correlation between two randomly selected observations from two participants within the same cluster (102) and is equivalent to ρ in a model with exchangeable correlation. In this context, there are three fundamental assumptions that are made:

1) Any two observations from participants within the same cluster are correlated.

2) The correlation is the same for all pairs of participants from within the same cluster regardless of the duration between the participants.

3) Any two observations made on participants from different clusters are independent.

To ensure a sufficiently powered study, an accurate estimate of the ICC is necessary. Dependent on the outcome type, there are numerous methods that can be used to estimate the ICC. Many previous trials in UK primary care have failed to report the ICCs in the post-trial analysis (73). As such, planned trials often use ad-hoc ICCs in the estimation of sample size (103), which may lead to an underpowered trial. Several papers have recommended that ICCs should be published routinely following the completion of a trial (54, 104). The CONSORT extension for cluster randomised trials recommends that authors report the ICCs in the post-trial analysis in addition to reporting the ICC used in the pre-trial sample size calculation (70). There are also examples of ICCs being presented from the analysis of large patient databases (105, 106).

4.1.2.2 *Within-cluster correlation coefficient*

In alternative study designs with repeated cross-sections, it is recognised that the correlation between observations made from within the same cluster and the same period will differ from the correlation between observations made from participants in the same cluster but at different periods (27, 61, 87, 107). In this context, the correlation is not exchangeable and the correlation is dependent on the timing of the observations. As shown by Figures 4.1 and 4.2, time is split into a number of (equal) time-periods. Within this framework, several assumptions are made:

1) Any two observations from the same cluster are correlated.

2) There is a constant correlation between any two observations within a cluster from the same time-period.

3) There is a constant correlation between any two observations within the same cluster but from different time-periods.

4) Any two observations from different clusters are independent.

Here, we define the within-cluster correlation (WCC) as the correlation between two randomly chosen observations within a cluster in this setting. Therefore, the WCC is equivalent to ρ when fitting a time-dependent model. Since the correlation between two observations within the same cluster and the same time-period is different to the correlation between observations within the same cluster at different time-periods, we conceptualise a within-period cluster correlation (WPC) and an inter-period cluster correlation (IPC).

The WPC is defined as the correlation between two randomly selected observations taken from the same cluster and the same period (Figure 4.1). The IPC is defined as the correlation between two randomly selected observations taken from participants in the same cluster but at different periods (Figure 4.2).

Figure 4.1: Comparing observations within the same cluster within the same period – the within-period correlation – in a cross-sectional SW-CRT

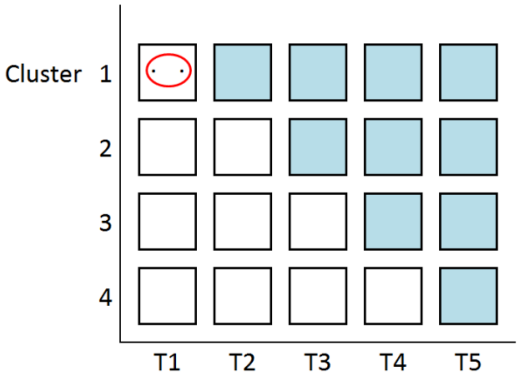
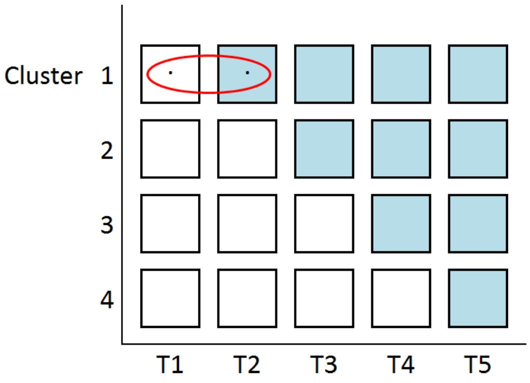


Figure 4.2: Comparing observations within the same cluster at different periods - the inter-period correlation – in a cross-sectional SW-CRT



The "." indicates an observation and the circle highlights the observations being compared.

In cluster cross-over trials, it has been established that the IPC influences the power of a design (100). In SW-CRTs, it has been shown that the IPC may impact the sample size and the number of clusters required (27).

4.1.2.3 *Correlation between observations with binary outcomes*

If the primary outcome for a trial is dichotomous, there is a lack of clarity surrounding the definition and calculation of the ICC. For dichotomous outcomes, the ICC can be estimated on a proportion scale or on a logistic scale (108) and the ICC may differ greatly between the two scales.

4.1.3 Why are estimates of the correlation important

CRTs crucially depend on values of the ICC for accurate sample size calculations (102). For continuous outcomes, the framework required to estimate the ICC is relatively straightforward. In longitudinal studies, the ICC may not be sufficient to describe the correlation between participants in a cluster, and the WCC should be used to replace ρ in a design effect. Early research suggests the IPC and WPC play a pivotal role in the power of a repeated cross-sectional CRT (27, 61, 87, 107). Moreover, a recent DE appropriate for SW-CRTs requires estimates of the IPC and WPC in order to estimate the sample size (26). Nevertheless, there is a lack of empirical literature in this area (27, 48), and so no published estimates of the IPC or the WPC exist, which has led to calls for values of the IPC and WPC to be published (26, 27). Furthermore, for binary outcomes, there is much confusion surrounding what type of ICC should be presented and used in sample size calculations. It is essential that the correct approach is underlined to ensure future sample size calculations are performed with the appropriate ICC.

4.1.4 Chapter aim

The main aim of this chapter is to highlight the different types of correlation that can be calculated in a longitudinal CRT. This includes evaluating the validity of the assumption that the correlation between observations within a cluster is independent of time. To this end, we will:

1. Demonstrate how the correlation between observations within a cluster can be described in time-dependent and time-independent settings.

2. Outline a method to estimate the IPC and WPC.

3. Illustrate using a case study what these correlations may look like in practice.

4. Investigate whether study length or period-length impacts the IPC and WPC.

In this chapter, we begin with the framework for estimating ICCs for continuous outcomes. It is then shown how the IPC and the WPC can be estimated. For illustrative purposes, estimates of the ICC, IPC and WPC are presented in relation to outcomes associated with type-2 diabetes. The methodology for the differing approaches to the ICC for binary outcomes is also considered.

4.2 Methods

4.2.1 Intra-cluster correlation coefficient

Typically, P-CRTs assume exchangeable correlation – that is, observations are identically correlated regardless of the duration between them. The ICC – defined as the correlation between two randomly selected observations from participants within the same cluster – is defined only when fitting this time independent model. For any two observations *k* and *k'* in cluster *i*, the correlation between them can be given as:

$$ICC = corr(Y_{ik}, Y_{ik'})$$

It is assumed that this correlation is independent of the time in which the observations are made. This concept is consistent with a decomposition of the total variance into two independent components – signifying variation between cluster (${\sigma_b}^2$) and between subjects (within clusters) (${\sigma_w}^2$). Below, we highlight the differences between ICCs for continuous and dichotomous outcomes.

4.2.1.1 *Continuous outcomes*

For continuous outcomes, the ICC provides a measure of the homogeneity of the outcome within clusters. In this setting, the ICC can alternatively be defined as the proportion of the total variance that is attributable to the between-cluster component. Assuming that the total variance in a CRT is equal to the sum of the between-cluster variance (${\sigma_b}^2$) and the between-subject (within cluster) variance (${\sigma_w}^2$), the ICC is given as:

$$ICC = \frac{{\sigma_b}^2}{{\sigma_b}^2 + {\sigma_w}^2} \qquad \textbf{4.2}$$

(See Appendix D).

In the standard analysis of a CRT (assuming exchangeable correlation), an ICC calculated using equation 4.2 is equivalent to ρ and can be used in the DE for a CRT (equation 4.1).

The analysis of a cluster trial is typically undertaken using a multi-level linear model (See Box 4.1 and section 4.2.4 for model details). Ratio of variances is then used to estimate the ICC. Throughout this work, this approach is taken whenever an estimated ICC is reported.

Box 4.1: Time-independent model for the analysis of a continuous outcome

$$\boldsymbol{Y_{ik} = \mu + \alpha_i + \varepsilon_{ik}}$$

Where Y_{ik} is the outcome for patient k in cluster i, μ is the mean outcome, α_i is a cluster effect, and $\varepsilon_{ik}$is the residual error.

It is assumed that the following distributions apply:

$\alpha_i \sim N(0, {\sigma_b}^2)$ $\varepsilon_{ijk} \sim N(0, {\sigma_w}^2)$

4.2.1.2 *Binary outcomes*

In clinical trials, outcome data is often dichotomous – usually the presence or absence of a particular clinical outcome. The ICC is more complex for binary outcomes, which can lead to a lack of clarity in its definition (102, 105, 108-111). Following our earlier definition, the ICC is a measure of the correlation between two dichotomous outcomes from two randomly selected participants within the same cluster, estimated on the proportions scale. Sample size calculations for trials with a binary outcome classically involve a normal approximation to the binomial distribution of grouped data. Nonetheless, multi-level logistic models are commonly used for the post-trial analysis of dichotomous outcomes in a CRT.

Latent intra-cluster correlation coefficient

Hypothetically, the observed binary outcome for a multi-level logistic model may be obtained via the dichotomisation of a continuous latent scale. Upon fitting a multi-level logistic model (See Box 4.2) in a statistical package (e.g. STATA); a type of ICC is reported. However, this ICC relates to the unobservable latent scale, and not to the observed binary outcomes. This ICC (ρ_L) takes the form:

$$\rho_L = \frac{{\sigma_L}^2}{{\sigma_L}^2 + \frac{\pi^2}{3}} \tag{4.3}$$

Where ${\sigma_L}^2$ is the between cluster variance (on the latent scale) and $\frac{\pi^2}{3}$ refers to the variance of the logistic distribution which is used to generate the binary model (108).

Throughout this work, we refer to an ICC estimated in this manner as a *latent ICC* - to reflect that it refers to the unobservable latent scale, rather than the correlation between the dichotomous outcomes of two participants from within the same cluster.

Box 4.2: Time-independent model for binary outcomes

$$\boldsymbol{ln\left(\frac{p_{ik}}{1 - p_{ik}}\right) = \mu + \alpha_i}$$

Where p_{ik} is the probability of the outcome for patient *k* in cluster i, μ is the mean outcome, α_i is a cluster effect, where: $\alpha_i \sim N(0, {\sigma_L}^2)$

Natural intra-cluster correlation coefficient

The latent ICC should not be used directly in the computation of a design effect for use in a sample size calculation, since it refers to the unobservable latent scale. An ICC in a design effect should refer to the correlation between observations in a cluster. For binary outcomes, we refer to this as the *natural ICC* (ρ_N). The natural ICC is estimated on the proportions scale via a mixed effects linear model (See Box 4.1), and is calculated as:

$$\rho_N = \frac{{\sigma_b}^2}{{\sigma_b}^2 + {\sigma_w}^2} \qquad 4.4$$

Box 4.1: Time-independent model for the analysis of a continuous outcome

$$\boldsymbol{Y_{ik} = \mu + \alpha_i + \varepsilon_{ik}}$$

Where Y_{ik} is the outcome for patient *k* in cluster i, μ is the mean outcome, α_i is a cluster effect, and $\varepsilon_{ik}$is the residual error.

It is assumed that the following distributions apply:

$$\alpha_i \sim N(0, {\sigma_b}^2) \qquad \varepsilon_{ijk} \sim N(0, {\sigma_w}^2)$$

By considering the prevalence of the outcome, the latent ICC can be converted into a natural ICC for the raw data – see, for example, the table presented by Eldridge et al. (1). Throughout this work, a clear distinction between the natural ICC and the latent ICC is maintained for binary outcomes.

4.2.2 Correlation in longitudinal studies

In longitudinal CRTs, the difference in time between two observations is likely to impact the degree of correlation and the assumption of exchangeable correlation may not hold. When observations are not identically correlated independent of time, a time dependent model can be fitted to the data. In this context, we define the within-cluster correlation (WCC) as the correlation between two randomly chosen observations within a cluster. In a time dependent model setting, the WCC is equivalent to ρ and can be used in a design effect.

In a time-dependent model setting, time is split into a number of (equal) periods. In this design, constant correlations are assumed for:

a) For any two observations within a cluster from the same time-period (WPC)

b) For any two observations from within the same cluster but from different time-periods (IPC).

These assumptions are consistent with a variance-decomposition into three independent components: between clusters (τ^2), between time-periods (within clusters) $\left(\sigma_p{}^2\right)$ and between-subjects (within time-period and cluster) $(\sigma_t{}^2)$. The time-dependent model fitted is given in Box 4.3

Box 4.3: Time-dependent model for continuous outcomes

$$\boldsymbol{Y_{ijk} = \mu + \alpha_i + \omega_{ij} + \varepsilon_{ijk}}$$

Where Y_{ijk} is the outcome for patient *k* in cluster *i* at time *j*, μ is the mean outcome, $\alpha_i$is the cluster effect, ω_{ij} is a random effect for cluster *i* at time *j*, and $\varepsilon_{ijk}$is the residual error.

It is assumed that the following distributions apply:

$\alpha_i \sim N(0, \tau^2)$ $\omega_{ij} \sim N(0, \sigma_t{}^2)$ $\varepsilon_{ijk} \sim N\left(0, \sigma_p{}^2\right)$

In a framework with a time dependent model, the WCC can be calculated using the IPC and the WPC, as:

$$WCC = IPC + \frac{1}{T}(WPC - IPC) \qquad 4.5$$

See Appendix E.

Here *T* is the number of time-periods in the study. It is assumed that each time-period contains an equal number of observations.

4.2.2.1 *Within-period cluster correlation*

The WPC is the correlation between two randomly selected observations from within the same cluster and from within the same time-period. For two participant's *k* and *k'* in cluster *i* at time *j*, the correlation between them can be given as:

$$WPC = corr\left(Y_{ijk}, Y_{ijk'}\right)$$

Now, the WPC can be written as:

$$WPC = corr\left(Y_{ijk}, Y_{ijk'}\right) = \frac{\tau^2 + {\sigma_t}^2}{\tau^2 + {\sigma_p}^2 + {\sigma_t}^2} \qquad 4.6$$

(See Appendix F)

4.2.2.2 *Inter-period cluster correlation*

The IPC indicates the correlation between two randomly selected observations from within the same cluster but from different time-periods. As such, it provides the correlation between participant's *k* and *k'* from cluster *i* at times *j* and *j'*, and can be given as:

$$IPC = corr\left(Y_{ijk}, Y_{ij'k'}\right)$$

Now, the IPC can be written as:

$$IPC = corr\left(Y_{ijk}, Y_{ij'k'}\right) = \frac{\tau^2}{\tau^2 + {\sigma_p}^2 + {\sigma_t}^2} \qquad \mathbf{4.7}$$

(See Appendix G)

4.2.2.3 ***Cluster autocorrelation***

The IPC and WPC describe the correlation of participants within a cluster. This has previously been presented as a cluster autocorrelation (CA) (27) and has been established as vital to sample size formulae in cluster cross-over designs (112). The CA is defined as the correlation between the cluster level outcome from two different time-periods from a fixed cluster (i.e. conditional on cluster) (48, 112). The CA can be calculated as the ratio of the within-cluster within-period variance to the within cluster variance, and can also be viewed as the ratio of the IPC to the WPC, as follows:

$$CA = \frac{IPC}{WPC} = \frac{\tau^2}{\tau^2 + {\sigma_t}^2}$$

In the absence of period effects, the CA = 1 (since ${\sigma_t}^2 = 0$), indicating that the time-dependent model is unnecessary. In this setting, WCC = WPC = IPC. Otherwise it follows from the definitions that WPC > WCC > IPC.

4.2.2.4 ***Summary of correlation structures***

To summarise, the ICC, defined only for a time-independent model (see Box 4.1), is the correlation between two randomly selected observations and is calculated as:

$$ICC = \frac{{\sigma_b}^2}{{\sigma_b}^2 + {\sigma_w}^2}$$

The WCC, defined only for a time-dependent model (see Box 4.3), is the correlation between two randomly selected observations and is calculated as:

$$WCC = IPC + \frac{1}{T}(WPC - IPC)$$

The WPC, defined only for a time-dependent model (see Box 4.3), is the correlation between two observations within the same cluster and within the same time-period and is calculated as:

$$WPC = \frac{\tau^2 + {\sigma_t}^2}{\tau^2 + {\sigma_p}^2 + {\sigma_t}^2}$$

The IPC, defined only for a time-dependent model (see Box 4.3), is the correlation between two observations within the same cluster but from different periods and is calculated as:

$$IPC = \frac{\tau^2}{\tau^2 + {\sigma_p}^2 + {\sigma_t}^2}$$

4.2.3 Estimating correlation between observations for type 2 diabetes

In section 4.2.1, we have discussed the methodology required to estimate the time independent correlation between participants within a cluster (the ICC). An extension to this methodology has been presented in section 4.2.2 for estimating the correlation between observations in longitudinal designs (the WPC and IPC) which may appropriate for SW-CRTs. Now, we introduce a scenario for which this methodology can be illustrated. Here, estimates of the ICC, IPC, and WPC are presented for typical outcomes related to type-2 diabetes.

Below, the necessity for these estimates is discussed, alongside a patient database that will allow for these estimates to be calculated.

Diabetes is progressively becoming a more influential disorder, and is a significant health issue (113). To overcome this, an increasing number of clinical trials are being conducted with the aim of lowering the risk of diabetes (114) or to improve care for pre-existing sufferers (91, 115, 116).

When conducting a trial related to diabetes, typical outcomes might be clinical measurements such as HbA1c (both as a continuous and dichotomised outcome) (116), body mass index (BMI) (117), cholesterol (118), blood pressure (119) or the incidence of macrovascular and microvascular outcomes (120). Regardless of the outcome, pre-trial sample size calculations require an accurate estimate of the correlation of the outcomes within a cluster (56, 121). As such, we provide estimates of the ICC, IPC, and WPC for typical trial outcomes related to type-2 diabetes. These are made using routinely collected data, obtained from The Health Improvement Network (THIN) database (122). This is a large archive of anonymised patient records from UK general practices. Using THIN database, data was obtained for a retrospective cross-section of patients with type-2 diabetes. Participating general practices contributed anonymised demographic, prescribing information and clinical data for more than 3.7 million patients throughout the UK (122).

4.2.3.1 *Inclusion criteria*

A set of practice level and patient level inclusion criteria was used. In order for data from the GP practices to be included, they were required to use the Vision computer system, and this must have been used for a minimum of one year. An acceptable mortality reporting date can be used to indicate whether a practice is reporting data sufficiently (123), and so all GP practices were required to have an AMR date recorded. Patients were included if a diagnosis of type-2 diabetes was made before the measurement period began. A diagnosis was indicated by the appropriate Read codes - a coded thesaurus of clinical terms (124). Read codes allow for the recording of clinical information in primary care electronic medical records in the UK (125). Generally, trials within diabetes focus on type-2 diabetes patients, as these make up the majority of diabetic patients (126), and so the focus here is on patients with type-2 diabetes only. Patients were required to be over 18 to ensure a correct diagnosis of type-2 diabetes, rather than a misdiagnosis of type-1 diabetes.

4.2.3.2 ***Outcome variables***

All variables clinically relevant to a trial in type-2 diabetes that are routinely recorded were included. The variables were divided into three distinct categories: clinical measures, medication prescription, and clinical outcomes. Clinical measures included: HbA1c; systolic blood pressure; diastolic blood pressure; body mass index (BMI); total cholesterol level; and high-density lipoprotein (HDL) cholesterol level. Medication measurements involved the prescribing of: Insulin; and other hypoglycaemic medications. The clinical outcomes deemed relevant were: atrial fibrillation; chronic kidney disease; chronic obstructive pulmonary disease; ischaemic heart disease; peripheral vascular disease; and stroke.

Whilst it is advisable to keep continuous outcomes in their raw form, trialists often dichotomise a continuous measure (127, 128). Since primary outcomes are more likely to be binary than continuous (74), cut-points have been chosen for each continuous outcome that reflects a clinically relevant value. This ensures that the results of this work will be useful for all trialists. A summary of the cut-points used in the dichotomisation of the clinical measures is given below in Table 4.1, and then discussed further.

Table 4.1: Cut points for dichotomising clinical measures

Outcome		Cut-point
Clinical measure		
HbA1c	(>)	7.5
Systolic blood pressure	(>)	140.0
Systolic blood pressure	(>)	130.0
Diastolic blood pressure	(>)	80.0
BMI	(>)	30.0
BMI	(>)	25.0
Total cholesterol	(>)	4.0
HDL Cholesterol	(<)	1.2

For HbA1c, a threshold value of 7.5% was selected as NICE guidelines state that levels above 7.5% indicate inadequate control (129). A threshold value of 7.5% has additionally been used in previous studies (130). Recommendations have been made that HDL cholesterol level should be above 1.2mmol/L (131). Equally, levels of total cholesterol should be below 4.0mmol/L (131). For systolic blood pressure (SBP), two values were chosen – 140mmHg is the upper limit recommended for patients with type-2 diabetes (129), whilst 130mmHg is the target who patients who have suffered a stroke, or those who suffer from kidney and eye problems (129). Two thresholds were chosen for BMI to match to the groupings of overweight (25kg/m^2) and moderately obese (30kg/m^2) (132).

4.2.3.3 *Measurement period used to estimate the ICC*

To estimate ICCs, a cross-section of patients was identified for use in the analysis. A 15 month cross-section was chosen, as this reflects the NICE quality and outcomes framework (QOF) (133), for which the measurements taken during a 15 month period is monitored. Since QOF offers a financial incentive to GP surgeries who exhibit high levels of care, a 15 month period should provide measurements for a large subset of the total patients in the database. Observations are taken from 1st February 2009 to 30th April 2010. If multiple observations were made for a patient's outcome during this period then the measurement closest to the end date was included in the analysis.

However, in 2009, the measuring unit for HbA1c changed to mmol/mol from % HbA1c. Naturally, the reporting consistency is likely to be poor around this period. To prevent confusion and irregularity, it is logical to include a different cross-section of measurements. To this end, a cross-section of HbA1c observations are taken from the period 1st January 2008 to 31st December 2008. A 12 month period, rather than a 15 month period, was included due to data limitations.

4.2.3.4 *Measurement period used to estimate time-dependent correlation*

To estimate the IPC, and WPC, the study must be divided into a number of (equal) periods. In a real-world SW-CRT, these periods will be derived by the repeated cross-section design. Since data here originates from a patient database, the period lengths are not fixed. The study design we therefore chose is an extension of the design used for estimation of the ICC. It consists of two-periods – with each period lasting 15 months – resulting in a 30 month study. Measurements taken between 1st November 2007 and 30th April 2010 contribute to the analysis. The 15 month period from 1st November 2007 to 31st January 2009 refers to period one, and observations from the 15 month period 1st February 2009 to 30th April 2010 contribute to period two. For HbA1c, observations from 1st January 2007 to 31st December 2007 contribute to period one, and from 1st January 2008 to 31st December 2008 to period two. Patients could contribute only one measurement to the analysis, with the measurements closest to the end date used.

Study length and period length

The IPC and WPC are based upon the dissection of a study into (equal) periods. Since they are measuring the correlation between participants in the same cluster within a period and between periods, the length of the period, and the length of the study, may influence their value.

Study length

To evaluate the impact of study duration on the IPC and WPC, we extend the 2-year cross-section used to evaluate HbA1c and use incremental year increases (but with a fixed 1-year period length) up to a 6-year study period. This provides IPC and WPC estimates that correspond to studies that are 2-years, 3-years, 4-years, 5-years, and 6-years in study length. In a 6-year study period (with 1-year period length) the WPC will measure the correlation between two observations in the same cluster and during the same year. The IPC will measure the correlation between two observations in the same cluster, but with one observation in one year, and the other observation from any of the other five years. The dates used to define each period are given on Table 9.3 in Appendix H.

Period length

To assess whether period length influences the estimates of IPC and WPC, we consider the 15-month cross-section 1st February 2009 to 30th April 2010 used to estimate the ICCs. This cross-section is then divided into a number of (equal) time-periods. Four sets of estimates of the IPC and WPC are calculated, corresponding to a design with 2 (equal) periods, 3 (equal) periods, 4 (equal) periods, and 5 (equal) periods. In a 5-period study (with 15-month study length) the WPC will measure the correlation between two observations in the same cluster and during the same time-period. The IPC will measure the correlation between two observations in the same cluster, but with one observation in one time-period, and the other observation from any of the other four time-periods. Details of the dates used to define each period are given on Table 9.4 in Appendix H.

4.2.3.5 *Data summary*

Patient and practice level characteristics were summarised using suitable summary statistics. General Practice (GP) characteristics include the total number of practices, location (country) of the practice and practice inclusion size (the number of patients from each practice satisfying the entry criteria). Patient characteristics included number of participants, age (years), gender, location (country of residence), deprivation quintiles (IMD score) and number of deaths. Potential trial outcomes were summarised using appropriate summary statistics. Outcomes included clinical measurements, the prescribing of medication and the onset of clinical outcomes.

4.2.4 Statistical models

In this section, we highlight the statistical models that are used to estimate the ICC, IPC, and WPC.

4.2.4.1 *Time-independent model for continuous outcomes*

In order to estimate ICCs, generalised linear mixed models are fitted to the data, with cluster modelled as a random effect. The model is fitted as:

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik} \qquad \mathbf{4.8}$$

Where Y_{ik} is the outcome for patient *k* in cluster *i*, μ is the mean outcome, α_i is a random cluster effect, and ε_{ik} is the residual error. It is assumed that the following distributions apply:

$$\alpha_i \sim N(0, {\sigma_b}^2) \qquad \varepsilon_{ik} \sim N(0, {\sigma_w}^2)$$

Following the fitting of equation 4.8, the ICC was estimated as the ratio of the between cluster variance (of the outcome) to the total variance (of the outcome) (equation 4.2).

4.2.4.2 *Time-independent model for binary outcomes*

For binary outcomes, a mixed effects linear model (as given above by equation 4.8) was fitted to estimate the natural ICC. To estimate the latent ICC, a mixed effects logistic regression model was fitted as:

$$ln\left(\frac{p_{ik}}{1 - p_{ik}}\right) = \mu + \alpha_i \qquad \mathbf{4.9}$$

Where p_{ik} is the probability of the outcome occurring for participant *k* in cluster *i*, μ is the mean outcome, α_i is a random cluster effect, where:

$$\alpha_i \sim N(0, {\sigma_L}^2)$$

The latent ICC can then be estimated using equation 4.3.

4.2.4.3 *Time-dependent model for continuous outcomes*

To estimate time-dependent correlations, the generalised linear mixed model given in section 4.2.4 must be extended to include an additional random effect. To this end, a random effect for cluster by period is included in the model. Now, the model is fitted as:

$$Y_{ijk} = \mu + \alpha_i + \omega_{ij} + \varepsilon_{ijk} \qquad \textbf{4.10}$$

Where Y_{ijk} is the outcome for patient *k* in cluster *i* at time *j*, μ is the mean outcome, $\alpha_i$is the cluster effect, ω_{ij} is a random effect for cluster *i* at time *j*, and $\varepsilon_{ijk}$is the residual error.

Now, it is assumed that the following distributions apply:

$$\alpha_i \sim N(0, \tau^2) \qquad \omega_{ij} \sim N(0, {\sigma_t}^2) \qquad \varepsilon_{ijk} \sim N\left(0, {\sigma_p}^2\right)$$

The WPC is then estimated as the ratio of the between-cluster variance (of the outcome) and between-subject variance (of the outcome) to the total variance (of the outcome) (equation 4.6). The IPC is estimated as the ratio of the cluster variance (of the outcome) to the total variance (of the outcome) (equation 4.7).

4.2.4.4 *Implementation of models*

When estimating ICCs, IPCs and WPCs for typical outcomes related to type-2 diabetes, both unadjusted and adjusted models were fitted to the data. Adjustments were made for age, sex, location, and deprivation quintiles. All analysis was performed using Stata 13 (StataCorp, Texas, USA). Linear models were fitted using the "*mixed*" command, fitted via a maximum likelihood, and logistic models using the "*melogit*" command. All point estimates and confidence intervals of the ICC, WPC, and IPC were calculated using the default version of "*estat icc*". The "estat icc" function estimates the latent ICC when a logistic model (equation 4.9) has been fitted to the data.

4.3 Results

4.3.1 Included patient and cluster demographics

Included patient characteristics from THIN database are summarised at the practice level (Table 4.2) and at the patient level (Table 4.3) for the 15-month period (1st February 2009 to 30th April 2010). All patients satisfying the inclusion criteria contributed to the analysis - with 112,633 patients from 430 contributing GP surgeries included. The majority of both practices (75%) and patients (79%) were from England.

Table 4.2: Summary of practice level characteristics of THIN study population for the 15-month study period 01/02/2009 to 30/04/2010

Practice Characteristics	
Number of GP Practices	430
GP Size[1], Median [IQR]	241 [150-351]
Location, N (%)	
England	322 (75)
Northern Ireland	21 (5)
Scotland	56 (13)
Wales	31 (7)

[1]: Here GP size corresponds to the number of patients within the GP practice who have satisfied the entry criteria (patients who had C10F Read (version 2) code for diabetes entered on the Vision GP patient management system or other codes specifying type-2 diabetes)

When considering HbA1c, it is usually described assuming normality. Here, the mean (7.35) was greater than the median (7.05) – emphasising the positive skewness that HbA1c exhibits. A large proportion (59.1%) were reported as being prescribed insulin, whilst 29.4% were taking other hypoglycaemic medication – though some patients may have been taking this in addition to insulin. When considering the first incident of clinical outcomes, chronic kidney disease was the least common (0.332%), whilst atrial fibrillation was the most common (1.034%).

Table 4.3: Summary of patient level characteristics of THIN study population for the 15-month study period 01/02/2009 to 30/04/2010

Patient Characteristics	
Number of patients	112,633
Age, Median [IQR]	70 [60 -78]
Sex (Male), N (%)	61,944 (55)
Death, N (%)	4,237 (4)
Location, N (%)	
England	88,838 (79)
Northern Ireland	3,464 (3)
Scotland	12,461 (11)
Wales	7,870 (7)
Deprivation quintiles, N (%)	
1 (most affluent)	23,853 (21)
2	23,106 (21)
3	23,031 (20)
4	22,054 (20)
5 (most deprived)	16,352 (15)
Unknown	4,237 (4)
Clinical measures	
HbA1c (%), Median [IQR]	7.05 [6.4-7.9]
HbA1c (%), Mean (SD)	7.35 (1.41)
Systolic blood pressure (mmHg), Mean (SD)	134 (16)
Diastolic blood pressure (mmHg), Mean (SD)	75 (10)
BMI (kg/m^2), Median [IQR]	29.9 [26.4 - 34.2]
Total cholesterol (mmol/L), Median [IQR]	4.1 [3.5 - 4.7]
HDL cholesterol (mmol/L), Median [IQR]	1.19 [1.00 - 1.40]
Medication, N (%)	
Insulin	66,520 (59.1)
Other hypoglycaemic medication	33,061 (29.4)
Clinical outcomes, N (%)	
Atrial fibrillation	1075 (1.034)
Chronic kidney disease	362 (0.332)
Chronic obstructive pulmonary disease	848 (0.804)
Ischaemic heart disease	924 (1.061)
Peripheral vascular disease	566 (0.537)
Stroke	441 (0.410)

BMI: Body mass index. Note: The percentage corresponds to the number of applicable patients, and so the total may not be identical for each outcome variable.

Table 4.4 presents the number of patients who exceed the dichotomised value that was deemed clinically relevant. Over one third (34.2%) of patients had a HbA1c value exceeding 7.5%. The lower recommendation for systolic blood pressure (130mmHg) was surpassed by 58.4% of patients, whilst over one quarter (27.0%) had a systolic blood pressure greater than 140mmHg. When describing BMI in categorical terms, 0.54% were underweight (BMI<18.5), 15.92% were normal weight (18.5<BMI<25), 34.38% were overweight (25<BMI<30), 27.46% were obese (30<BMI<35) and 21.71% were morbidly obese (BMI>35).

Table 4.4: Number of participants exceeding the clinical measures cut off during the 15-month study period 01/02/2009 to 30/04/2010

Outcome	Number of observations	Number exceeding measurement N (%)
Clinical Measures		
HbA1c (%) (>7.5)	101,412	34,723 (34.2)
Systolic blood pressure (mmHg) (>140)	105,147	28,415 (27.0)
Systolic blood pressure (mmHg) (>130)	105,147	61,423 (58.4)
Diastolic blood pressure (mmHg) (>80)	105,147	25,307 (24.1)
BMI (kg/m^2) (>30)	97,469	47,922 (49.2)
BMI (kg/m^2) (>25)	97,469	81,429 (83.5)
Total cholesterol (mmol/L) (>4)	101,108	50,813 (50.3)
HDL cholesterol (mmol/L) (<1.2)	78,985	39,800 (50.4)

4.3.2 Intra-cluster correlation coefficient

4.3.2.1 *Continuous outcomes*

In Table 4.5, estimates of the ICC, along with a 95% confidence interval, are presented for all continuous outcomes. The median unadjusted ICC was 0.026 [IQR: 0.020 – 0.032], whilst adjusting for covariates had little effect (median = 0.025 [IQR: 0.020 – 0.032]). The main outcome for trials relating to type-2 diabetes is HbA1c. Here, the unadjusted ICC was estimated as 0.032 (95% CI: 0.027 to 0.037), and 0.032 (95% CI: 0.028 to 0.037) after adjusting for covariates.

Table 4.5: Intra-cluster correlation coefficients (ICCs) for continuous outcomes associated with type-2 diabetes from THIN database for study period 01/02/2009 to 30/04/2010

Outcome	Unadjusted model		Adjusted model[1]	
	ICC	95% CI	ICC	95% CI
Clinical measures				
HbA1c	0.032	0.027 to 0.037	0.032	0.028 to 0.037
Systolic blood pressure	0.030	0.026 to 0.035	0.028	0.024 to 0.033
Diastolic blood pressure	0.039	0.034 to 0.045	0.039	0.034 to 0.045
BMI	0.019	0.016 to 0.023	0.022	0.018 to 0.026
Total cholesterol	0.020	0.017 to 0.024	0.020	0.016 to 0.023
HDL cholesterol	0.021	0.018 to 0.025	0.020	0.017 to 0.024

[1]: Adjusted for age, sex, location, and deprivation quintiles.CI: Confidence interval.

4.3.2.2 *Dichotomous outcomes*

Estimates of the ICC and corresponding 95% confidence interval for the dichotomous outcomes are reported in Table 4.6. Estimates of the ICC from a logistic model (latent ICC) and from a linear model (natural ICC) are both presented. For HbA1c, the latent unadjusted ICC was estimated as 0.0350 (95% CI: 0.0298 to 0.0410), whilst the natural ICC was 0.0260 (95% CI: 0.0222 to 0.0304).

The median natural ICC for all binary outcomes was 0.0150 (IQR: 0.0023 to 0.0366) and median latent ICC was 0.0456 (IQR: 0.0255 to 0.1135). For clinical measures only, the medium natural ICC was 0.0274 (IQR: 0.0178 to 0.0368) and medium latent ICC was 0.0353 (IQR: 0.02365 to 0.0538). Clinical outcomes a medium natural ICC of 0.0021 (IQR: 0.0009 to 0.0042) and medium latent ICC was 0.0930 (IQR: 0.0256 to 0.1330). The latent ICC has noticeable greater values than the natural ICC. This is highlighted further in Figure 4.3. Natural ICCs were smaller than latent ICCs for all outcome variables, and also contained within a smaller range (Figure 4.3 and Table 4.6). The results are predominantly similar when adjusting for covariates (Table 9.6 in Appendix I).

Figure 4.3: Box plot of ICC estimates of continuous and binary outcomes

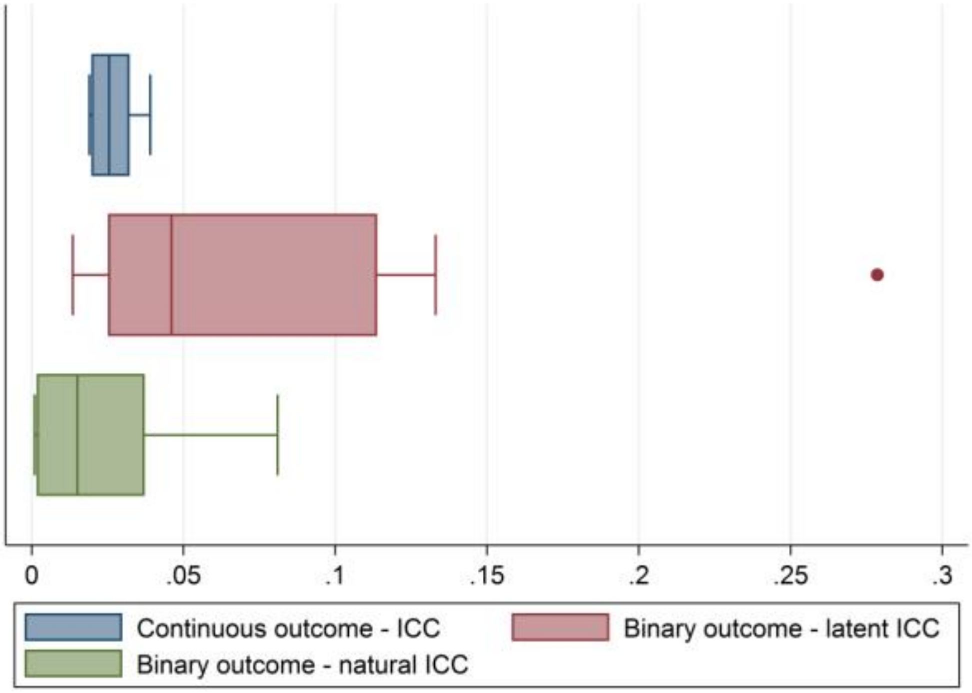


Table 4.6: Intra-cluster correlation coefficients (ICCs) for binary outcomes (medication prescribing, clinical outcomes, and dichotomised clinical measures) associated with type-2 diabetes from THIN database for study period 01/02/2009 to 30/04/2010

Outcome	Prevalence of outcome	Latent ICC (95% CI)	Natural ICC (95% CI)
Clinical measures			
HbA1c (>7.5)	0.34240	0.035 (0.030 to 0.041)	0.026 (0.022 to 0.030)
Systolic blood pressure (>140)	0.27024	0.062 (0.054 to 0.072)	0.0369 (0.032 to 0.043)
Systolic blood pressure (>130)	0.58416	0.046 (0.039 to 0.053)	0.037 (0.032 to 0.043)
Diastolic blood pressure (>80)	0.24068	0.082 (0.071 to 0.095)	0.046 (0.040 to 0.053)
BMI (>30)	0.49166	0.0186 (0.016 to 0.022)	0.015 (0.013 to 0.018)
BMI (>25)	0.83543	0.0218 (0.018 to 0.027)	0.010 (0.008 to 0.013)
Total cholesterol (>4)	0.50256	0.0255 (0.022 to 0.030)	0.0205 (0.017 to 0.024)
HDL cholesterol (<1.2)	0.50389	0.0356 (0.030 to 0.042)	0.029 (0.024 to 0.034)
Medication			
Taking of Insulin	0.59059	0.1135 (0.010 to 0.129)	0.081 (0.071 to 0.093)
Clinical outcomes			
Atrial fibrillation	0.01034	0.0135 (0.004 to 0.047)	0.001 (0.000 to 0.002)
Chronic kidney disease	0.00332	0.133 (0.089 to 0.193)	0.002 (0.002 to 0.003)
Chronic obstructive pulmonary disease	0.00804	0.0617 (0.0387 to 0.097)	0.002 (0.001 to 0.003)
Ischaemic heart disease	0.01061	0.0256 (0.0107 to 0.060)	0.001 (0.000 to 0.002)
Peripheral vascular disease	0.00537	0.124 (0.090 to 0.1688)	0.004 (0.003 to 0.006)
Stroke	0.00410	0.2786 (0.220 to 0.3459)	0.011 (0.009 to 0.013)

CI: Confidence interval

4.3.3 Correlation in longitudinal studies

Estimates of the WPC, the IPC, and the CA, for a 30-month study period are given in Table 4.7. The IPC is lower than the WPC (median IPC: 0.0188, median WPC: 0.0259), indicating the presence of a period effect. This is further highlighted by the CA (median CA 0.5910). The CA is smallest for total cholesterol, reflected the large difference between the WPC (0.021) and the IPC (0.010). This seems to indicate that total cholesterol is the outcome most affected by period effects. The IPC for HbA1c is 0.019 (95% CI 0.014 to 0.026) which indicates that the correlation between two participants within the same cluster at different time-periods is 0.019. The WPC for HbA1c is 0.035 (95% CI 0.030 to 0.040) which signifies that the correlation between two participants in the same cluster at the same period is 0.035. The results are predominantly similar when adjusting for covariates (Table 9.6 in Appendix I).

Table 4.7: Estimates of the within-period correlation and inter-period correlation for continuous outcomes associated with type-2 diabetes from THIN database for study period 01/11/2007 to 30/04/2010

Outcome	Unadjusted model		
	WPC (95% CI)	IPC (95% CI)	CA
HbA1c (%) [1]	0.035 (0.030 to 0.040)	0.019 (0.014 to 0.026)	0.546
Systolic blood pressure (mmHg) [2]	0.030 (0.026 to 0.035)	0.019 (0.014 to 0.026)	0.626
Diastolic blood pressure (mmHg) [2]	0.039 (0.034 to 0.045)	0.0297 (0.024 to 0.037)	0.760
BMI (kg/m^2) [2]	0.022 (0.018 to 0.026)	0.012 (0.009 to 0.017)	0.556
Total cholesterol (mmol/L) [2]	0.021 (0.018 to 0.025)	0.010 (0.007 to 0.016)	0.492
HDL cholesterol (mmol/L) [2]	0.021 (0.018 to 0.025)	0.0186 (0.015 to 0.023)	0.870

WPC: Within-period correlation. IPC: Inter-period correlation. CA: Cluster autocorrelation. CI: Confidence interval. BMI: Body mass index [1]: Two consecutive 12-month periods were used (01/01/2007 – 31/12/2007 & 01/01/2008 – 31/12/2008). [2]: Two consecutive 15-month periods were used.

4.3.3.1 *Impact of study length on the IPC and WPC*

Table 4.8 highlights the impact of study length on the WPC and IPC for HbA1c%. Here, a fixed 1-year period length was used, and the overall study length was changed. The results show that increasing the study length does not affect the IPC, whereas the WPC increased slightly (0.035 to 0.041) – which is reflected in the CA which decreased as the study length increased (0.5462 to 0.4366). That is, with a fixed period-length, increasing the study length does affect the correlation between participants in the same cluster at different time-periods, but increases the correlation between participants in the same cluster at the same time-period. The results are predominantly similar when adjusting for covariates (Table 9.7 in Appendix I).

Table 4.8: Impact of increasing the study length on the within-period correlation and inter-period correlation for HbA1c from included participants from THIN database when maintaining a one-year period length

Study length (years)	Number of periods	WPC (95% CI)	IPC (95% CI)	CA
2	2	0.035 (0.030 to 0.040)	0.019 (0.014 to 0.026)	0.546
3	3	0.038 (0.033 to 0.045)	0.018 (0.014 to 0.025)	0.480
4	4	0.040 (0.039 to 0.047)	0.019 (0.014 to 0.025)	0.468
5	5	0.040 (0.034 to 0.048)	0.018 (0.013 to 0.024)	0.441
6	6	0.041 (0.035 to 0.048)	0.018 (0.013 to 0.024)	0.437

WPC: Within-period correlation. IPC: Inter-period correlation. CA: Cluster autocorrelation. CI: Confidence interval.

4.3.3.2 *Impact of time-period length on the IPC and WPC*

The impact of period length on the WPC, IPC, and CA is presented in Table 4.10. Here, the 15-month study period is split into 1-month, 3-month, 3.75 month, 5-month, and 7.5-month time-periods. The median WPC for a 7.5-month period was 0.023, with a median IPC of 0.018. The WPC and IPC increased as the period length decreased. That is, for a fixed study length, decreasing the period length leads to an increase in the correlation between two observations from the same cluster and the same time-period – the WPC – and an increase in the correlation between two observations from the same cluster and at different time-period – the IPC. The CA remains relatively high for all outcomes, highlighting that there are only small differences between the IPC and WPC for all period lengths. This would indicate that for this 15-month study period; there are only minimal period effects. The WPC, IPC, and CA generally remained unaffected when adjusting for covariates (Table 9.8 in Appendix I).

Table 4.9: Impact of period length on the within-period correlation and inter-period correlation for continuous outcomes associated with type-2 diabetes from THIN database for study period 01/02/2009 to 30/04/2010 from an unadjusted model

Period Length	WPC (95% CI)	IPC (95% CI)	CA
7.5 month period length			
Systolic blood pressure (mmHg)	0.040 (0.035 to 0.045)	0.016 (0.011 to 0.022)	0.395
Diastolic blood pressure (mmHg)	0.040 (0.035 to 0.046)	0.036 (0.030 to 0.042)	0.886
BMI (kg/m^2)	0.022 (0.019 to 0.026)	0.016 (0.013 to 0.020)	0.742
Total cholesterol (mmol/L)	0.021 (0.018 to 0.025)	0.018 (0.015 to 0.022)	0.855
HDL cholesterol (mmol/L)	0.023 (0.020 to 0.027)	0.019 (0.016 to 0.024)	0.846
Median	0.023	0.018	0.846
5 month period length			
Systolic blood pressure (mmHg)	0.042 (0.037 to 0.047)	0.021 (0.017 to 0.027)	0.512
Diastolic blood pressure (mmHg)	0.041 (0.035 to 0.047)	0.037 (0.032 to 0.043)	0.911
BMI (kg/m^2)	0.024 (0.020 to 0.027)	0.016 (0.013 to 0.020)	0.699
Total cholesterol (mmol/L)	0.021 (0.018 to 0.025)	0.019 (0.016 to 0.023)	0.881
HDL cholesterol (mmol/L)	0.024 (0.021 to 0.028)	0.020 (0.017 to 0.024)	0.826
Median	0.024	0.020	0.826
3.75 month period length			
Systolic blood pressure (mmHg)	0.043 (0.038 to 0.048)	0.023 (0.019 to 0.028)	0.538
Diastolic blood pressure (mmHg)	0.040 (0.035 to 0.047)	0.038 (0.032 to 0.044)	0.933
BMI (kg/m^2)	0.023 (0.020 to 0.027)	0.018 (0.015 to 0.021)	0.748
Total cholesterol (mmol/L)	0.022 (0.019 to 0.026)	0.019 (0.016 to 0.022)	0.852
HDL cholesterol (mmol/L)	0.025 (0.021 to 0.029)	0.020 (0.017 to 0.024)	0.819
Median	0.025	0.020	0.819
3 month period length			
Systolic blood pressure (mmHg)	0.044 (0.039 to 0.049)	0.024 (0.020 to 0.029)	0.543
Diastolic blood pressure (mmHg)	0.041 (0.035 to 0.047)	0.038 (0.033 to 0.044)	0.928
BMI (kg/m^2)	0.025 (0.022 to 0.029)	0.018 (0.015 to 0.021)	0.703
Total cholesterol (mmol/L)	0.022 (0.019 to 0.026)	0.019 (0.016 to 0.023)	0.866
HDL cholesterol (mmol/L)	0.026 (0.022 to 0.030)	0.020 (0.017 to 0.025)	0.802
Median	0.026	0.020	0.802
1 month period length			
Systolic blood pressure (mmHg)	0.046 (0.042 to 0.052)	0.028 (0.023 to 0.032)	0.593
Diastolic blood pressure (mmHg)	0.043 (0.037 to 0.049)	0.038 (0.033 to 0.044)	0.897
BMI (kg/m^2)	0.027 (0.024 to 0.032)	0.018 (0.015 to 0.022)	0.674
Total cholesterol (mmol/L)	0.022 (0.019 to 0.026)	0.020 (0.016 to 0.023)	0.880
HDL cholesterol (mmol/L)	0.026 (0.022 to 0.030)	0.021 (0.018 to 0.025)	0.829
Median	0.027	0.021	0.829

WPC: Within-period correlation. IPC: Inter-period correlation. CA: Cluster autocorrelation. CI: Confidence interval.

4.4 Discussion

The key difference between a CRT and a RCT is the correlation between observations within a cluster – denoted as ρ – which is needed for pre-trial sample size calculations. In the absence of period effects, the ICC is sufficient to denote the correlation between patients in a cluster – for which there is a lack of reporting of in post-trial analysis (73). The lack of clarity of the ICC for binary outcomes has led us to differentiate between the latent ICC (from a logistic model) and a natural ICC (from a linear model). Nonetheless, the decay of this correlation structure over time may be equally important (27, 31) – leading us to the WCC, IPC, and WPC. This chapter has highlighted the methodology required to estimate the IPC (alongside the WPC and CA) in addition to the ICC, and reported estimates of these correlations using data from The Health Improvement Network for outcomes associated to type-2 diabetes. It has shown that in longitudinal studies, researchers should consider time-dependent models to fully depict the correlation between participants over multiple time-periods.

4.4.1 Intra-cluster correlation coefficient

When publishing a CRT, it is recommended that the ICC used in the sample size calculation and the ICC relating to the primary outcome in the analysis are reported (70). However, the number of studies that comply to this is still sub-optimal (73). To ensure a CRT is sufficiently powered, trialists rely heavily on accurate estimates of the ICC being available. A substandard quality of post-trial reporting of the ICC will inevitably lead to trialists using ad-hoc estimates. When considering a continuous outcome, the estimation of the ICC – as a ratio of variances – is relatively straightforward and the formula is well established (102).

4.4.1.1 *Dichotomous outcomes*

The differing approaches to estimating the ICC for a dichotomous outcome has led to previous methodological articles and trial reports using both natural ICCs (estimated from a linear model) and latent ICCs (from a logistic model) (105, 106, 108-111, 134-137). There is often a lack of clarity as to whether a natural ICC or latent ICC has been presented.

A latent ICC refers to an unobservable latent scale, opposed to the correlation between observations within a cluster – which is needed for a sample size calculation. In its raw form, the latent ICC is presented on a log-odds scale – rather than the natural scale used by the natural ICC (105, 108) – and so must be converted into a natural ICC for use in a sample size calculation. A sample of conversions has been presented by Eldridge et al. (102).

The results of this chapter are similar to previous research, highlighting that the natural ICC is generally smaller than the latent ICC (20). Furthermore, it has also been established that dichotomous outcomes with small prevalence's will have small natural ICCs (105, 138, 139). In light of this, clinical outcomes (e.g. stroke) which have low prevalence are likely to have small natural ICCs. In contrast, clinical measures dichotomised using clinically relevant cut-points will have a higher prevalence and so will have larger natural ICCs. The results of this chapter agree with existing research in this area (106, 140). One of the other noticeable differences between the latent ICC and natural ICC observed in this chapter was the increased variation in the values of the latent ICC. This is perhaps because ICCs follow the same pattern as proportions, whereby there is a greater scope for variability in values as they increase towards 0.5. As such, since the latent ICC is larger than the natural ICC, it also has a greater scope for variability in its values.

Although the prevalence influences the natural ICC – which may make the latent ICC preferable (139) – since the natural ICC refers to the correlation between observations within a cluster, it should be reported. If a reported natural ICC reflects a prevalence different to a planned trial, it could be converted to a latent ICC before being back transformed to a natural ICC using the appropriate prevalence (102). Reporting of ICCs should be presented on the natural scale for continuity with continuous outcomes.

4.4.1.2 *Alternative databases of intra-cluster correlation coefficients*

In some settings, it may be relatively simple to obtain an ICC from a previously published trial. Yet, in many settings, this may not be possible. The concept of using a large database of patient information to estimate ICCs is increasing in practice. One such database that has been established is the Health Service Research Unit, who have combined ICCs from previous trials and databases (141). ICCs are presented for a variety of continuous and dichotomous outcomes using GP surgery, hospital, and physician as the level of clustering. When limiting to only GP surgery as the cluster level, the mean ICC was 0.042 (SD: 0.042) for continuous outcomes and 0.064 (SD: 0.062) for dichotomous outcomes. Since the ICCs are collated from multiple sources, it is unclear whether they refer to latent ICCs, natural ICCs, or a combination of both. The ICCs were generally lower for continuous outcomes than dichotomous outcomes, which may suggest logistic regression was used to estimate a latent ICC.

4.4.1.3 *Confidence interval for intra-cluster correlation coefficients*

For all estimates of the ICC, a 95% confidence interval is reported. This confidence interval is calculated using a logit transformation to the ICC to ensure that the interval is contained between zero and one. Nevertheless, it is possible to calculate a confidence interval for the ICC on the natural scale (142). Standard statistical packages (e.g. Stata) report as default, a confidence interval that has been calculated using a logit transformation. As such, it may not be appropriate to simple report an estimate of the standard error of the ICC, since this value may be used by future trialists assuming that the ICC is on the natural scale, and not on the logit scale. It is possible that alternative methodology should be implemented when estimating a confidence interval for an ICC when fitting a model that contains covariate adjustment (142).

4.4.2 Correlation in longitudinal studies

In longitudinal CRTs, there is a necessity to describe the decay of correlation over time within a cluster. This can be done using the IPC and WPC. Currently, there is little or no empirical literature to inform likely values for these parameters at the design stage (27, 48). As such, many sample size formulae assume that they are equal (32). Here, we have shown that for all outcomes, there was evidence of decay of correlation over time – indicating the presence of a period effect that should be accounted for in the design of a CRT.

As the IPC and WPC are parameters to describe a time dependent correlation, it is likely that both the study length and the period length will influence their values. This is the first piece of research to investigate what impact the study and period duration has on the IPC and WPC.

Increasing the study length (whilst maintaining the period length) leads to a slight decrease in the IPC. Increasing the study duration would lead to observations that can now originate from further apart in time, and so it is likely that these observations would be less correlated than observations made closer together in time. Furthermore, the CA becomes smaller as the study length increases – highlighting a greater disparity between the IPC and WPC. Perhaps unexpectedly, the WPC increases slightly as the study length increases, despite maintaining the period length.

Decreasing the length of the period leads to an increase in both the IPC and WPC. As observations will now stem from a shorter period of time, it is expected that they will be greater correlated – highlighted by the increasing WPC. Consequently, observations from different periods can be relatively closer together in time if the period length is shorter – reflected in the greater IPC for smaller study period lengths. The impact of altering the period length affects the IPC and WPC relatively equally, and so the CA is not greatly affected by changes in period length.

A summary of these key results is presented below (Table 4.10). Since it is known that SW-CRTs are more efficient for greater values of the ICC (37, 61), it is likely that they will perform better for greater values of the WPC and IPC. It is therefore likely that decreasing the period length (where possible) will lead to an increase in power.

Table 4.10: Impact of increasing study and period length on the inter-period correlation, within-period correlation, and cluster autocorrelation

Parameter	Increasing study length	Increasing period length
Inter-period correlation	Decreases	Decreases
Within-period correlation	Increases	Decreases
Cluster autocorrelation	Decreases	Remains the same

The decay of correlation over time has previously been reported as a cluster autocorrelation (48). The CA directly influences the sample size in studies employing a cluster cross-over design (112). Expected values of the CA are not readily available, but an expected value of 0.8 has been reported previously (48). However, within the SW-CRT, current formulae assume a CA of 1 (32), which may be an underestimate of the decay of correlation over time.

Within the study designs highlighted here, it has been shown that the study design has vital influences on the estimates IPC and CA. Failure to acknowledge the IPC or the CA in the sample size calculation of a study with a repeated cross-section design may produce an overestimate of the power of the study (48), else lead to an incorrect estimate of the required number of clusters for a CRT (27) .

The methodology for estimating the power in a SW-CRT has been established (9, 32, 45) and is discussed in detail elsewhere (see Chapter 6). However, this methodology considers a time-independent correlation, and so does not currently allow for the inclusion of the IPC or CA when estimating the power in a SW-CRT. Furthermore, it is not known what impact their inclusion would have and whether it would be beneficial or detrimental to power. Simulation methods may provide a solution to include this additional correlation.

As a minimum, if the IPC differs to the WPC, then studies should ensure that the estimate of ρ for use in a DE (equation 4.1) stems from the WCC (estimated via equation 4.5), and not from the ICC (estimated via equation 4.2).

4.4.3 Future research

It has been well established that the ICC is vital for sample size calculations for CRTs. Now, it is becoming increasingly recognised that the IPC and CA are vital for designs with repeated cross-sections. In P-CRTs, the decay of clustering over time is typically ignored. However, this deterioration of correlation should be acknowledged within sample size calculations. As the correlation deteriorates over time, a P-CRT naturally increases in power. Recognising this decay in the framework will therefore increase power if period effects are acknowledged in a P-CRT. However, there has been no research previously into the degree of this increase.

4.4.4 Limitations

Using routine data from a large patient database allows for the estimation of the correlation between patients for a multitude of outcomes. However, some limitations are present when using data in this form. One such example is the difficulty in establishing the difference between follow-up care for the first instance of a clinical event (e.g. stroke) from a second event. Patient databases rely on the coding of events, and it may be challenging to distinguish between events coded in an identical manner without further consultation. To prevent confusion and to eradicate this problem, all patients whom had suffered an event prior to the study inclusion period were excluded from the analysis for that event. Naturally, patients excluded may have suffered a second (or more) event during the inclusion period, but to ensure accurate results consistency, they were excluded. Only the first instance of an event was considered. The reliance of a coding system can also lead to misclassification of type-2 diabetes rather than type-1 diabetes. The minimum age of 18 was used to minimise this risk.

The Health Improvement Network data stems from patient information collated from GP surgeries. As such, only data recorded by the practices is available. When considering a model framework that includes covariate adjustment, the choice of covariates is limited by the reporting of each practice. The quality of reporting and coding is likely to differ between practices. This can also create inconsistencies in terms of the recording of clinical measurements, and the degree of patient monitoring. The inclusion of a minimum acceptable mortality reporting date was done to ensure practice level reporting was of a sufficient standard.

The study period length used in this analysis ranged from 15 months to 72 months. In the context of a P-CRT, the 15 month closed cohort may be a shorter period than is normal for a trial in diabetes. The extension up to 72 months was considered in the context of the SW-CRT, but only for HbA1c.

In the analysis of a trial, the primary aim is to assess whether a treatment performs differently to a control – usually performed by including a fixed effect for the treatment arm. When estimating ICCs from this model, the treatment arm is a covariate of interest. A large patient database can allow for the estimation of ICCs. However, adjustments cannot be made for treatment arm as there is not one present. As such, the model framework differs between analyses used on a large patient database to that used in a trial. In some settings, the intervention may influence the degree of clustering, which would create an ICC that differs in the treatment arms. The differing degree of clustering in the trials arms has previously been described as differential clustering (143, 144). Should the intervention influence the ICC, it is not known whether an ICC obtained from patient data would lead to an accurate estimate of the sample size required.

The estimates of the IPC and WPC in this chapter are limited to continuous outcomes in a cross-sectional study framework. Many SW-CRTs employ a cohort design and consist of dichotomous outcomes (see Chapter 3). The estimation of correlation for dichotomous outcomes is more complex than for continuous outcomes, due to the necessary transformation from a latent scale to a proportion scale. The use of a cohort design would require an additional random effect for the within-participant correlation over time, and so would provide added complexity.

4.5 Conclusions

In CRTs, the ICC is required for an accurate sample size calculation (54, 94, 102, 140). In CRTs with repeated cross-sections – such as a SW-CRT – observations are taken over multiple time-periods. In such designs, it is possible that two observations within the same cluster and from the same time-period will be greater correlated than two observations from within the same cluster but from different time-periods. As such, the ICC may not be sufficient to describe the correlation, and additional correlation types are required – the IPC, the WPC, and the WCC. The WPC is defined as the correlation between two observations within a cluster and within the same time-period. The IPC is defined as the correlation between two observations within the same clusters but from different time-periods. Together the IPC and WPC can be used to calculate the WCC – the correlation between two random observations in a cluster – and give an indication of the decay of correlation over time. The IPC and WPC are becoming increasingly recognised as important to SW-CRTs and other longitudinal designs (26, 48). Nevertheless, there has been no evidence of the publication of any empirical values, and so it has been repeatedly reported that a ratio of the IPC to the WPC can be expected to be 0.8 (27, 48). This chapter is the first reporting of the IPC and WPC and shows that the ratio of the IPC to the WPC may be smaller in some instances than previously expected. This emphasises that SW-CRTs should acknowledge the decay of correlation over time. Furthermore, the results of this chapter are a useful resource that be used in future methodological work, and in the sample size calculation of future CRTs.

The methodological review (Chapter 3) highlighted that no published SW-CRTs had acknowledged the IPC and WPC in the sample size calculation. The results of this chapter have shown that the ICC alone may not be sufficient to describe the correlation within a cluster and that the IPC and WPC should be accounted for. The review also highlighted that varying cluster size is poorly reported in pre-trial sample size calculations and no study has used appropriate methods to adjust for varying cluster size. This may stem from a lack of research into unequal cluster sizes in a SW-CRT. In the next chapter, we present a simulation study to illustrate the impact of varying cluster size in a SW-CRT.

CHAPTER 5: THE IMPACT OF VARYING CLUSTER SIZE IN STEPPED-WEDGE CLUSTER RANDOMISED TRIALS

5.1 Introduction

A large focus of the methodological research in cluster randomised trials (CRTs) centres on the assessments of power (52, 57, 74, 145-147). Equally important is the precision of the treatment effect estimate, which is independent of the treatment effect estimate, but is used to derive the power. The precision indicates how wide the confidence interval for a treatment effect estimate will be – the larger the precision, the narrower the confidence interval. It has been established that the correlation between observations within a cluster impact the precision (and power) in a parallel CRT (P-CRT) (55). Also associated with the precision is the variability in the cluster sizes (52, 148). Often, CRTs are planned supposing that an equal number of participants are present in each cluster (equal cluster size) – which is evident in the methodological review in Chapter 3. Nevertheless, without limiting the number of participants per cluster, it is unlikely in practice that each cluster will contain the same number of participants. As such, the majority of CRTs contain unequal sized clusters (58, 146).

In P-CRTs, varying cluster size can be accounted for in the design stage through an adjusted design effect (DE), examples of which have been presented earlier (Chapter 2). However, these conventional methods may be conservative (57), and always assume a fixed value of the precision by assuming a balance in the number of observations in the control and intervention arms (60, 62). Nevertheless, it has been shown that the degree of variation in cluster size has an inverse relationship with the precision (57, 145) – so that the greater the degree of variation in cluster sizes, the smaller the precision (62).

Current DEs for SW-CRTs (21, 26) are only applicable for designs with equal sized clusters – and so may not be appropriate when clusters vary in size. However, the matrix approach shown by Hussey and Hughes (32) can be used to estimate the power for varying cluster size if the cluster sizes are known, but this has not been investigated. The absence of an appraisal of varying cluster size in a SW-CRT consequently implies that it is unknown whether the SW-CRT is more (or less) robust than a P-CRT when the cluster sizes vary.

When clusters vary in size in a SW-CRT, the randomisation order of the clusters is likely to affect the precision – since a different number of observations will be contributing to the intervention and control conditions. As such, the precision is likely to be better represented as a distribution of values, rather than a single value. This thought has led us to consider whether the precision in a P-CRT should also be represented as a distribution of values, rather than a single value – which is usually assumed.

5.1.1 Chapter aim

The main aim of this chapter is to investigate the impact of between-cluster variability in a SW-CRT. Results will also be compared to a P-CRT to assess whether a P-CRT and a SW-CRT are affected in the same manner by varying cluster size. Our objectives are split into four sections:

1) Examine the precision in a SW-CRT with between-cluster variability in size and compare it to the precision of a SW-CRT with equal cluster size to highlight:

    a) The average precision in a design with unequal cluster sizes.

    b) The variation in precision in a SW-CRT with unequal cluster size.

    c) The impact of the design on the precision (i.e. number of steps, number of clusters, etc.).

2) Examine the precision in a P-CRT with between-cluster variability in size and compare it to the precision of a P-CRT with equal cluster size to assess:

    a) The average precision in a design with unequal cluster sizes.

    b) The variation in precision in a SW-CRT with unequal cluster size.

    c) The impact of the design on the precision (i.e. number of steps, number of clusters, etc.).

3) Investigate how a SW-CRT with varying cluster size compares to a P-CRT with varying cluster size.

4) Examine the precision in a SW-CRT with between-cluster variability in size and within-cluster variability in size over time and compare it to the precision of a SW-CRT with equal cluster size to evaluate:

    a) The average precision in a design with unequal cluster sizes.

    b) The variation in precision in a SW-CRT with unequal cluster size.

c) The impact of the design on the precision (i.e. number of steps, number of clusters, etc.).

In this chapter, we seek to fully assess the influence of varying cluster size in a SW-CRT, and how this compares to a P-CRT. To do this, we begin with the analytical framework used to analyse P-CRTs and SW-CRTs, and describe how the precision of a treatment effect estimate can be estimated for both trial designs. We then describe a simulation study that allows the precision in a SW-CRT to be estimated when clusters vary in size. The precision is then compared back to a SW-CRT with equal cluster size to form a relative efficiency. We investigate the impact of unequal cluster size on the average relative efficiency. It is then discussed whether an average value of the relative efficiency is sufficient, or whether the full distribution of possible efficiency values is required. An extension to the SW-CRT framework is also presented, that considers both between-cluster variation in size and within-cluster variation in size over time. Throughout this chapter, we are assuming that the SW-CRT and P-CRT follow a cross-sectional design, and so participants contribute only one observation to the study.

5.2 Methods

In this section, we present established methodology used in the design and analysis of P-CRTs and SW-CRTs and discuss how they can be extended to evaluate the impact of cluster size variation in a SW-CRT. Models used in the analysis of P-CRTs and SW-CRTs are presented, alongside the methods used to estimate the precision of the treatment effect estimate. Different types of variation in cluster size are described alongside the necessary adaptations to estimate the precision when these variations are present. We then discuss the simulation process and the factorial design that is used to evaluate the impact of cluster size variation on the efficiency of a SW-CRT.

5.2.1 Analytical model used to analyse SW-CRTs

In a SW-CRT, clusters exposed to the intervention provide, on average, observations at a later period of time. As such, calendar time must be accounted for in the model – by adopting a fixed effect for each time-period, independent of the cluster. To account for correlation between participants within a cluster, a random effect for cluster is included. A model has been proposed by Hussey and Hughes (32) to analyse a SW-CRT, given as:

$$Y_{ijk} = A + \beta_j + X_{ij}\delta + \alpha_i + \varepsilon_{ijk} \qquad \mathbf{5.1}$$

Where, Y_{ijk} is the outcome for participant *k* in cluster *i* at time *j*, *A* is the mean outcome in the unexposed period in the first time-period, β_j is a time effect, fixed for time-periods j = 2,…, T ($\beta_1 = 0$ for identifiability), δ is the treatment effect, α_i is a random effect for cluster *i* $\alpha_i \sim N(0, {\sigma_b}^2)$, ε_{ijk} is the residual error $\left(\sim N(0, {\sigma_w}^2)\right)$, and X_{ij} is an indicator of treatment, where:

$$X_i \begin{cases} 1 & \text{if cluster } i \text{ is exposed to the intervention at time } j \\ 0 & \text{if cluster } i \text{ is not exposed to the intervention at time } j \end{cases}$$

When using this framework, there are some assumptions that are being made:

i. Constant correlation over time:

That is, two observations from within the same cluster will have the same correlation, regardless of the time-period in which the observations are made. That is, it is assumed the cluster-by-period variance component discussed in Chapter 4 is zero, and so the ICC is sufficient to describe the correlation between observations within a cluster.

ii. Single layer of clustering:

Only one level of clustering is included – participants within clusters. We do not discuss the framework for multi-level clustering designs.

iii. The full intervention effect is realised in the same time interval in which the intervention is introduced:

This implies that there is no delay in the treatment effect and so the effect is visible immediately.

iv. Common secular trend across clusters:

Each cluster follows the same underlying trend in the outcome over time.

5.2.2 Estimating precision in a SW-CRT

In any trial, a measure of the uncertainty surrounding a treatment effect estimate should be presented (68, 149). The precision – the inverse of the variance of the treatment effect estimate – is one such measure of uncertainty. As the variability in the estimate decreases, the estimate becomes more accurate (precise) and the precision increases. The precision is independent of the treatment effect, and reflects only the study design, allowing direct comparisons of differing study designs without the requirement of a treatment effect. Before formulae are presented to estimate the precision for a SW-CRT, we outline some terminology.

The design of a SW-CRT can be depicted by a design pattern matrix (DPM). This matrix indicates whether a cluster is in the control (0) or intervention (1) for any given time-period. An example of a DPM for a SW-CRT with 3 clusters (C_1, C_2, and C_3) and 4 time-periods (t_1, t_2, t_3, and t_4) is presented in Figure 5.1.

Figure 5.1: Example design pattern matrix

$$\begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

This DPM can be converted into a design matrix, denoted as Z. This matrix contains $T + 1$ columns, and $C \times T$ rows, where T is the number of time-periods, and C the number of clusters. The first column indicates whether a cluster is in the control (0) or intervention (1), and the remaining T columns indicate the time-period that is being referred to (1 if yes, 0 if no). The $C \times T$ rows can be split into C groups of T rows, where each row refers to a particular cluster at one time-period. For the DPM given in Figure 5.1, the design matrix *Z* can be given by Figure 5.2. In this example, row 7 indicates that the second cluster (since it's in the second set of C rows) is in the intervention (since the first column is a 1) during the third time-period (since the 1 in the remaining T columns correspons to t_3).

Figure 5.2: Example matrix *Z*

$$Z = \begin{array}{c} \begin{matrix} I & t_1 & t_2 & t_3 & t_4 \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Now, we refer to matrix *V*, which is the variance-covariance matrix of the cell means ordered by time within cluster (9, 32). V is a $CT \times CT$ block diagonal matrix, made up *C* blocks – each called V_i and each $T \times T$ in size. Here each $T \times T$ block refers to a particular cluster (*i*) and describes the correlation between the cluster means over time, taking the form:

Figure 5.3: Example matrix V$_i$

$$V_i = \begin{pmatrix} \frac{\sigma_w^2}{m_{i1}} + {\sigma_b}^2 & {\sigma_b}^2 & \cdots & {\sigma_b}^2 \\ {\sigma_b}^2 & \frac{\sigma_w^2}{m_{i2}} + {\sigma_b}^2 & \cdots & {\sigma_b}^2 \\ \vdots & \vdots & \ddots & \vdots \\ {\sigma_b}^2 & {\sigma_b}^2 & \cdots & \frac{\sigma_w^2}{m_{iT}} + {\sigma_b}^2 \end{pmatrix}$$

Here m_{ij} represents the size of cluster i at time j which we refer to as the cluster-period size.

Matrix V can then be summarised as a $C \times C$ block diagonal matrix (Figure 5.4), in which each entry V_i is itself a $T \times T$ matrix given by Figure 5.3:

Figure 5.4: Example matrix V

$$V = \begin{pmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_C \end{pmatrix}$$

The zeroes in this matrix refer to the independence of participants between clusters

If assuming the random effects are known, when fitting model 5.1, there are T unknown parameters – the intervention effect (δ) and a time effect for each time-period (excluding the first period for identifiability) $(\beta_2, \beta_3, \dots, \beta_T)$. These estimates can be obtained from a weighted least squares approach (32). The variance-covariance matrix for the estimated values of these parameters can be calculated as $(Z'V^{-1}Z)^{-1}$. The variance of the treatment effect estimate is given by the entry in the first column and first row (denoted as [1, 1]) of this variance-covariance matrix. As such, the precision of the treatment effect estimate can be calculated as:

$$\varphi = \frac{1}{(Z'V^{-1}Z)^{-1}[1,1]} \quad 5.2$$

When designing a study, the precision is vital for two reasons, as it influences: (i) how the study compares to other study designs, and (ii) what the estimate of the power is. In this setting, we are interesting in (i) and, more specifically, how a CRT is affected by varying cluster size. The precision allows a direct comparison between study designs, without the prerequisite of a treatment effect estimate that is necessary to estimate power. Here, the precision can be used to compare a CRT with unequal cluster size to a CRT with equal cluster size when the total sample size is fixed. As such, the relative efficiency (RE) of the design is used here, given as the ratio of the precision for a design with unequal cluster size to the precision of a design with equal cluster size:

$$RE = \frac{Precision_{unequal}}{Precision_{equal}} \quad 5.3$$

5.2.2.1 *Types of variation in cluster sizes in a SW-CRT*

In a SW-CRT, time is split into a number of (equal) time-periods, and so the size of a cluster can be denoted by the total cluster size (size of a cluster over the whole study) (M) or by a cluster-period size (size of a cluster at one time-period) (m_{ij}). When referring to cluster size variation, there are two forms that this can take: (1) between-cluster variation in size; and (2) within-cluster variation over time.

The main source of variation will likely stem from the variation between clusters. That is, the number of participants will differ cluster to cluster, often as a result of logistical limitations

(57) – such as each cluster containing a different number of beds which limits the maximum participants for that cluster, such as in the ERFIC trial (see section 1.6.3) (25). The differing in size between clusters is described as between-cluster variation in size.

Since a SW-CRT is split into time-periods, it is possible that the number of participants in a cluster will vary over time. This could be due to fluctuations in work load or recruitment. This is referred to as the within-cluster variation in size over time.

Therefore, cluster size variation can be in the form of between-cluster variation or within-cluster variation (or both). To begin, we consider only the between-cluster variation in size, and assume that there is no within cluster-variation in size (i.e. within a cluster, there is no variation in size over time) – since it is likely that between-cluster variation will account for the majority of the variation. The methodology will then be extended to include within-cluster variation in size over time.

In the discussion of cluster sizes, we have previously referred to the cluster-period size. To establish the terminology and prevent confusion, a summary of the terms and notations used throughout are given in Table 5.1.

Table 5.1: Summary of key terminology used in reference to cluster size

Term	Notation	Description
Design matrix	(Matrix) Z	The design of a SW-CRT, indicating whether clusters are in the control or intervention at each time-period
Variance-covariance matrix	(Matrix) V	The variance-covariance of the cell means ordered by time within cluster
True cluster-period size	μ	The (known) average cluster-period size – used to simulate all other cluster-period sizes
Coefficient of variation	cv	The (known) coefficient of variation in cluster sizes – calculated as the ratio of the standard deviation of cluster sizes to the mean cluster size
Cluster-period size	m_{ij}	The (simulated) size of cluster *i* at time *j*
Average cluster-specific cluster-period size	$\overline{m}_{i\bullet}$	The average (simulated) size of cluster *i* over all time-periods
Average cluster-period size	$\overline{m}_{\bullet\bullet}$	The average (simulated) size of a cluster at one time-period

Throughout this work, we use the notation of "dot-bar", so that the following hold:

$$\overline{m}_{i\bullet} = \frac{1}{T}\sum_{j=1}^{T} m_{ij} \qquad\qquad \overline{m}_{\bullet\bullet} = \frac{1}{CT}\sum_{i=1}^{C}\sum_{j=1}^{T} m_{ij}$$

Where $\overline{m}_{i\bullet}$ is the average cluster-period size for cluster i, and $\overline{m}_{\bullet\bullet}$ is the average cluster-period size. Both are calculated from the simulated cluster-period sizes (m_{ij}).

Throughout this work, μ refers to a value that is being used to simulate cluster-period sizes. The term m_{ij} will refer to the simulated cluster-period sizes. The terms $\overline{m}_{i\bullet}$ and $\overline{m}_{\bullet\bullet}$ refer to an average, which is estimated from a set of simulated values.

Below, we discuss the steps taken to evaluate the impact of between-cluster variation in size on the precision of a SW-CRT and a P-CRT, and assess the effect of within-cluster variation on the precision of a SW-CRT.

5.2.3 Impact of between-cluster variation in size in a SW-CRT

In the design stage of a CRT, a measurement of the total cluster size (number of participants in a cluster over the whole study period) and the cluster-period size (number of participants in a cluster during one time-period) are likely to be known. If the clusters vary in size, then an average total cluster size or average cluster-period size may be known, alongside a measure of dispersion – such as the coefficient of variation of cluster sizes.

The impact of varying cluster size in a SW-CRT can be shown by comparing the precision of a SW-CRT with unequal cluster sizes to the precision of a SW-CRT with equal cluster sizes, when both SW-CRTs have the same total sample size. If the sample size differs between the designs, any change in efficiency may due to a difference in sample size, opposed to the effect of clusters varying in size. This also ensures that each estimate of the RE is from a SW-CRT with an identical sample size – and so the distribution of the RE and any loss in average RE will be due to varying cluster size.

To this end, we simulate cluster-period sizes that vary for each of C clusters at each of T time-periods, and calculate the precision for the simulated cluster-period sizes. Alongside this, the precision is estimated for a SW-CRT with clusters of equal size, and of the same total sample size. The precision of the unequal design and equal design are then compared as a relative efficiency (RE). Below, in sections 5.2.3.2 to 5.2.3.6, we discuss the steps taken to simulate cluster-period sizes and estimate the RE. Prior to that, an example is presented to illustrate the outline of the methods, and then each step is discussed in more detail.

5.2.3.1 *Example of estimating the relative efficiency of a SW-CRT with unequal cluster sizes*

We begin with a SW-CRT with equal cluster sizes, with 12 clusters, 6 time-periods, and a true cluster-period size $\mu = 50$. Since $m_{ij} = m_{i`j`}\ \forall\ i, i'$ and j, j', it follows that $m_{ij} = m_{i'j'} = 50\ \forall\ i, i'$ and j, j', and so the sample size in the SW-CRT with equal cluster sizes (S_e) will be:

$$S_e = \mu \times C \times T = 50 \times 12 \times 6 = 3600.$$

Now, consider a SW-CRT with unequal cluster size, with 12 clusters, 6 time-periods, a true cluster-period size $\mu = 50$ and a coefficient of variation of cluster sizes, $cv = 0.5$. Now, we assume that the average cluster-specific cluster-period sizes $(\overline{m}_{i\bullet})$ follow a Gamma distribution, and so we use μ and cv to simulate the average cluster-specific cluster-period size $(\overline{m}_{i\bullet})$ for each cluster i. A set of potential $\overline{m}_{i\bullet}$ values are:

$$\begin{array}{llll}
m_{1\bullet} = 47.10 & m_{2\bullet} = 68.90 & m_{3\bullet} = 33.82 & m_{4\bullet} = 56.24 \\
m_{5\bullet} = 47.44 & m_{6\bullet} = 71.67 & m_{7\bullet} = 26.81 & m_{8\bullet} = 83.26 \\
m_{9\bullet} = 28.17 & m_{10\bullet} = 12.09 & m_{11\bullet} = 63.87 & m_{12\bullet} = 60.64
\end{array}$$

These values indicate the average cluster-period size for that cluster, so that $m_{1\bullet} = 47.10$ indicates that the average size for cluster 1 at one time-period is 47.10. The values of $m_{i\bullet}$ can then be used to obtain the size of a cluster at a particular time-period – the cluster-period size $\left(m_{ij}\right)$. Since there is no within-cluster variation, each cluster is the same size at each time period, so that $m_{ij} = m_{ij'}\ \forall\ j, j'$. Now, to estimate the RE, the sample size should be identical in the SW-CRT with unequal cluster size and the SW-CRT with equal cluster size. In a SW-CRT with unequal cluster sizes, the total sample size S_u is the sum of all of the cluster-period sizes m_{ij}, so that:

$$S_u = \sum_{i=1}^{C} \sum_{j=1}^{T} m_{ij}.$$

Now, under the assumption of $S_u = S_e$, we calculate the precision for the SW-CRT with unequal cluster sizes and the precision for the SW-CRT with equal cluster sizes using equation 5.2 by replacing the m_{ij} values in each V_i matrix. Following this, the RE is calculated as the ratio of the two precisions.

The steps taken to estimate the RE can be broken down into five sections: simulating cluster-period sizes, scaling cluster-period sizes to maintain the total sample size, estimating the precision, estimating the relative efficiency, and compiling the relative efficiency estimates. Each step is discussed in more detail below.

5.2.3.2 *The simulation of cluster sizes*

Since the cluster-period sizes are unknown, we use the true cluster-period size (μ) and the coefficient of variation of cluster sizes (cv) to simulate potential cluster-period sizes $\left(m_{ij}\right)$. We simulate the average cluster-specific cluster-period size, $\overline{m}_{i\bullet}$, for each cluster *i*, using a Gamma distribution with mean μ and coefficient of variation cv, so that the following holds:

$$\overline{m}_{i\bullet} \sim \Gamma(\alpha, \beta) \qquad E(\overline{m}_{i\bullet}) = \alpha \times \beta = \mu \qquad V(\overline{m}_{i\bullet}) = \alpha \times \beta^2 = {\sigma_m}^2$$

Here $E(\overline{m}_{i\bullet})$ indicates the expected value of the average cluster-specific cluster-period size for cluster *i*, and $V(\overline{m}_{i\bullet})$ represents the variance of the average cluster-specific cluster-period sizes. Rather than assume cluster sizes follow a Normal distribution, which is highly unlikely, we would expect clusters sizes to exhibit a positive skew. Furthermore, we require that the distribution contains no negative values – since clusters cannot contribute a negative number of patients in a time-period. In this context, it is not essential that the cluster sizes are integer sizes. Since the role of this chapter is to consider how variation in cluster sizes affects a stepped-wedge design, it is vital that a fixed average cluster size (and fixed total sample size) can be used, whilst changing the degree of variation in the cluster sizes. As such, we have chosen a Gamma distribution to reflect the distribution of cluster sizes.

The values of α and β, the parameters of the Gamma distribution, can be calculated as:

$$\alpha = \frac{1}{cv^2} \qquad\qquad \beta = \mu \times cv^2$$

By specifying the values of μ and cv, values of α and β can be calculated to simulate $\overline{m}_{i\bullet}$ values. In this setting, there is no variation in size within a cluster, so that:

$$m_{ij} = m_{ij'} \; \forall \, j, j'.$$

To ensure an unbiased comparison of precision between the unequal cluster size design and the equal cluster size design, the total sample size in the SW-CRT with equal cluster size should be the same as the sample size in the SW-CRT with unequal cluster size. To ensure a fixed total sample size is maintained between the two designs, we describe a scaling factor that is used.

5.2.3.3 *Scaling cluster-period sizes to maintain the total sample size*

In the unequal cluster size design, the total sample size (S_u) is the sum of the cluster-period sizes m_{ij} from every cluster and from every time-period:

$$S_u = \sum_{i=1}^{C} \sum_{j=1}^{T} m_{ij} = \overline{m}_{\bullet\bullet} \times C \times T$$

In a design with equal cluster sizes, the total sample size (S_e) is:

$$S_e = \mu \times C \times T$$

When simulating m_{ij}, the total sample size in the unequal cluster size design will differ from the sample size in the equal cluster size design, and so $\sum_{i=1}^{C} \sum_{j=1}^{T} m_{ij} \neq \mu \times C \times T$. As such, a scaling factor is required to maintain the fixed sample size.

An inherent ability of the Gamma distribution is that it can be scaled and still follow a Gamma distribution. That is, if $\mathrm{X} \sim \Gamma(a, b)$, the distribution of $Y = \omega X$ (where ω is a scalar) also follows a Gamma distribution (see Appendix J).

Since the average cluster-specific cluster-period sizes $(\overline{m}_{i\bullet})$ follow a Gamma distribution, it is possible to scale the $\overline{m}_{i\bullet}$ values, and they will still follow a Gamma distribution. As such, we scale the $\overline{m}_{i\bullet}$ values to ensure the total sample size is the same in the unequal cluster size design as the equal cluster size design.

The scaling factor (ω) is calculated as the ratio of the total sample size in the equal cluster size design (S_e) to the total sample size in the unequal cluster size design (S_u), as:

$$\omega = \frac{S_e}{S_u} = \frac{\mu \times C \times T}{\overline{m}_{\bullet\bullet} \times C \times T} = \frac{\mu}{\overline{m}_{\bullet\bullet}}$$

Now, each of the cluster-period sizes, m_{ij}, are multiplied by ω, so that each of cluster-period sizes become ωm_{ij}, and $\sum_{i=1}^{C}\sum_{j=1}^{T}\omega m_{ij} = \mu \times C \times T$. The ωm_{ij} values are then used to calculate the precision in a SW-CRT with varying cluster size.

5.2.3.4 *Estimating the precision*

To estimate the precision, the study design – the number of clusters, time-periods, DPM – is used to create the design matrix, Z. Then, the simulated ωm_{ij} values are inputted into each V_i matrix, to create the variance-covariance matrix, V. From this, the precision is estimated using equation 5.2.

5.2.3.5 *Estimating the relative efficiency*

To estimate the RE, the precision of the unequal cluster size design must be compared to the precision of an equal cluster size design. The design matrix Z will be identical in both the equal cluster size and unequal cluster size scenarios. For the equal cluster size design, all clusters are the same size, for each and every time-period, so that $m_{ij} = \mu = m_{i'j'}\ \forall\ i, i'$ and j, j', and contain a total sample size of $S_e = \mu \times C \times T$ which is identical to the unequal cluster size design. The values of m_{ij} can be inputted into each V_i matrix, and the precision can be estimated using equation 5.2.

The precision from a SW-CRT with unequal cluster size is then compared to the precision of a SW-CRT with equal cluster size as the RE (equation 5.3).

5.2.3.6 *Compiling estimates of the relative efficiency*

The methodology described above generates one estimate of the RE for a particular set of simulated m_{ij} values. After this estimate of the RE has been made, α and β are used to simulate a new set of m_{ij} values, to obtain a new estimate of the precision. This estimate is compared to the precision of a SW-CRT with equal cluster size to obtain a new estimate of the RE.

The simulation of m_{ij} to estimate the RE is repeated a large number of times, in order to form a distribution of possible efficiency values, from which the average RE can be obtained.

5.2.3.7 *Simulation study*

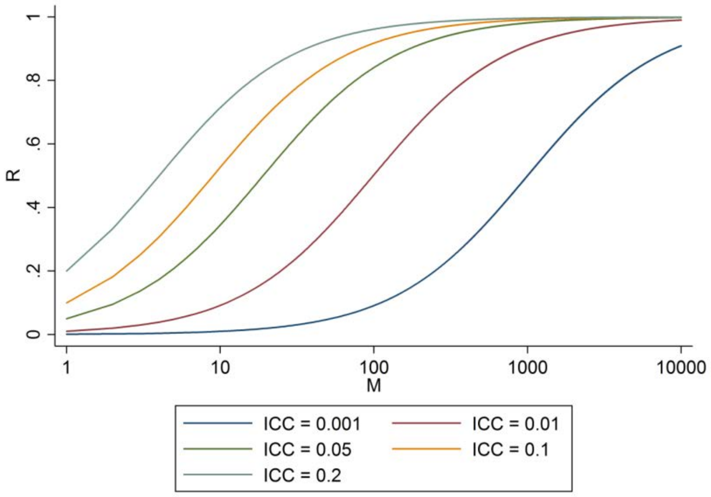
There are many design features that can impact the precision in a SW-CRT, such as: the number of clusters, the cluster size, the number of steps, the ICC, and the coefficient of variation of cluster sizes. However, we combine the ICC and the cluster size as the cluster mean correlation – which is described below – and investigate the impact of the cluster mean correlation on the precision of a SW-CRT.

The cluster mean correlation, denoted as R, represents the correlation between the cluster means of two repeated sets of observations taken from the same cluster (61) and is defined as (9, 48, 61):

$$R = \frac{M \times \rho}{1 + (M - 1)\rho}$$

Generally, $0 \leq R \leq 1$, and it has been shown that R can approach 1, even for small values of the ICC (Figure 5.5).

Figure 5.5: The cluster-mean correlation (R) as a function of the total cluster size (M)



In the simulation study, four factors are considered: (1) the number of steps, (2) the number of clusters, (3) the cluster mean correlation, and (4) the coefficient of variation of cluster sizes (Table 5.2). For the number of steps, values of 2, 3, 4, 6, and 12 were chosen. In the methodological review (Chapter 3), it was identified that over 50% of studies used a design with 4 or less randomisation steps. To capture the full effect, the values of 6 and 12 were also chosen. For simplicity, we maintain an equal number of clusters randomised at each step, so it was necessary that the total number of clusters be a multiple of 12, and so values of 12, 24, 48, and 96 were chosen – which is appropriate since the majority of SW-CRTs include studies with more than ten clusters. For ease, we assumed a total cluster size of 72 participants, which provides an integer cluster-period size for the equal cluster size design regardless of the number of steps. As such, the total number of participants (total sample size) in the 12-cluster, 24-cluster, 48-cluster, and 96-cluster designs are 864, 1728, 3456, and 6912. This sample size is fixed for equal and unequal cluster designs, for all values of the cluster mean correlation, and irrespective of the number of steps. Since it has been established that the cluster mean correlation impacts the precision of the SW-CRT, a wide spectrum of values were chosen (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). The degree of variation in cluster sizes ranged from small (cv = 0.25) to large (cv = 1.5), with incremental values chosen between this range.

Table 5.2: Summary of parameter values used in simulation

Variables	Values chosen
Number of steps	2, 3, 4, 6, 12
Number of clusters	12, 24, 48, 96
Cluster mean correlation	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
Coefficient of variation of cluster sizes	0.25, 0.5, 0.75, 1.0, 1.25, 1.5

In total, 4 x 4 x 11 x 6 = 1,056 combinations of variables were considered.

The number of simulations used to estimate the distribution of the RE is determined by a Monte Carlo estimate of the error around the precision (44). Here, to maintain an error smaller than 1%, 4,000 simulations will be used to form a detailed understanding of the distribution of efficiency.

5.2.3.8 *Summary of methods*

The steps taken to estimate the RE are summarised in Box 5.1.

Box 5.1: Outline of the steps taken to estimate the relative efficiency of a stepped-wedge cluster randomised trial with varying cluster size compared to a stepped-wedge cluster randomised trial with equal cluster size.

Background: Define the SW-CRT design

Define the design of the SW-CRT such as number of time-periods, number of clusters, number of steps, and the cluster mean correlation (i.e. select one of the scenarios from Table 5.2).

Step 1: Simulate cluster-period sizes

Simulate for each cluster, its cluster-period size, m_{ij} such that $m_{ij} = m_{ij'}\ \forall\, j, j'$ using a gamma distribution, with mean μ and coefficient of variation, cv.

Step 2: Scale cluster-period sizes

- Calculate the average cluster-period size ($\bar{m}_{\bullet\bullet}$) as:

$$\bar{m}_{\bullet\bullet} = \frac{1}{CT}\sum_{i=1}^{C}\sum_{j=1}^{T} m_{ij}.$$

- Calculate ω which is the ratio of the total sample size in an equal cluster size design (S_e) to the total sample size in the unequal cluster size design (S_u) as:

$$\omega = \frac{S_e}{S_u} = \frac{\mu\times C\times T}{\bar{m}_{\bullet\bullet}\times C\times T} = \frac{\mu}{\bar{m}_{\bullet\bullet}}.$$

- Scale the cluster-period sizes m_{ij}, as $m_{ij} \times \omega$.

Step 3: Estimating the precision

- Calculate matrices V_i (i=1, ... , C) using the ωm_{ij} from step (2).
- Form matrix V using the block matrices V_i .
- Calculate the precision using equation 5.2.

Box 5.1 continued…

Step 4: **Estimating the relative efficiency**

- Estimate the precision for a SW-CRT with equal cluster sizes with a total sample size $S_e = \mu \times C \times T$, so that each cluster-period is size μ.
- Calculate the RE of the design as a ratio of the precision of the unequal cluster size design (from step 3) compared to the equal cluster size design (from step 4).

Step 5: **Compiling the efficiency estimates**

- Repeat steps (1) - (4) 4,000 times.
- Collate the estimates of the RE to form the distribution (and average) of the RE.

5.2.4 Between cluster variation in a P-CRT

Conventional DEs for a P-CRT assume a singular RE value and the methods are approximate. Here, we want to establish whether the RE is a singular value, and use an exact method to obtain the RE. To this end, the methodology described in section 5.2.3 can be applied to a P-CRT, by describing a P-CRT in the DPM as:

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

By describing a P-CRT in this way, an estimate of the precision of the treatment effect estimate for a P-CRT can then be obtained using equation 5.2. Throughout this work, we consider estimates of the precision for a P-CRT calculated via this approach, and not via a design effect. This approach allows us to relax the assumption of a balance of observations between the intervention and control arms, and provides a distribution of possible RE values – rather than a fixed value.

A P-CRT is a study design with only one time-period, and so essentially, the cluster-period size m_{ij} refers to the size of cluster i over the whole study duration. The methodology described in sections 5.2.3.2 and 5.2.3.3 can be used to simulate the m_{ij} values, and the precision can be estimated using equation 5.2. The precision is then calculated for a P-CRT with equal cluster size and the same total sample size $(S_e = \mu \times C \times T)$. The precision in the P-CRT with unequal cluster size is then compared to the precision in a P-CRT with equal cluster size as a RE. This is repeated a large number of times, and the estimates of the RE are compiled to produce a distribution of possible RE estimates – similar to section 5.2.3.6.

5.2.4.1 *Simulation study*

In this simulation study, three factors are considered: (1) the number of clusters, (2) the cluster mean correlation, and (3) the coefficient of variation of cluster sizes. The parameter values used in this simulation study were based on the values used in the earlier simulation study for the SW-CRT and are summarised below (Table 5.3).

Table 5.3: Summary of parameter values used in simulation

Parameter	**Values chosen**
Number of clusters	12, 24, 48, 96
Cluster mean correlation	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
Coefficient of variation of cluster sizes	0.25, 0.5, 0.75, 1.0, 1.25, 1.5

The number of simulations used for each combination of variables was maintained from the SW-CRT, so that 4,000 simulations were used to maintain an error rate smaller than 1%.

5.2.4.2 *Summary of methods*

The steps taken to estimate the RE are summarised in Box 5.2.

Box 5.2: Outline of the steps taken to estimate the relative efficiency of a parallel cluster randomised trial with varying cluster size compared to a parallel cluster randomised trial with equal cluster size.

Background: Define the P-CRT design

Define the design of the P-CRT such as the number of clusters and the cluster mean correlation (i.e. select one of the scenarios from Table 5.3).

Step 1: Simulate cluster-period sizes

Simulate for each cluster, its cluster-period size, m_{ij}, using a gamma distribution, with mean μ and coefficient of variation, cv.

Step 2: Scale the cluster-period sizes

- Calculate the average cluster-period size ($\bar{m}_{\bullet\bullet}$) as:

$$\bar{m}_{\bullet\bullet} = \frac{1}{CT}\sum_{i=1}^{C}\sum_{j=1}^{T} m_{ij}.$$

- Calculate ω which is the ratio of the total sample size in an equal cluster size design (S_e) to the total sample size in the unequal cluster size design (S_u) as:

$$\omega = \frac{S_e}{S_u} = \frac{\mu \times C \times T}{\bar{m}_{\bullet\bullet} \times C \times T} = \frac{\mu}{\bar{m}_{\bullet\bullet}}.$$

- Scale the cluster-period sizes m_{ij}, as $m_{ij} \times \omega$.

Box 5.2 continued…

Step 3: **Estimating the precision**

- Calculate matrices V_i (i=1, … , C) using the ωm_{ij} from step (2).
- Form matrix V using the block matrices V_i .
- Calculate the precision using equation 5.2.

Step 4: **Estimating the relative efficiency**

- Estimate the precision for a P-CRT with equal cluster sizes with a total sample size $S_e = \mu \times C \times T$, so that each cluster-period is size μ.
- Calculate the RE of the design as a ratio of the precision of the unequal cluster size design (from step 3) compared to the equal cluster size design (from step 4).

Step 5: **Compiling the efficiency estimates**

- Repeat steps (1) - (4) 4,000 times.
- Collate the estimates of the RE to form the distribution (and average) of the RE.

5.2.5 Impact of within-cluster variation in size over time

In section 5.2.3, a method was described to estimate the RE of a SW-CRT with unequal cluster sizes compared to a SW-CRT with equal cluster sizes – with the assumption of each cluster being in fixed in size over time. In this section, we now extend the method to allow individual clusters to independently vary in size over time.

Here, we compare the precision in a SW-CRT with unequal cluster sizes to the precision in a SW-CRT with equal cluster sizes, when the sample size is the same in both designs. To do this, we use the true cluster-period size, the between-cluster coefficient of variation of cluster sizes, and the within-cluster variation in cluster sizes, to simulate potential cluster-period sizes. The between-cluster coefficient of variation in cluster sizes describes the variation in the average cluster-specific cluster-period sizes. That is, it describes the variability in the average cluster size at one time-period. The within-cluster coefficient of variation describes the variability in size of one cluster over time.

Below, sections 5.2.5.2 to 5.2.5.6 detail the steps taken to simulate cluster-period sizes with between-cluster and within-cluster variation in size, and how the RE is estimated. Initially, we illustrate the methods though an example, before elaborating on each step in more depth.

5.2.5.1 *Example of estimating the relative efficiency in a SW-CRT with between-cluster and within-cluster variation in size*

We start with a SW-CRT with 12 cluster, 3 time-periods, and a true cluster-period size $\mu = 50$. Since $m_{ij} = m_{i'j'}\ \forall\ i, i'$ and j, j', it follows that $m_{ij} = m_{i'j'} = 50\ \forall\ i, i'$ and j, j', and so the sample size in the SW-CRT with equal cluster sizes (S_e) will be:

$$S_e = \mu \times C \times T = 50 \times 12 \times 3 = 1800.$$

Now consider a SW-CRT with 12 clusters and 3 time-periods, with $\mu = 50$, $cv = 0.5$, and $cv_w = 0.1$. Assuming that $\bar{m}_{i\bullet}$ follows a Gamma distribution, we can use μ and cv to simulate the average cluster-specific cluster-period size $(\overline{m}_{i\bullet})$ for each cluster *i*. A set of potential $\overline{m}_{i\bullet}$ values are:

$$m_{1\bullet} = 47.10 \qquad m_{2\bullet} = 68.90 \qquad m_{3\bullet} = 33.82 \qquad m_{4\bullet} = 56.24$$

$$m_{5\bullet} = 47.44 \qquad m_{6\bullet} = 71.67 \qquad m_{7\bullet} = 26.81 \qquad m_{8\bullet} = 83.26$$

$$m_{9\bullet} = 28.17 \qquad m_{10\bullet} = 12.09 \qquad m_{11\bullet} = 63.87 \qquad m_{12\bullet} = 60.64$$

These values indicate the average cluster-period size for that cluster, so that $m_{1\bullet}$= 47.10 indicates that the average cluster-period size over the study for cluster 1 is 47.10. The average cluster-specific cluster-period sizes $(\overline{m}_{i\bullet})$ can then be used to obtain the cluster-period sizes (m_{ij}). Here, each cluster is modelled using an independent Gamma distribution, with mean, $\overline{m}_{\mathrm{i}\bullet}$, and coefficient of variation cv_w. Using the values of $\overline{m}_{\mathrm{i}\bullet}$ above, and $cv_w = 0.1$, the simulated cluster-period sizes (m_{ij}) for each cluster *i* may be:

	Time-period 1	**Time-period 2**	**Time-period 3**
Cluster 1	$m_{1\,1} = 43.80$	$m_{1\,2} = 49.85$	$m_{1\,3} = 47.65$
Cluster 2	$m_{2\,1} = 64.53$	$m_{2\,2} = 77.27$	$m_{2\,3} = 64.89$
Cluster 3	$m_{3\,1} = 31.88$	$m_{3\,2} = 36.06$	$m_{3\,3} = 33.53$
Cluster 4	$m_{4\,1} = 54.75$	$m_{4\,2} = 57.98$	$m_{4\,3} = 55.98$
Cluster 5	$m_{5\,1} = 42.03$	$m_{5\,2} = 49.21$	$m_{5\,3} = 51.08$
Cluster 6	$m_{6\,1} = 75.56$	$m_{6\,2} = 60.79$	$m_{6\,3} = 78.67$
Cluster 7	$m_{7\,1} = 26.29$	$m_{7\,2} = 26.78$	$m_{7\,3} = 27.36$
Cluster 8	$m_{8\,1} = 97.55$	$m_{8\,2} = 80.53$	$m_{8\,3} = 71.71$
Cluster 9	$m_{9\,1} = 28.30$	$m_{9\,2} = 27.63$	$m_{9\,3} = 28.57$
Cluster 10	$m_{10\,1} = 12.31$	$m_{10\,2} = 10.00$	$m_{10\,3} = 13.95$
Cluster 11	$m_{11\,1} = 63.75$	$m_{11\,2} = 64.82$	$m_{11\,3} = 63.04$
Cluster 12	$m_{12\,1} = 58.16$	$m_{12\,2} = 60.70$	$m_{12\,3} = 63.06$

These values correspond to cluster-period sizes for cluster *i* at time *j*, so that $m_{4\,2}$ = 57.98 indicates that cluster 4 at time 2 has 57.98 observations.

In the SW-CRT with unequal cluster sizes, the total sample size S_u is the sum of all of the cluster-period sizes m_{ij}, so that:

$$S_u = \sum_{i=1}^{C} \sum_{j=1}^{T} m_{ij}.$$

By assuming that $S_u = S_e$, the precision is calculated for a SW-CRT with unequal cluster sizes and the precision for a SW-CRT with equal cluster sizes using equation 5.2 by inputting each m_{ij} value into each V_i matrix. The RE is calculated as the ratio of the two precisions.

Similar to section 5.2.3, the steps used to estimate the RE can be split into five sections – which are discussed in depth, below.

5.2.5.2 *The simulation of cluster sizes*

Since the cluster-period sizes are unknown, we use the true cluster-period size (μ), the between-cluster coefficient of variation of cluster sizes (cv), and the within-cluster coefficient of variation (cv_w) to simulate potential cluster-period sizes $\left(m_{ij}\right)$. Firstly, we simulate the average cluster-specific cluster-period size $(\overline{m}_{i\bullet})$ for each cluster *i*, using a Gamma distribution with mean μ and coefficient of variation cv. Following this, the cluster-period sizes $\left(m_{ij}\right)$ are simulated for each cluster independently using a Gamma distribution with mean $\overline{m}_{i\bullet}$ and coefficient of variation cv_w, so that the following holds:

$$m_{ij}\big|\overline{m}_{i\bullet} \sim \Gamma\left(\alpha_{w_i}, \beta_{w_i}\right) \quad E\left(m_{ij}\big|\overline{m}_{i\bullet}\right) = \alpha_{w_i} \times \beta_{w_i} = \overline{m}_{\mathrm{i}\bullet} \quad V\left(m_{ij}\big|\overline{m}_{i\bullet}\right) = \alpha_{w_i} \times {\beta_{w_i}}^2 = {\sigma_{w_i}}^2$$

$$m_i \sim \Gamma(\alpha_m, \beta_m) \qquad E(m_i) = \alpha_m \times \beta_m = \mu \qquad V(m_i) = \alpha_m \times {\beta_m}^2 = {\sigma_m}^2$$

Where the values of α_m, α_{w_i}, β_m, and β_{w_i} can be calculated as follows:

$$\alpha_m = \frac{1}{cv^2} \qquad \alpha_{w_i} = \frac{1}{{cv_w}^2}$$

$$\beta_m = \mu \times cv^2 \qquad \beta_{w_i} = \overline{m}_{i\bullet} \times {cv_w}^2$$

By specifying the values of μ, cv, and cv_w, values of α_m and β_m, can be calculated to simulate $\overline{m}_{i\bullet}$ and then m_{ij} values.

An unbiased comparison of the precision in an unequal cluster size design and the precision in an equal cluster size design relies on the total sample size being equal in the two designs – to ensure loss in precision is due to the unequal cluster size structure and not a difference in sample size. A scaling factor is used to maintain a fixed total sample size, which is described below. Although some aspects of the scaling factor were reported in section 5.2.3.3, it is reiterated here for clarity.

5.2.5.3 *Scaling factor to maintain a fixed total sample size*

Because m_{ij} are simulated, the total sample size will differ between the unequal and equal cluster size designs, so that $\sum_{i=1}^{C}\sum_{j=1}^{T} m_{ij} \neq \mu \times C \times T$. Since there is between-cluster and within-cluster variation, two scaling factors are required to maintain the fixed sample size – one to maintain the total sample size, and one to maintain the total cluster size.

To this end, we firstly assume that there is no within-cluster variation in size (only between-cluster variation in size), and use a scaling factor that was described in section 5.2.3.3. Then a second scaling factor is applied to each individual cluster, and compares a cluster with variation in size over time to a cluster with no variation in size over time. Essentially, this is the same scaling factor that is used in section 5.2.3.3 but it is applied cluster-periods within an individual cluster, rather than average cluster-specific cluster-period sizes across clusters.

Scaling factor for total sample size

We begin by considering the total sample size (S_e) in a SW-CRT with equal cluster sizes:

$$S_e = \mu \times C \times T$$

In a design with unequal clusters, the total sample size (S_u) is the sum of the cluster-period sizes m_{ij} from every cluster and from every time-period:

$$S_u = \sum_{i=1}^{C} \sum_{j=1}^{T} m_{ij} = \bar{m}_{\bullet\bullet} \times C \times T$$

Firstly, the scaling factor (ω) is calculated for the average cluster-specific cluster-period sizes $(\bar{m}_{i\bullet})$ as the ratio of the total sample size in the equal cluster size design (S_e) to the total sample size in the unequal cluster size design (S_u), as:

$$\omega = \frac{S_e}{S_u} = \frac{\mu \times C \times T}{\bar{m}_{\bullet\bullet} \times C \times T} = \frac{\mu}{\bar{m}_{\bullet\bullet}}$$

Now, each of the cluster-period sizes, m_{ij}, are multiplied by ω, so that each of cluster-period sizes become ωm_{ij}, and $\sum_{i=1}^{C} \sum_{j=1}^{T} \omega m_{ij} = \mu \times C \times T$.

Scaling factor for total cluster size

For each cluster, the average cluster-period size must be calculated - using the ωm_{ij} values. We refer to this as $\overline{\omega m}_{\mathrm{i}\bullet}$, which is calculated as:

$$\overline{\omega m}_{\mathrm{i}\bullet} = \frac{1}{T} \sum_{j=1}^{T} \omega m_{ij}$$

Following this, for cluster i, we simulate new values of m_{ij} from a Gamma distribution with mean $\overline{\omega m}_{\mathrm{i}\bullet}$ and coefficient of variation cv_w. As such, the new values of m_{ij} now contain variation in size over time, and so differ from those simulated earlier. As a result, the average cluster-specific cluster-period size $\overline{m}_{i\bullet}$ will differ from its earlier value.

Now, for a SW-CRT with no within-cluster variation in size, the total cluster size for cluster *i* $\left(M_{e_i}\right)$ is:

$$M_{e_i} = \overline{\omega m}_{i\bullet} \times T$$

In a design with within-cluster variation in size over time, the total cluster size for cluster *i* $\left(M_{u_i}\right)$ is the sum of the cluster-period sizes for cluster *I* from each time period:

$$M_{u_i} = \sum_{j=1}^{T} m_{ij} = \overline{m}_{i\bullet} \times T$$

Now, since m_{ij} are simulated, the total cluster size for cluster *i* with within-cluster variation will differ from the total cluster size with no within-cluster variation, so that $\sum_{j=1}^{T} m_{ij} \neq \overline{\omega m}_{i\bullet} \times T$, and a scaling factor is required to maintain a fixed total cluster size.

The scaling factor for cluster *i* (ω_i) is calculated as the ratio of the two total cluster sizes as:

$$\omega_i = \frac{M_{e_i}}{M_{u_i}} = \frac{\overline{\omega m}_{i\bullet} \times T}{\overline{m}_{i\bullet} \times T} = \frac{\overline{\omega m}_{i\bullet}}{\overline{m}_{i\bullet}}$$

Now, each of the cluster-period sizes $\left(m_{ij}\right)$ in cluster *i* are multiplied by ω_i, so the cluster-period sizes m_{ij} become $\omega_i m_{ij}$. The $\omega_i m_{ij}$ values are then used to calculate the precision in a SW-CRT with varying cluster size.

5.2.5.4 *Estimating the precision*

To estimate the precision, the study design – the number of clusters, time-periods, and DPM – is used to create the design matrix, Z. The simulated cluster-period sizes $\left(m_{ij}\right)$ are inputted into each V_i matrix, to create the variance-covariance matrix, *V*. The precision is then estimated using equation 5.2.

5.2.5.5 *Estimating the relative efficiency*

To estimate the RE, the precision of an unequal cluster size design is compared to the precision of an equal cluster size design. In both designs, the design matrix Z will be identical. In the equal cluster size design, all clusters are the same size, for each and every time-period, so that $m_{ij} = m_{i'j'}\ \forall\ i, i'$ and j, j', and so the total sample size is $S_e = \mu \times C \times T$ which is identical to the unequal cluster size design. Each m_{ij} can be inputted into each V_i matrix, and the precision can be estimated using equation 5.2.

The precision from a SW-CRT with unequal cluster size is then compared to the precision of a SW-CRT with equal cluster size as the RE (equation 5.3).

5.2.5.6 *Compiling estimates of the relative efficiency*

The above methodology generates one RE estimate for a particular set of m_{ij} values. After an estimate of the RE has been made, α and β are used to simulate a new set of m_{ij} values (which are corrected to ensure the same total sample size) and a new estimate of the precision is obtained. This estimate is compared to the precision of a SW-CRT with equal cluster size to obtain a new estimate of the RE.

The simulation of m_{ij} to calculate the RE is repeated a large number of times, to form a distribution of possible RE values, and an estimate of the average RE.

5.2.5.7 *Simulation study*

Table 5.4 highlights the scenarios used in this simulation study. Since SW-CRTs typically employ a small number of clusters, only the two smaller numbers of clusters were used (12 and 24). To encapsulate the impact of the number of steps, only the two extreme values were used (2 and 12). All values of R and cv used in the primary analysis were included in this secondary analysis. For the within-cluster variation component, we used only small (cv_w = 0.25) and medium (cv_w = 0.50) levels of within-cluster variation.

Table 5.4: Summary of parameter values used in the variation over time extension to the simulation study

Parameter	**Values chosen**
Number of steps	2, 12
Number of clusters	12, 24
Cluster mean correlation	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
Coefficient of variation of cluster sizes	0.25, 0.5, 0.75, 1.0, 1.25, 1.5
Coefficient of variation of cluster-period sizes	0.25, 0.50

For each combination of variables, the number of simulations was maintained from that used for the earlier SW-CRT simulation study and the P-CRT simulation study, so that 4,000 simulations were used to maintain an error rate smaller than 1%.

5.2.5.8 *Summary of methods*

The steps taken to estimate the RE are summarised in Box 5.3.

Box 5.3: Outline of the steps taken to estimate the relative efficiency of a stepped-wedge cluster randomised trial with varying cluster size (between-cluster and within-cluster) compared to a stepped-wedge cluster randomised trial with equal cluster size.

Background: Define the SW-CRT design

- Define the design of the SW-CRT such as number of time-periods, number of clusters, number of steps, and the cluster mean correlation (i.e. select one of the scenarios from Table 5.4).

Step 1: Simulate cluster-period sizes with no within-cluster variation in size

- Simulate for each cluster, its cluster-period size, m_{ij} such that $m_{ij} = m_{ij'}\ \forall\ j, j'$ using a gamma distribution, with mean μ and coefficient of variation, cv.

Step 2: Scale the cluster-period sizes

- Calculate the average cluster-period size ($\bar{m}_{\bullet\bullet}$) as:

$$\bar{m}_{\bullet\bullet} = \frac{1}{CT}\sum_{i=1}^{C}\sum_{j=1}^{T} m_{ij}.$$

- Calculate ω which is the ratio of the total sample size in an equal cluster size design (S_e) to the total sample size in the unequal cluster size design (S_u) as:

$$\omega = \frac{S_e}{S_u} = \frac{\mu \times C \times T}{\bar{m}_{\bullet\bullet} \times C \times T} = \frac{\mu}{\bar{m}_{\bullet\bullet}}.$$

- Scale the cluster-period sizes m_{ij}, as $m_{ij} \times \omega$.

Box 5.3 continued…

- Calculate for each cluster, the average cluster-period size $\overline{\omega m}_{i\bullet}$ as:

$$\overline{\omega m}_{i\bullet} = \frac{1}{T}\sum_{j=1}^{T} \omega m_{ij}$$

Step 3: **Simulate cluster-period sizes with within-cluster variation in size**

Simulate, for each cluster independently, a new cluster-period size (m_{ij}) using a Gamma distribution with mean $\overline{\omega m}_{i\bullet}$ and coefficient of variation cv_w.

Step 4: **Apply a scaling factor**

- For cluster i, calculate the average cluster-specific cluster-period size $(\overline{m}_{i\bullet})$, as:

$$\overline{m}_{i\bullet} = \frac{1}{T}\sum_{j=1}^{T} m_{ij}$$

- For cluster i, calculate ω_i which is the ratio of the total cluster size in a design with no within-cluster variation in size over time (M_{e_i}) to the total cluster size in a design with within-cluster variation in size over time (M_{u_i}) as:

$$\omega_i = \frac{M_{e_i}}{M_{u_i}} = \frac{\overline{\omega m}_{\mathrm{i}\bullet} \times T}{\overline{m}_{\mathrm{i}\bullet} \times T} = \frac{\overline{\omega m}_{\mathrm{i}\bullet}}{\overline{m}_{\mathrm{i}\bullet}}$$

- Scale the cluster-period sizes m_{ij}, as $\omega_i m_{ij}$.

Step 5: **Estimate the precision**

- Calculate matrices V_i (i=1, … , C) using $\omega_i m_{ij}$ from step (4)
- Form matrix V using the block matrices V_i .
- Calculate the precision using equation 5.2.

Box 5.3 continued…

Step 6: Estimating the relative efficiency

- Estimate the precision for a SW-CRT with equal cluster sizes with a total sample size $S_e = \mu \times C \times T$ so that each cluster-period is size μ.
- Calculate the RE of the design as a ratio of the precision of the unequal cluster size design (from step 4) compared to the equal cluster size design (from step 5).

Step 7: Compiling the efficiency estimates

- Repeat steps (1) - (6) 4,000 times.
- Collate the estimates of the RE to form the distribution (and average) of the RE.

5.3 Results

5.3.1 Relative efficiency plots used to display the results

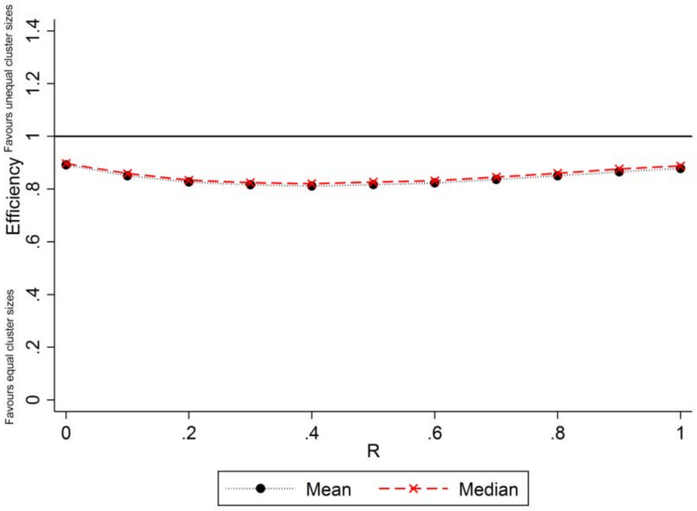
For each scenario described earlier (sections 5.2.3, 5.2.4, and 5.2.5), estimates of the precision were simulated with unequal cluster size and compared to the precision for an equivalent design with equal cluster size. The total sample size is identical in both the equal and unequal settings, to ensure that any differences are to the unequal cluster size structure. All results are presented as the relative efficiency (RE) of the unequal cluster size scenario compared to the equal cluster size scenario. These results provide an approximate upper bound for the RE of a given design. All estimates are plotted as efficiency curves, with RE plotted against the cluster mean correlation (denoted as R). For all efficiency plots, each value of R contains 4,000 estimates of the RE, so that each graph contains 44,000 simulations.

There are two types of graphs presented here. We firstly consider estimates of the mean and median values of the RE, calculated using the 4,000 simulated values. This allows us to consider how the design is affected, on average. Secondly, we consider the full distribution of the RE estimates, and present the quintiles of these estimates.

On all graphs, a RE of 1 would indicate that the unequal cluster size scenario is equally efficient as the equal cluster size setting. This is indicated by the bold horizontal line on Figure 5.6. An estimate of the RE greater than 1 would indicate that the unequal cluster setting is more efficient than an equal cluster size design, and a RE less than one would favour the equal cluster size setting. Essentially RE>1 indicates more precision in the unequal cluster size design, and RE<1 indicates more precision in the equal cluster size design.

The mean and median are plotted on all graphs, with the mean of the RE presented as a black dotted line with circles representing the data points, whilst the median is presented as a red dashed line with a cross indicating the data points. In all scenarios, the data is skewed towards zero, and so the mean is always less than the median. Because of this skewness, we present both the mean and median in all plots.

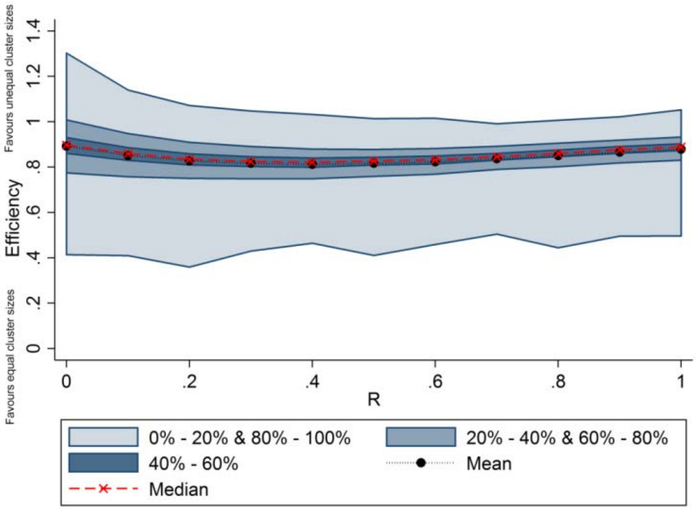
Figure 5.6: Explanation of mean and median efficiency curves



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

To highlight the full distribution of the precision, we present the full range of simulated estimates of the RE, in which different shadings are used to represent the quintiles of the data – see, for example Figure 5.7. The lightest shaded region represents the outside 40% of the RE estimates, signifying the 0 – 20% region (1st quintile) and the 80 – 100% region (5th quintile) of the RE values. The slightly darker region indicates the next 40% of the data, referring to the 20 – 40% region (2nd quintile) and the 60 – 80% region (4th quintile). The darkest shaded area reflects the middle 20% of the RE values (3rd quintile). An example is given below in Figure 5.7. The mean and median values are presented alongside the shaded quintile regions. Again, a value of the RE greater than one would favour the unequal cluster size design, and a value of RE less than one would favour the equal cluster size design. In the design below, when R = 0, the RE varies from 0.4 to 1.3, highlighted by the edges of the 0% and 100% percentiles. This would indicate that at best, a SW-CRT with unequal cluster size may offer 30% more precision than a SW-CRT with equal cluster sizes. However, at worst, a SW-CRT with unequal cluster sizes may offer 60% less precision than a SW-CRT with equal cluster sizes. The middle 20% of the RE values are between 0.89 and 0.96.

Figure 5.7: Explanation of quintiles on efficiency curves



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

In this section, the key results are highlighted, and we present a selection of graphical illustrations to support these. Full results of the simulation study are provided via efficiency curves showing the full range of the RE estimates, and can found in Appendix K, Appendix L, and Appendix M.

5.3.2 Impact of varying cluster size on stepped-wedge cluster randomised trials

In this section, we present the results highlighting the impact of within-cluster variability on the precision of a SW-CRT by comparing the precision in a SW-CRT with varying cluster size to the precision of a SW-CRT with equal cluster size as a RE. We present results of the average efficiency change and the distribution of efficiency. It is highlighted how changes to the number of steps, the number of clusters, and R may affect the efficiency of a SW-CRT with varying cluster size.

5.3.2.1 *Average values of the relative efficiency*

Firstly, only the mean and median values of the RE are presented. In all scenarios, both the mean and median RE estimates are never greater than 1 (Appendix K). This leads us to our first key finding, which is given in Box 5.4.

Box 5.4: Key finding for SW-CRT in relation to the average efficiency loss when clusters vary in size

The unequal cluster setting is, on average, less efficient than a setting with equal cluster size. This holds true for all designs, and so is irrespective of the number of steps, the number of clusters, and the value of R.

Figure 5.8 shows that for small cv, the RE of the design is not heavily impacted by the cluster mean correlation – the average RE remains close to the reference line. For larger cv values, there is visibly more influence of R on the efficiency (Figure 5.9). Between R = 0 and R = 0.4, the average efficiency is decreasing. However, between R = 0.4 and R = 1, the average RE is increasing. The RE is at its lowest when R = 0.4, and at its highest when R = 1.

Figure 5.8: Average efficiency vs R for a SW-CRT with 12 clusters, 2 randomisation steps, and cv = 0.25

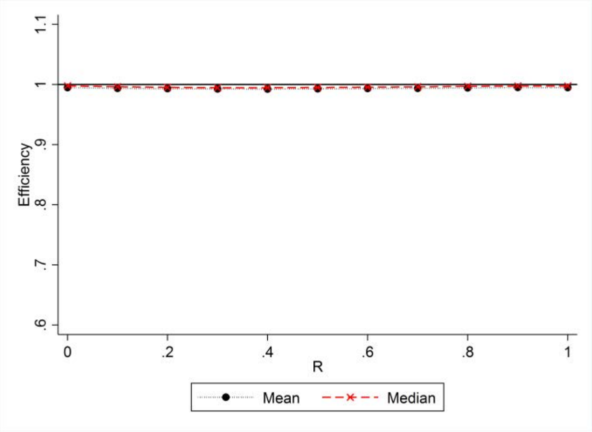
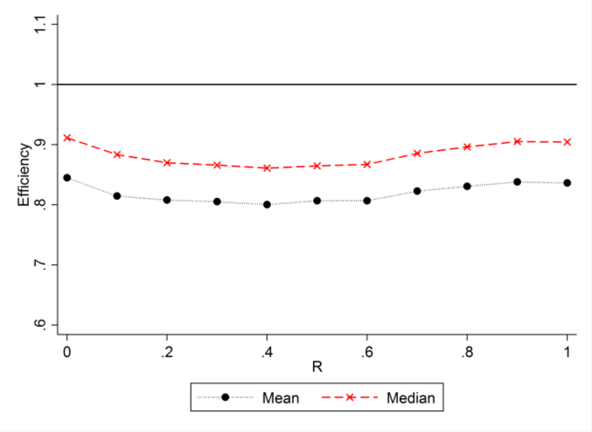


Figure 5.9: Average efficiency vs R for a SW-CRT with 12 clusters, 2 randomisation steps, and cv= 1.50



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

This leads us to our next key point, which is given in Box 5.5.

Box 5.5: Key finding of the cluster mean correlation on the average precision

When the variability between cluster sizes is small, there is little loss in precision, for any value of the cluster mean correlation. When the variability between cluster sizes is large, the amount of precision lost is related to the cluster mean correlation.

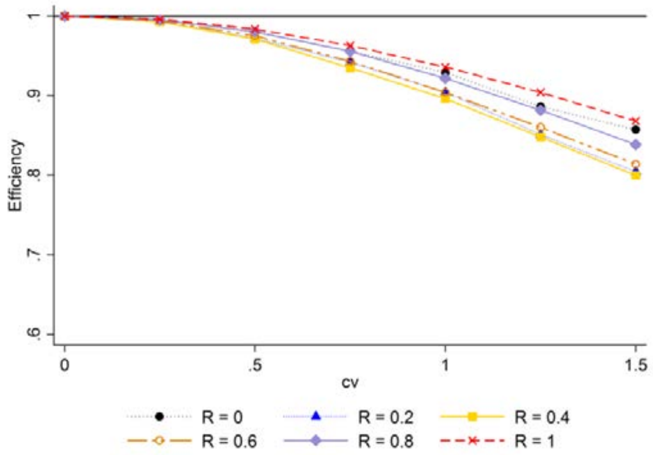
For many scenarios, there is a large difference between the mean and median RE – which is highlighted further by the distribution plots which show a large negative skew. As such, when describing the average efficiency, we refer to the median RE.

Our third key result is given in Box 5.6 and highlighted in Figure 5.10, which shows that as cv increases, the average efficiency deviates further from the reference value of 1. Noticeably, whilst the general shape is similar for each R value, the rate of decay in the efficiency varies depending on the value of R. When considering the average efficiency, the design with R at its extreme values (R = 1 and R = 0) seems to lose the least amount of efficiency as the cv increases, whilst R = 0.4 loses the most (Figure 5.10). This result is consistent regardless of the number of clusters or the number of randomisation steps.

Box 5.6: Key finding of increasing between-cluster variability on the average precision

The greater the cluster size variability, the smaller the average precision will be.

Figure 5.10: Median efficiency vs coefficient of variation (cv) for a SW-CRT with 12 clusters and 12 randomisation steps



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

5.3.2.2 *Distribution of efficiency estimates*

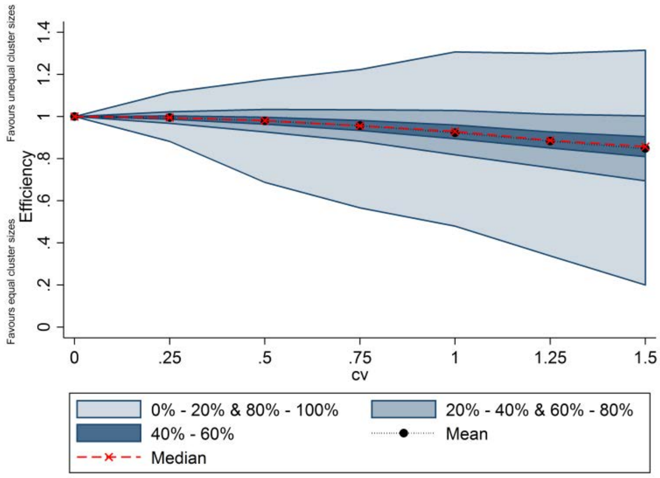
A full depiction of the efficiency of the SW-CRT comes when we consider the complete efficiency curves, which highlight the full distribution of the efficiency. That is, they display every estimate of the RE that is calculated for that particular scenario. The complete efficiency curves lead us to our next key finding – given in Box 5.7. The efficiency curves show that it is possible for the RE to be greater than 1. This would indicate the unequal cluster size setting being more efficient than a design with equal cluster size conditional on a particular randomisation order. That is, a SW-CRT with unequal cluster sizes can have a precision that is greater than in a SW-CRT with equal cluster sizes.

Box 5.7: Key finding of the distribution of precision in a SW-CRT with unequal cluster size

In a SW-CRT with varying cluster size, it is possible for the precision to be greater than in a SW-CRT with equal cluster size.

Figure 5.11 shows that as cv increases, the amount of variability in RE increases, as well as the magnitude of the possible change in RE. In this scenario, even for small values of cv, there could be deviations from RE = 1 when considering the full distribution of RE values.

Figure 5.11: Efficiency vs coefficient of variation (cv) for a SW-CRT with 12 clusters and 12 randomisation steps and R = 0



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

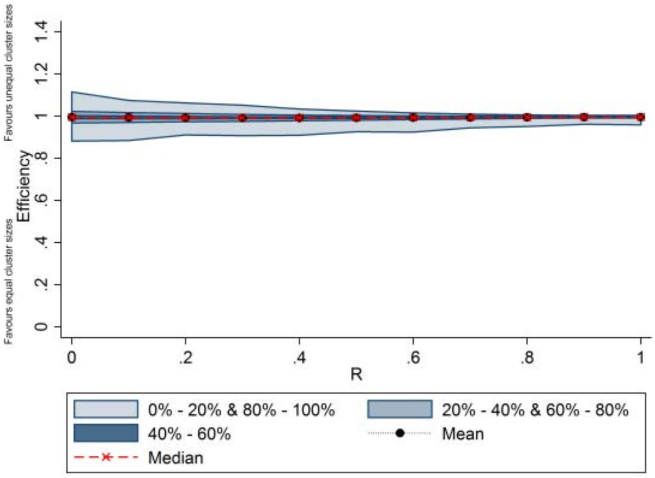
When cv is small, there is little loss in the average efficiency, and little variation in the potential efficiency. That is, unequal cluster sizes has little impact on the precision when cv is small. However, for large cv, there is a loss in the average precision and a large variation in the potential efficiency. As such, we now present distribution curves separately for scenarios with small and large values of cv.

Small values of the coefficient of variation

Figure 5.12 shows that for small cv, the potential change in RE is quite small. The largest potential change in RE is at R = 0 (RE range: 0.9 - 1.1), and so the precision in an unequal cluster size may be between 10% lower or 10% higher than the precision of an equal cluster size scenario. The likelihood of an unequal cluster size scenario offering greater precision than the equal cluster size scenario is almost half.

As R increases, the potential change in RE decreases further, so that at R = 1, there is little deviation in RE from the reference value. That is to say the precision in the unequal cluster size scenario will be almost identical to the precision in an equal cluster size scenario. This leads to our next key finding (Box 5.8).

Figure 5.12: Efficiency vs R for a SW-CRT with 12 clusters, 12 randomisation steps, and cv = 0.25



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Box 5.8: Impact of small cluster size variability on the precision

If the variability in cluster sizes is small, then the potential loss or gain in precision in a SW-CRT with varying cluster size compared to a SW-CRT with equal cluster size is very small. As such, varying cluster size is likely to have minimal impact on the power when the coefficient of variation of cluster sizes is small.

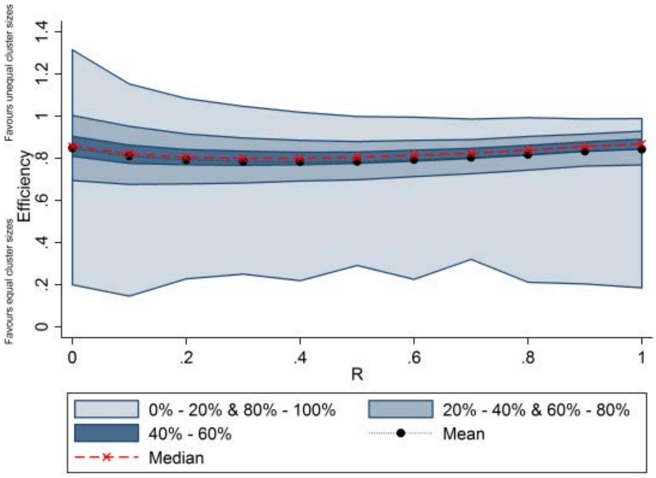
Large values of the coefficient of variation

Figure 5.13 shows that for large cv, the RE is very variable. At R = 0, the RE varies between 0.2 and 1.3 – indicating that the unequal cluster size scenario may increase the precision by 30% compared to an equal cluster size scenario, or decrease the precision by 80%. This magnitude is much greater than for small cv (Figure 5.12). However, the likelihood of RE>1 is only 20%, which is much less than for small cv (Figure 5.12).

As R increases, the upper RE value decreases, so that at R = 1, the upper RE value is just above 1 – which would indicate that at best, the unequal cluster size scenario will have slightly more precision than an equal cluster size scenario – though this increase would be tiny in magnitude. The lower RE value does not change much as R increases.

For large cv, the plots are less symmetrical – since the upper RE value decreases as R increases. However, the inner 60% of the RE values are relatively unaffected by changes to R.

Figure 5.13: Efficiency vs R for a SW-CRT with 12 clusters, 12 randomisation steps, and cv = 1.5

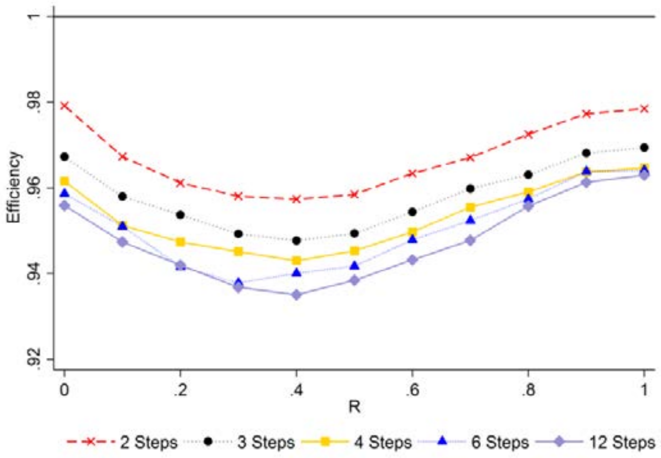


Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

5.3.2.3 *Impact of increasing number of steps*

The average RE is influenced by the number of randomisation steps. As the number of steps increases, the average RE decreases. That is, for a fixed total sample size, a SW-CRT with 2-steps loses less precision when the clusters vary in size than a SW-CRT with any other amount of steps (Figure 5.14). Conversely, as the number of steps increases, the greater the potential increase in RE – which would indicate a greater increase in precision in the unequal cluster size scenario from the equal cluster size scenario.

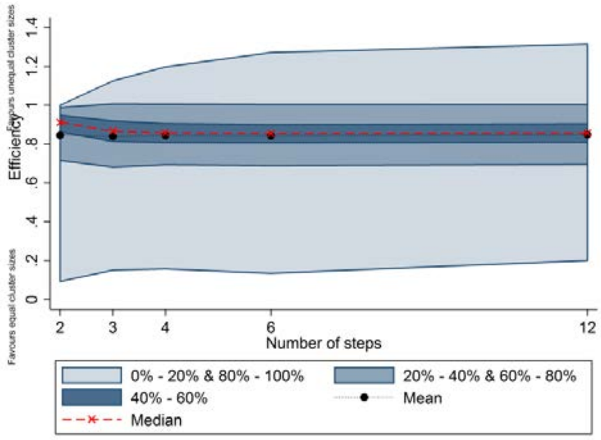
Figure 5.14: Median efficiency vs R for a SW-CRT with 12 clusters, and cv = 0.75



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Figure 5.15 shows that for large cv when R = 0, increasing the number of steps can lead to greater RE values. With 2 steps, the maximum RE is approximately one, which would indicate the unequal cluster size scenario would have at best the same precision as an equal cluster size scenario. However, an unequal scenario with 12 steps could have 30% more precision than an equal cluster size scenario with 12 steps. The potential loss in efficiency is relatively constant regardless of the number of steps – the unequal cluster size scenario may have only 20% of the precision of an equal cluster size scenario. When the $cv = 1.5$, the likelihood of RE>1 is approximately 20%, regardless of the number of randomisation steps.

Figure 5.15: Efficiency vs steps for a SW-CRT with 12 clusters, R = 0, and cv = 1.50



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Comparing RE across R shows similar results. Figure 5.16 and Figure 5.13 show that an increase in the number of steps does not greatly change the lower 80% of the RE values, though they do decrease slightly. The upper RE values are greater as the number of steps increases.

Figure 5.16: Efficiency vs R for a SW-CRT with 12 clusters, 2 randomisation steps, and cv = 1.50

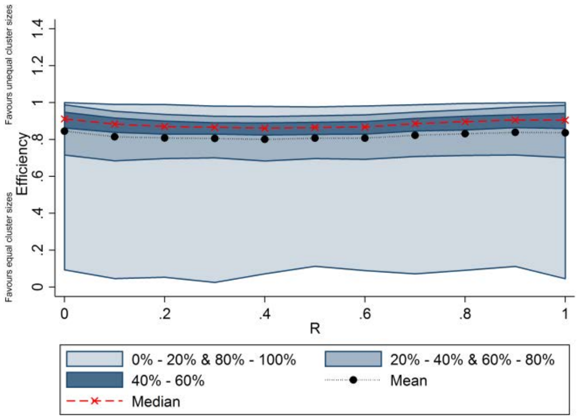
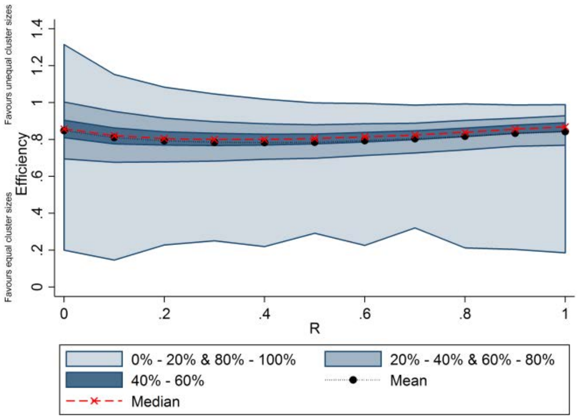


Figure 5.13: Efficiency vs R for a SW-CRT with 12 clusters, 12 randomisation steps, and cv = 1.5



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

The most variability in RE is found in scenarios with a greater number of steps. However, the additional variability is due to the potential for the RE >1 – so whilst it is more variable, the unequal cluster size scenario may offer more precision than an equal cluster size scenario, with scenarios with a greater number of steps offering the largest increases in precision.

This leads us to our next key result, given in Box 5.9.

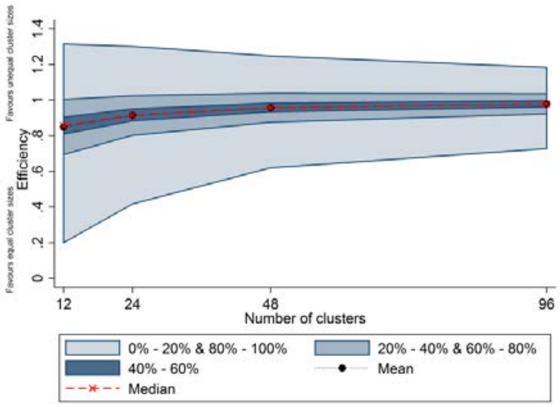
Box 5.9: Impact of the number of randomised steps on the precision

The average relative efficiency is affected less in designs with fewer randomisation steps, i.e. designs with a greater number of steps will have a smaller average relative efficiency. Furthermore, increasing the number of randomisation steps increases the variability in the relative efficiency. Consequently, designs with a greater number of randomisation steps can see greater increases and decreases in the relative efficiency.

5.3.2.4 *Impact of increasing number of clusters*

The impact of increasing the number of clusters is quite unambiguous – an increase in the number of clusters leads to a decrease in the variability of the RE estimates. That is, the distribution becomes narrower as more clusters are included – so that the potential gains or losses in precision are minimised (Figure 5.17). This is to be expected, as an increase in the number of clusters will lead to more clusters being randomised at each step, and so there is a lower likelihood of an imbalance between the number of participants contributing to the intervention and control conditions.

Figure 5.17: Efficiency vs number of clusters for a SW-CRT with 12 randomisation steps, R = 0 and cv = 1.5



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Figure 5.13 and Figure 5.18 show the decrease in variability in the RE as the number of clusters increases – and the shift towards 1 for the average RE. This mostly comes from a shift towards 1 from the bottom 0-20% percentile. Generally, increasing the number of clusters has minimal impact on the top 80-100% quintile. When the cv is large, both the 12-cluster and 96-cluster scenarios could offer a 20% increase in precision in the unequal cluster size scenario compared to an equal cluster size scenario. However, the 12-cluster unequal cluster size scenario may lose 80% of the precision of the equal cluster size scenario, whereas the 96-cluster scenario only loses up to 25% of its precision.

Figure 5.13: Efficiency vs R for a SW-CRT with 12 clusters, 12 randomisation steps, and cv = 1.5

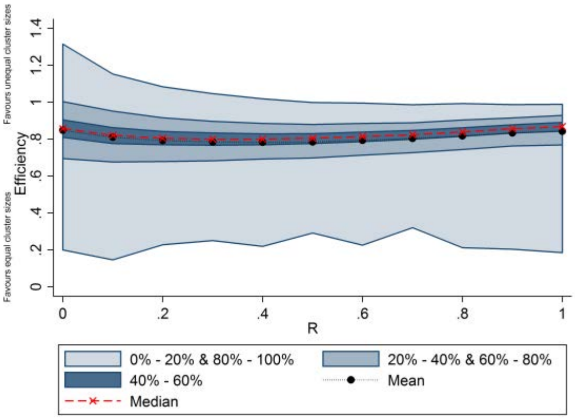
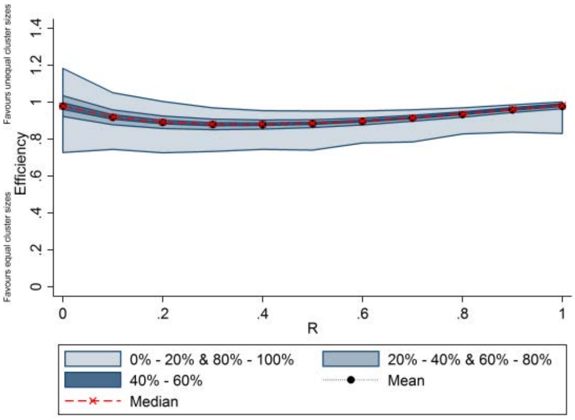


Figure 5.18: Efficiency vs R for a SW-CRT with 96 clusters, 12 randomisation steps, and cv = 1.50



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

5.3.3 Impact of varying cluster size on parallel cluster randomised trials

As a comparator, plots of the RE against R are also presented for the parallel design. Since there are no randomisation steps in a P-CRT, there are only 4 main scenarios – representing the RE of a study with 12, 24, 48, and 96 clusters.

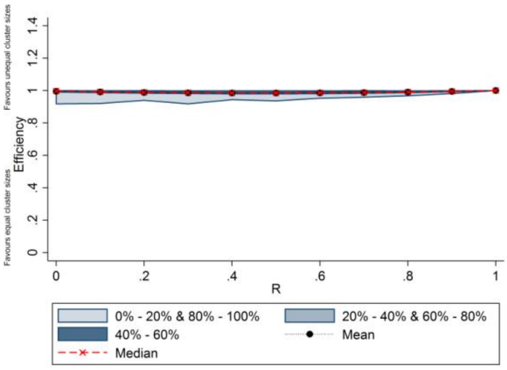
The average efficiency is less than 1 for all scenarios (Appendix L), indicating that a P-CRT with unequal cluster size is, on average, less efficient than a P-CRT with equal cluster size. As expected, if the variability in cluster sizes is small, there is little efficiency loss. As the variability in cluster sizes increases, there is a much greater degree of efficiency loss. Noticeably, the results show that the common assumption of the RE for a P-CRT being a single value may not be appropriate – and the RE can vary in a P-CRT and is dependent on the randomisation of the clusters. This leads us to our next key finding (Box 5.10). For all scenarios, the efficiency curves show that the RE is bounded by 1 (Appendix L).

Box 5.10: Key finding for the relative efficiency in a P-CRT

The precision (and hence the relative efficiency) is not a single value in a P-CRT with unequal cluster sizes. Instead the precision can be viewed as a distribution, which is influenced by the randomisation order of the clusters.

The efficiency curves also corroborate that there is little efficiency loss in a P-CRT when cv is small. For example, Figure 5.19 shows that cv = 0.25, there is little variation in the distribution of the RE. The most variability is at R = 0, with RE between 0.9 and 1.0. This corresponds to a 10% decrease in precision in the unequal cluster size scenario compared to the equal cluster size scenario. As R increases, we see a slight decrease in the variability of the RE. At R = 1, there is almost no variation in RE, so the unequal cluster size scenario will almost always have the same precision as an equal cluster size scenario.

Figure 5.19: Efficiency vs R for a P- CRT with 12 clusters, and cv = 0.25

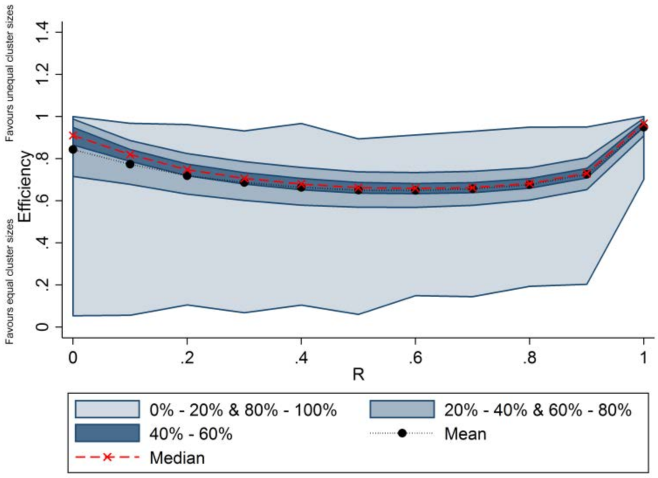


Efficiency is calculated as the ratio of the precision in a P-CRT with unequal cluster sizes compared to the precision in a P-CRT with equal cluster sizes. R represents the cluster mean correlation.

An increase to the cv can create a large variability in the potential RE in a P-CRT. Figure 5.20 shows that for $cv = 1.5$ there is huge variability in the possible RE values. At R = 0, the RE varies between 0.1 and 1.0, which would indicate that a P-CRT with unequal cluster size may only have 10% of the precision of a P-CRT with equal cluster size.

As R increases, the middle 60% of the RE values generally follow the same parabolic trajectory as the median RE – and are most variable at R = 0 and least variable at R = 1. However, the upper and lower values remain generally unaffected by R, and remain hugely variable.

Figure 5.20: Efficiency vs R for a P-CRT with 12 clusters, and cv = 1.5

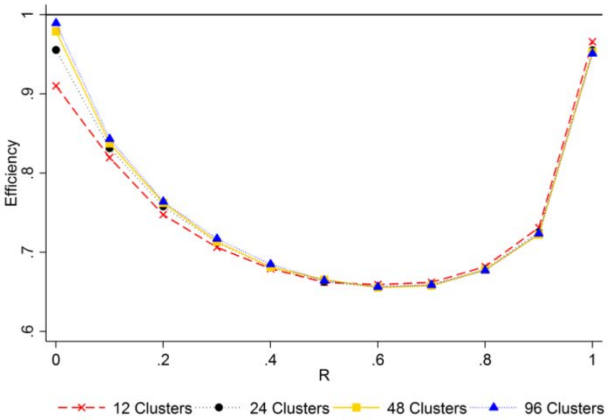


Efficiency is calculated as the ratio of the precision in a P-CRT with unequal cluster sizes compared to the precision in a P-CRT with equal cluster sizes. R represents the cluster mean correlation.

5.3.3.1 *Impact of increasing the number of clusters*

Figure 5.21 shows that increasing the number of clusters in the P-CRT does not influence the median RE. The RE generally follows the same parabolic trajectory regardless of the number of clusters, apart from at R = 0 – in which the RE increases in scenarios with more clusters.

Figure 5.21: Median efficiency vs R for a P-CRT with cv = 1.5



Efficiency is calculated as the ratio of the precision in a P-CRT with unequal cluster sizes compared to the precision in a P-CRT with equal cluster sizes. R represents the cluster mean correlation.

Figure 5.20 and Figure 5.22 show that increasing the number of clusters leads to a decrease in the variability of RE. At R = 0, scenarios with 12 clusters and cv = 1.50 may have lose 90% of the precision of an equal cluster size scenario – this is in contrast to the 96 cluster scenario in which the precision may decrease by approximately 20%. The middle 60% of RE values becomes much narrower as the number of clusters increases.

Figure 5.20: Efficiency vs R for a P-CRT with 12 clusters, and cv = 1.5

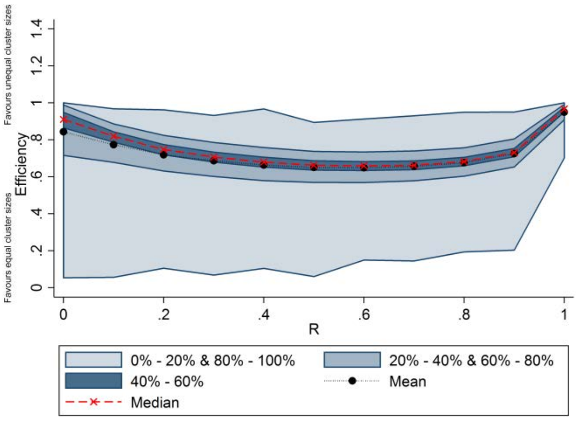
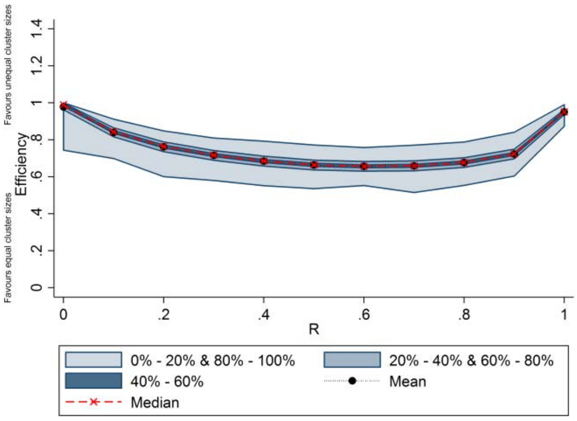


Figure 5.22: Efficiency vs R for a P-CRT with 96 clusters, and cv = 1.5



Efficiency is calculated as the ratio of the precision in a P-CRT with unequal cluster sizes compared to the precision in a P-CRT with equal cluster sizes. R represents the cluster mean correlation.

5.3.4 Comparison of P-CRTs and SW-CRTs

In this section we compare the RE of a SW-CRT to the RE of a P-CRT. In this section, the RE of a SW-CRT will refer to a SW-CRT with unequal cluster size compared to a SW-CRT with equal cluster size, and the RE of a P-CRT will refer to a P-CRT with unequal cluster size compared to a P-CRT with equal cluster size.

5.3.4.1 *Comparison of the average relative efficiency*

The median RE estimates are greater in the SW-CRT than in the P-CRT for most values of R. However, when R is close to its extreme values (R = 0 and R = 1), the P-CRT has greater RE. This is particularly evident in cases with large values of the cv. These results are consistent for designs with a small number of steps (Figure 5.23) and designs with a greater number of steps (Figure 5.24). Essentially, this indicates that, on average, the SW-CRT is more efficient than the P-CRT.

Figure 5.23: Comparison of average efficiency of a 24 cluster CRT under a stepped-wedge design (2 steps) and a parallel design with cv = 1.5

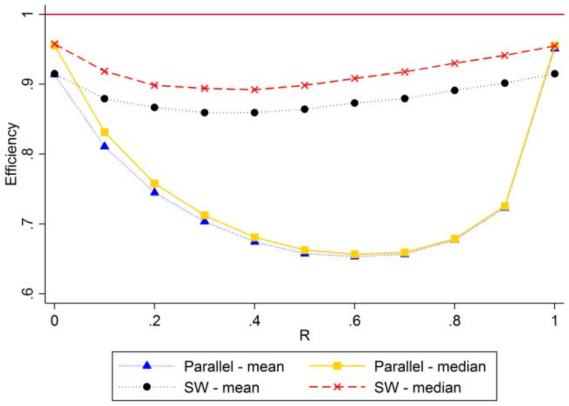
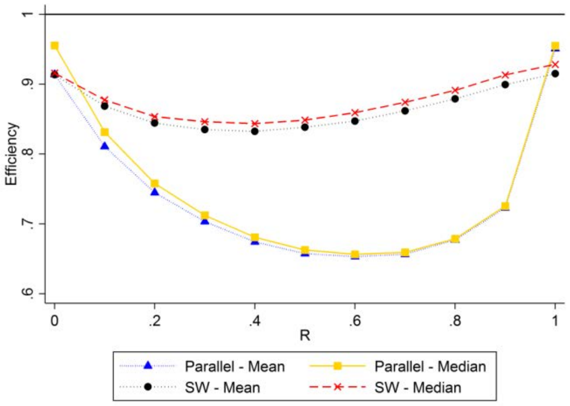


Figure 5.24: Comparison of average efficiency of a 24 cluster CRT under a stepped-wedge design (12 steps) and a parallel design with cv = 1.5



Efficiency is calculated as the ratio of the precision in a CRT with unequal cluster sizes compared to the precision in a CRT with equal cluster sizes. R represents the cluster mean correlation.

5.3.4.2 *Comparison of the distribution of relative efficiency*

When considering the full distribution of RE estimates, there is a noticeable difference between the P-CRT and the SW-CRT. There were many instances in the SW-CRT when unequal cluster size presented a design more precision than one with equal cluster size (i.e. RE>1) (Appendix K). However, in the P-CRT, there is no circumstances in which the RE > 1 (Appendix L), and so an unequal cluster size scenario can never offer more precision than an equal cluster size scenario. Figure 5.13 and Figure 5.20 show that at R = 0, the SW-CRT RE can lie between 0.2 and 1.3, whereas the P-CRT lies between 0.1 and 1.0 – indicating that the SW-CRT can lose 80% of the precision of an equal cluster size scenario but the P-CRT could lose 90%. However, the SW-CRT could have 30% more precision than an equal cluster size scenario, whereas the P-CRT cannot increase in precision. As R increases, the variability in the SW-CRT RE decreases, whereas the variability in the P-CRT RE only decreases at R = 1. The middle 60% of RE values is narrower in the P-CRT than the SW-CRT, which emphasises that the P-CRT RE is less variable than the SW-CRT RE.

Figure 5.13: Efficiency vs R for a SW-CRT with 12 clusters, 12 randomisation steps, and cv = 1.5

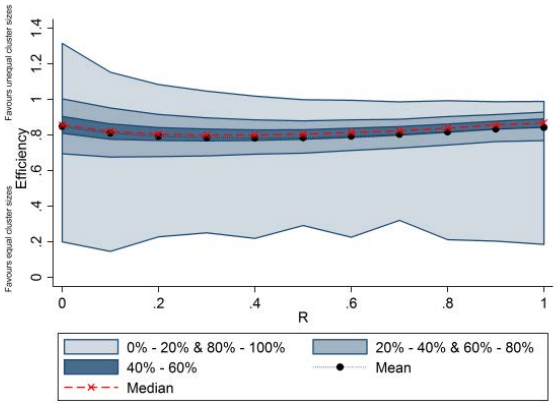
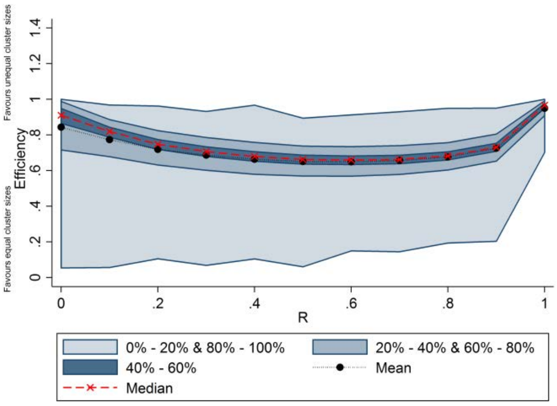


Figure 5.20: Efficiency vs R for a P-CRT with 12 clusters, and cv = 1.5



Efficiency is calculated as the ratio of the precision in a CRT with unequal cluster sizes compared to the precision in a CRT with equal cluster sizes. R represents the cluster mean correlation.

The RE estimates are more widely distributed for a SW-CRT than a P-CRT, principally when the number of steps is greater than 2. This includes the range of RE estimates and the quintiles. This leads to the key result for the impact of varying cluster size in the SW-CRT compared to the P-CRT, which is given in Box 5.11.

Box 5.11: Key finding of the impact of varying cluster size in a stepped-wedge cluster randomised trial compared to a parallel cluster randomised trial.

On average, a SW-CRT is affected less by varying cluster size than a P-CRT. Potentially, a SW-CRT can offer a more efficient design with unequal cluster sizes than equal cluster sizes, which is impossible in a P-CRT. However, there is a greater degree of variability in the possible RE of the SW-CRT.

5.3.4.3 *Comparison when increasing the number of clusters*

When increasing the number of clusters in the P-CRT the median RE is not influenced. This is in stark contrast to the SW-CRT in which the RE increases as more clusters are included. The full distribution of RE values becomes much narrower in both the SW-CRT and P-CRT when more clusters are included in the study.

5.3.5 Impact of within-cluster variation in size over time

In this section, we compare the precision of a design with between-cluster variation and within-cluster variation to a design with equal cluster size – with a fixed total sample size in both designs – as a RE. The added cv_w component reflects additional variation in this section than in section 5.3.2, and remarks will be made to compare to the cv_w = 0 scenario (i.e. no within-cluster variation). Results presented consider both small (cv_w= 0.25) and medium

(cv_w = 0.50) values of cv_w. Estimates of the RE were also calculated, assuming that $cv_w = cv$, but the results were hugely variable, and showed levels of variance in cluster sizes unlikely to be seen in practice. Efficiency curve plots of $cv_w = cv$ can be found in Appendix M.

The results still consider how the efficiency is affected by increases to the between-cluster variation, which is still referred to as cv and should not be confused with the within-cluster variation, denoted as cv_w.

In all scenarios, the median RE is never greater than 1 (Appendix M), and so the unequal design is, on average, less efficient than a design with equal sized clusters. When considering the full distribution of the RE, it can be noted that the RE can be greater than 1 for all designs that were considered, indicating that unequal cluster size can offer more precision than an equal cluster size design (Appendix M).

The inclusion of within-cluster variability leads to a greater variability in the potential RE. The degree of impact is influenced by the number of randomisation steps, and so the results are presented separately for designs with few steps and designs with a large number of steps.

5.3.5.1 *Small number of steps*

For design with few randomisation steps, when the between-cluster variation (cv) is small, there is little loss in the average efficiency as the within-cluster (cv_w) increases. That is, when comparing to a scenario with equal cluster size, the median precision lost in the cv_w = 0 scenario and the median precision lost in the cv_w = 0.50 scenario is relatively negligible. As the between-cluster variation increases, there is a large disparity between the median RE for the cv_w = 0, 0.25, and 0.50 scenarios. The loss in RE between the cv_w = 0 and cv_w = 0.25 scenarios remains relatively constant across R.

The full distribution of RE estimates is influenced by the within-cluster variation. In a design with small between-cluster variation in size and no within-cluster variation in size (Figure 5.25), there is little variability in the RE. However, in a design with small between-cluster variation in size and a large within-cluster variation in size (Figure 5.26), there is large variability in RE. This result is consistent for all R values.

Figure 5.25: Efficiency vs R for a SW-CRT with 12 clusters, 2 randomisation steps, and cv = 0.25 (no within-cluster variation)

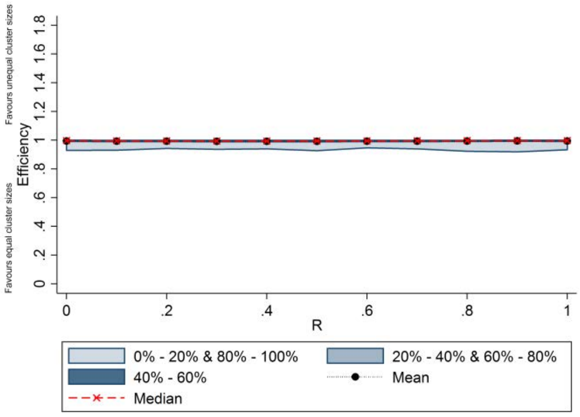
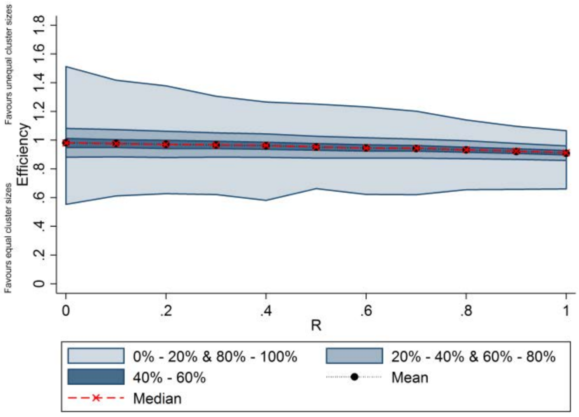


Figure 5.26: Efficiency vs R for a SW-CRT with 12 clusters, 2 randomisation steps, cv = 0.25, and cv$_w$ = 0.50



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

As the between-cluster variability increases, the within-cluster variability can still add a large degree of variability to the RE. Noticeably, Figure 5.16 and Figure 5.27 show that for scenarios with large cv, the inclusion of within-cluster variability allows the RE>1, which does not occur when cv_w = 0. That is to say, it is possible to have more efficient designs when $cv_w > 0$ than scenarios with cv_w = 0. However, the lower RE values remain the same regardless of the cv_w value.

Figure 5.16: Efficiency vs R for a SW-CRT with 12 clusters, 2 randomisation steps, and cv = 1.50 (no within-cluster variation)

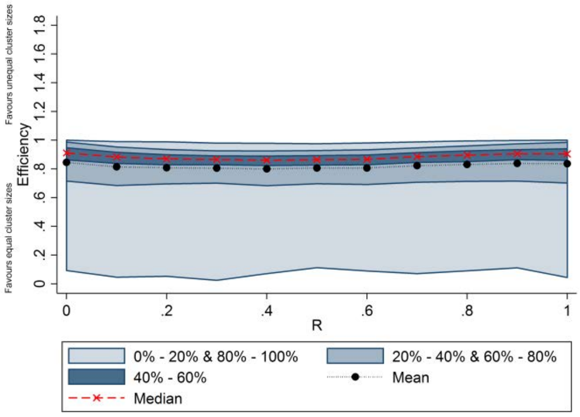
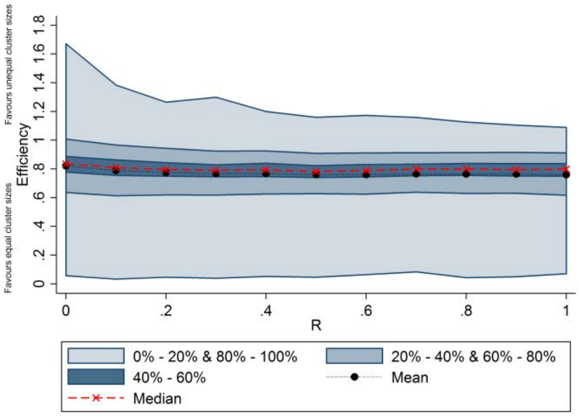


Figure 5.27: Efficiency vs R for a SW-CRT with 12 clusters, 2 randomisation steps, cv = 1.50 and cv_w = 0.50



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

5.3.5.2 *Large number of steps*

In designs with a large number of randomisation steps, an increase to the within-cluster variation (cv_w) has minimal effect on the average RE regardless of the degree of between-cluster variability (cv). That is, when comparing to a scenario with equal cluster size, the average precision lost in the cv_w = 0 scenario and the median precision lost in the cv_w = 0.5 scenario is relatively negligible.

The full distribution of RE is also not greatly impacted by changes to the RE as cv_w increases, irrespective of the amount of between-cluster variation in size. Figure 5.12 and Figure 5.28 shows that the within-cluster variation has little impact when cv is small, with the RE for scenarios with cv_w = 0 ranging between 0.9 and 1.1, whilst scenarios with cv_w = 0.5 ranging between 0.8 and 1.2. Noticeably, as R increases, the RE quintiles remains relatively constant for cv_w = 0.5, whereas the RE tends to 1 for all quintiles when cv_w = 0. As such, when R is close to 1, the precision in an unequal cluster size scenario, with cv_w = 0, will be very similar to the precision in an equal cluster size scenario, whereas if cv_w = 0.5, the precision may be 20% higher or lower than the precision in an equal cluster size scenario.

Figure 5.12: Efficiency vs R for a SW-CRT with 12 clusters, 12 randomisation steps, and cv = 0.25 (no within-cluster variation)

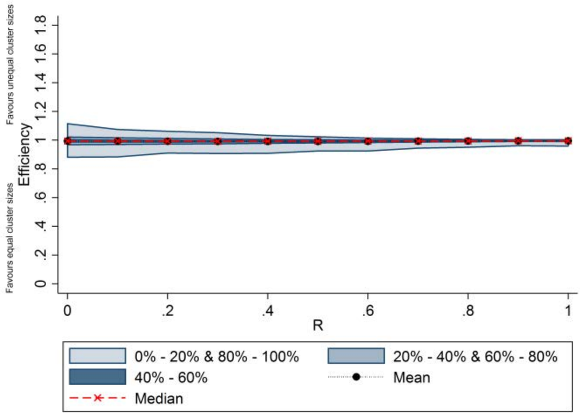
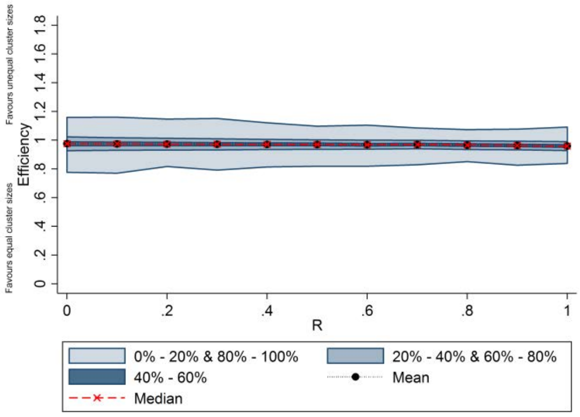


Figure 5.28: Efficiency vs R for a SW-CRT with 12 clusters, 12 randomisation steps, cv = 0.25, and cv_w = 0.50



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Figure 5.13 and Figure 5.29 show that for large cv, the inclusion of within-cluster variation does not greatly influence the distribution of RE. At R = 0, the variability in RE is almost identical for cv_w = 0 and cv_w = 0.5 – varying between 0.2 and 1.3. The only noticeable differences in the RE distribution is at R = 1, in which scenarios with within-cluster variation can obtain RE>1 – that is to say, at R=1, it is possible for an unequal cluster size scenario (with within-cluster variation) to have more precision than an SW-CRT with equal cluster size, whereas it is not possible for an unequal cluster size scenario with cv_w = 0 to obtain more precision than a SW-CRT with equal cluster size.

Figure 5.13: Efficiency vs R for a SW-CRT with 12 clusters, 12 randomisation steps, and cv = 1.5 (no within-cluster variation)

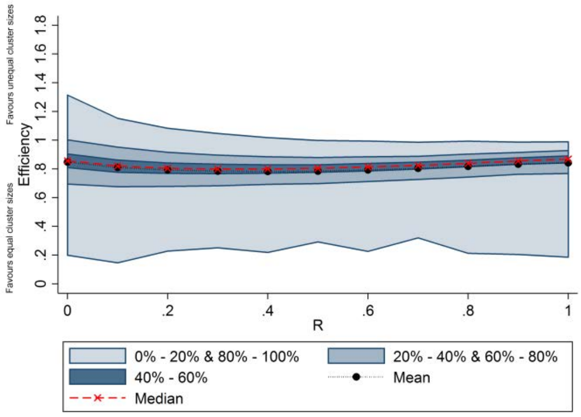
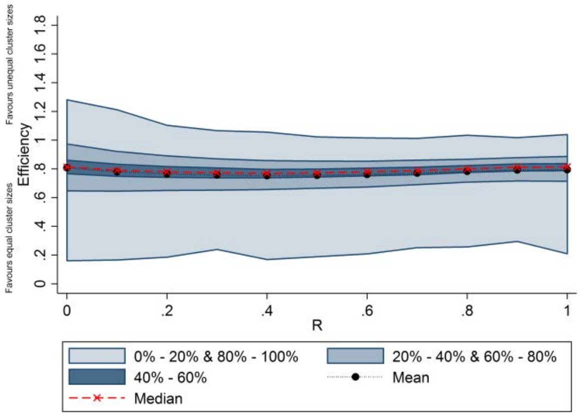


Figure 5.29: Efficiency vs R for a SW-CRT with 12 clusters, 12 randomisation steps, cv = 1.50 and cv$_W$ = 0.50



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

5.4 Discussion

In this chapter, we have examined the impact of cluster size variability on SW-CRTs by comparing a SW-CRT with unequal cluster size to a SW-CRT with equal cluster size. The methodology used centres upon the fitting of a mixed-effect model to a continuous outcome, and so the results may only be applicable in this setting. A simulation study was used to establish the influence of the cluster mean correlation, the number of clusters, and the number of randomisation steps. The results show that it is possible for a SW-CRT with varying cluster size to have only 20% of the efficiency of a SW-CRT with equal cluster size. That is to say, on average, a trial that is powered using methodology that assumes equal cluster size when the cluster sizes are not equal may have an underpowered trial.

We have shown that the average efficiency loss in a SW-CRT with varying cluster size compared to a SW-CRT with equal cluster size may not be large on average, but the variation in the RE is large, and so the actual efficiency loss is conditional on the randomisation order of the clusters. Noticeably, we found that it is possible for a SW-CRT with unequal cluster size to offer greater precision than a SW-CRT with equal cluster size. Since it is typically assumed that the RE of a P-CRT with varying cluster size compared to a P-CRT with equal cluster size is a single value (62) – based on the assumption of an equal number of observations in the control and intervention period – we utilised the same methodology as for a SW-CRT to highlight the effect of varying cluster size in a P-CRT. It was shown here that the RE may actually be a distribution of values. A P-CRT with unequal cluster size can never offer more precision than a P-CRT with equal cluster size. When comparing the results from the SW-CRT to the P-CRT, the median RE is greater in the SW-CRT than in the P-CRT for most values of the cluster mean correlation. This indicates that on average, the SW-CRT loses less efficiency than the P-CRT when the clusters vary in size. However, the variability of the RE is greater in the SW-CRT than the P-CRT, with the SW-CRT able to offer extremely efficient and extremely inefficient designs. Even if a equal cluster size design would be sufficiently powered, the volatility of the RE of a SW-CRT with large cv makes it difficult to give a high degree of certainty that a study will be sufficiently powered for all possible randomisation permutations.

Within CRTs, there are often settings in which there is variation in the size of the clusters (57). This may be to variation in actual size of the centre (such as hospital or school), or due to drop-out and non-response. In P-CRTs, equal cluster size optimises the estimation of variables and variance components. Previous research has considered the loss in efficiency when considering varying cluster size in P-CRTs compared to the equal sized cluster case (62). This has led to some simple approximations for the RE to consider the impact of cluster size variation (60).

SW-CRTs are an increasingly used type of CRT, but one in which there is much less research than the traditional parallel design. As such, there is little or no research that considers the impact of varying cluster size in SW-CRTs, with previous methodological paper highlighting this flaw (45). In this chapter, existing methodology that allows the estimation of the precision of a treatment effect estimate for a SW-CRT is considered. It is shown how simulations can allow this methodology to be extended to examine the effects of varying cluster size. We introduce the idea of a between-cluster variation in cluster size and a within-cluster variation in cluster size. In a trial, the between-cluster variation is likely to be larger, and so will have most impact on the RE of a design. A simulation study allowed varying degrees of cluster size variation to be considered on a variety of designs.

In pre-trial sample size calculations and post-trial analysis, it is recommended that trialists report whether the clusters are expected to vary in size (70). Often, trials may be sufficiently powered for a design with no cluster size variability, whereas the actual cluster sizes may vary, and the study may be underpowered. The large amount of variability in RE also means that adjusting for the average loss in efficiency still may not create a study with sufficient power, as the randomisation of clusters may leave the study with a lower power. The post-trial reporting of cluster sizes in published SW-CRTs – which is underreported (see Chapter 3) – would allow the conduction of a post-trial power calculation to assess a study's true power (150).

Although the methods and results described here have been for continuous outcomes, the methods can be used as an approximation for binary outcomes. As such, similar results are likely for binary outcomes. However, the approach is conditional on the number of observations being sufficiently large so that a Normal approximation can be used. Furthermore, the approximation is likely to be appropriate only when the time and intervention effects are small – since large effects would lead to greater differences in the prevalence and would violate the assumption of constant variance.

This chapter has assumed the fitting of a generalised linear mixed model (GLMM). An alternative method would have been to use generalised estimating equations (GEE) (31). GEE models are more robust for any misspecification in the variance structure (90) – which may be important given the continuing research into the most appropriate variance structure in a SW-CRT. However, GEEs require a large number of clusters to fit correctly. Recent systematic reviews have shown that many SW-CRTs have few clusters (31). Therefore, GEE models may not always be appropriate to fit.

5.4.1 Loss of precision in SW-CRTs when clusters vary in size

Little research has previously been presented on the inclusion of varying cluster size in a SW-CRT, and so there is little evidence of potential variations in the precision based upon the randomisation of clusters. We have shown here that the RE is much more variable in a SW-CRT than a P-CRT – and so a SW-CRT can be greater affected by varying cluster size than a P-CRT. Notably, there were many scenarios in which a SW-CRT with unequal cluster size could offer more precision than a SW-CRT with equal cluster size. Conversely, P-CRTs can never obtain a greater precision in an unequal cluster size setting than the equal cluster size setting, and so the RE can never exceed 1 in a P-CRT.

When the between-cluster variation in cluster sizes is small, there is little efficiency lost in a SW-CRT with unequal cluster size compared to a SW-CRT with equal cluster size, and there is little variation in the potential efficiency change. As the amount of variability in cluster sizes increases, there in more efficiency lost on average, and so the unequal design is, on average, more inefficient. However, an increase to the between-cluster variability can also lead to designs that offer much greater precision when the cluster sizes vary than designs with equal cluster size. The potential increases in precision become greater as the between-cluster variability increases.

5.4.1.1 *Impact of the SW-CRT design features on the efficiency loss*

There has been much research on the impact of the SW-CRT design features (number of steps, clusters, etc.) on the power of a SW-CRT – but no evidence on how they may impact a SW-CRT when clusters vary in size. It was shown here that the efficiency lost in a SW-CRT with unequal cluster size compared to a SW-CRT with equal cluster size is affected by the number of steps. On average, a greater the number of steps leads to a greater decrease in the RE. However, a greater number of steps can offer greater potential increases in RE, and so offer greater increases in precision than a SW-CRT with equal sized clusters. As expected, an increase in the number of clusters led to a decrease in the variability of the potential efficiency change. That is, the range of potential values of the precision is narrower as more clusters are included in the study.

The cluster mean correlation (R) – a function of the cluster size, and the ICC – has previously been shown to be pivotal in deciding which of the SW-CRT and the P-CRT has greater precision, and power, for any given design (61). Whilst here, we were not showing which design was most favourable, in terms of precision, R has been shown to also influence the median RE of both designs when the clusters vary in size. For values of R close to its extreme (0 and 1), the P-CRT may be the most efficient design. However, for the majority values of R, the SW-CRT is, on average, more efficient. As such the SW-CRT is, on average, less affected by varying cluster size. The value of R also heavily impacts the distribution of the RE. For all scenarios, the most variation in RE was at R = 0, regardless of the degree of variability between-clusters and within-clusters. Furthermore, the least variability tended to be at R = 1. For SW-CRTs, the likelihood of a design producing a RE > 1 was influenced by R, with R = 0 offering the greatest probability and prospective RE value. That is, the precision in an unequal cluster size scenario could be much more than the precision in an equal cluster size scenario when R = 0.

5.4.2 Within-cluster variation in size over time

P-CRTs only have between-cluster variation in cluster sizes. In this chapter, we have shown that in SW-CRTs, there are two types of cluster size variation – between-cluster and within-cluster. The inclusion of time is crucial in the design and analysis of SW-CRTs, but the impact of time on the size of each cluster has not previously been considered. The design of SW-CRTs is usually conducted under the assumptions of all clusters being equal in size, and each cluster remaining a fixed size over time. This is the first instance in which two types of cluster size variation have been considered, and the additional variation component leads to an increased range of potential RE values, even for small amounts of within-cluster, and between-cluster variations. The inclusion of within-cluster variation leads to possible increases in RE for all scenarios – regardless of the number of steps. That is, it is possible for studies with unequal cluster size to offer more precision than an equal cluster size scenario.

Previously, only between-cluster variation in size had been considered important when conducted power calculations. As such, there are many examples of expected values of cv to describe the between-cluster variation – such as cv = 0.65 for UK general practices (57). Since within-cluster variation has not previously been considered, it is unknown what the likely values of cv_w would be. However, it is likely that the within-cluster variation in size will be much smaller than the between-cluster variation. Whilst in this research, values of 0.25 and 0.50 were chosen for cv_w, it may be that in practice, smaller values are necessary.

5.4.3 Unequal sized clusters in P-CRTs

In P-CRTs, typically a singular RE value is used (60) – under the assumptions that the intervention and control arms are balanced in terms of number of observations – and small cv values only minimally impact the power (57, 150), with no adjustment needed if cv < 0.23 (57). If $cv \leq 1$, then the median RE in a P-CRT may be between 0.8 and 0.9 (62) – which is also shown in this chapter. However, when considering the distribution of RE values, it is possible to have RE close to 0.1 in extreme scenarios. That is to say, it is possible for a P-CRT with varying cluster size to have only 10% of the precision of a P-CRT with equal cluster size. This may be due to the imbalance of observations in the intervention and control arms, and so may be motivation for an advancement on simple randomisation (i.e. cluster sizes should be used in a balancing algorithm). However, there is no evidence of earlier work considering the variability of RE in P-CRTs and the necessity to use cluster sizes as part of a balancing algorithm. For P-CRTs, ignoring varying cluster sizes could leave a trial seriously underpowered, depending on the randomisation of clusters.

P-CRTs with few clusters and with small values of the cluster mean correlation were most affected by varying cluster size. This result has previously been reported for P-CRTs for designs with few clusters and small values of the ICC (150).

5.4.4 Future research

Previous research in P-CRTs has shown that it is always most efficient to have a design with equal cluster size. In SW-CRTs, there is the possibility of increases in the precision for designs with unequal cluster size. This therefore offers a potential route for further research, with the aim of investigating designs that offer the greatest RE under specific constraints. This can be extended to consider which designs would minimise losses in RE. This could potentially lead to suggestions in how the randomisation process should be done in SW-CRTs with varying cluster size. For example, it may be necessary to match clusters based on estimated cluster size in order to prevent the possibility of a large efficiency loss. As part of this, the use of optimal designs could be considered. Whilst there has been research into optimal designs for cluster trials with equal cluster size (61), there is potential for optimal designs of stepped cluster studies with unequal cluster size. If a study has a small number of large sized clusters, then it may be necessary to randomise large sized clusters separately to smaller sized clusters to ensure a balance of participants in the intervention/control arms and prevent a loss in efficiency.

Although the approach described is for continuous outcomes, it may be used as an approximation for binary outcomes. However, the approach is conditional on the number of observations being sufficiently large so that a Normal approximation can be used. Furthermore, the approximation is likely to be appropriate only when the time and intervention effects are small – since large effects would lead to greater differences in the prevalence and would violate the assumption of constant variance. Further research is necessary to consider the applicability of this work in studies with dichotomous outcomes.

A key limitation of this work is that model chosen includes only a random effect for cluster. The previous chapter has illustrated that a more complex correlation structure may be necessary. However, even the model discussed in Chapter 4 may not be an adequate representation of the correlation structure in a SW-CRT and could be improved. In this model, the correlation between observations in the same cluster but at different time-periods is a fixed value (the inter-period correlation), and is independent of the duration of the time between the observations. Recent research has suggested using a more complex correlation structure that allows for the correlation between observations to become smaller as the length of time between the observations increases (151). In this framework, the correlation within a cluster exhibits an exponential decay. An alternative option would be to consider time as continuous (151). Under this approach, patients that have observations one week apart would have higher correlation than patients whose observations are one month apart which would be greater than those one year apart. However, a correct specification of the time component may be difficult in the design stage of a trial. Nevertheless, the Hussey and Hughes model is the most complex tractable model that we know, and so it is important to establish the impact of varying cluster size within this framework, before considering an extended framework. Further work may be necessary to establish the most appropriate correlation structure for SW-CRTs before being combined with the work here on varying cluster size.

5.5 Conclusion

There is an abundance of research into the effects of varying cluster size in P-CRTs, and it is has been established that small degrees of variation in cluster sizes leads to little impact on the design (57). However, there is a dearth of literature surrounding the impact of varying cluster size in a SW-CRT – which this research aimed to address. Three simulation studies were presented to investigate the impact of between-cluster variability in a SW-CRT, and in a P-CRT, and to assess the impact of within-cluster variability over time in a SW-CRT. For this cluster-period sizes were simulated for to form a CRT with unequal cluster sizes, and this was comparted to a CRT with equal cluster sizes.

The key result of this chapter is that the precision of a CRT with varying cluster size is actually a distribution of possible values, which are influenced by the randomisation of clusters. This allows the precision of a CRT with unequal cluster sizes to be much greater or much less than a CRT with equal sized clusters. Whilst the key findings have been reported in the results section, a summary of them is presented in Box 5.12. The next chapter aims to implement the methods presented in this chapter in a practical setting, by creating a Stata function to estimate power in a SW-CRT with varying cluster size.

Box 5.12: Key findings from simulation study on the impact of varying cluster size in a SW-CRT and a P-CRT

- *The precision (and hence the power) is not a single value in a CRT with unequal cluster sizes. Instead the precision can be viewed as a distribution, which is influenced by the randomisation order of the clusters.*
- *A CRT with unequal cluster sizes will, on average, have less precision than a CRT with equal sized clusters, regardless of the study design.*
- *On average, a SW-CRT is affected less by varying cluster size than a P-CRT.*
- *Potentially, a SW-CRT can offer a more efficient design with unequal cluster sizes than equal cluster sizes, which is impossible in a P-CRT.*
- *There is a greater degree of variability in the possible precision of a SW-CRT with unequal cluster sizes than in a P-CRT.*
- *The greater the cluster size variability, the smaller the average precision will be.*
- *The greater the cluster size variability, the larger the variation in the possible precision*
- *With little between-cluster variability in size, the potential loss or gain in precision in a SW-CRT with varying cluster size compared to a SW-CRT with equal cluster size is very small, and so only minimally impacts the precision.*
- *The variability in the efficiency shows that the precision and power in a CRT is impacted by the randomisation order of the clusters.*

CHAPTER 6: A STATA FUNCTION TO ESTIMATE POWER IN STEPPED-WEDGE CLUSTER RANDOMISED TRIALS WITH VARYING CLUSTER SIZE

6.1 Introduction

6.1.1 Background

The central focus of most research into cluster randomised trials (CRTs) is the assessment of power (52, 57, 74, 145-147) – which is defined as the probability of detecting a treatment effect, when a difference between the arms exists. The power in a CRT is influenced by the correlation between observations within a cluster (55) and by the cluster sizes and their variability (58). In the previous chapter, we showed that a stepped-wedge CRT (SW-CRT) with unequal cluster size will have less precision (and therefore power), on average, than a SW-CRT with equal cluster sizes. When using current statistical packages, the power calculation for a SW-CRT with varying cluster size would require the assumption of equal cluster sizes (63) or acknowledging varying cluster size through a design effect only appropriate for parallel CRTs (P-CRTs) (152). Current DEs to estimate sample size for a fixed power in a SW-CRT (21, 26) are only appropriate if the clusters are of equal size. Since unequal cluster sizes may decrease the power in a SW-CRT, power calculations should account for varying cluster size.

6.1.2 Chapter aim

This chapter aims to implement the methods described in Chapter 5 in a Stata function – that will allow the power to be estimated in a SW-CRT with varying cluster size. To this end, we seek to:

1. Show how the power can be estimated in a SW-CRT with known cluster size, for a given randomisation order and for all possible randomisation orders.
2. Illustrate a Stata function that can estimate the power for a SW-CRT with equal or varying cluster size.
3. Demonstrate how the Stata function works through a selection of examples.

In this chapter, we reiterate the methods used to estimate the power in a SW-CRT with varying cluster size using the framework proposed in Chapter 5, which is appropriate for continuous outcomes. We then illustrate how this methodology can be implemented in practice, through a Stata function. The Stata function can estimate the power in a SW-CRT with unequal clusters sizes – including both known and unknown cluster sizes – which we highlight through a number of examples. The reporting of power includes both the average power and a measure of dispersion (inter-quartile range and range) where appropriate. For SW-CRTs with known cluster size, the power can be estimated for all randomisation orders, or one particular order.

6.2 Statistical Methods

In the previous chapter, we presented a method to estimate the precision in a SW-CRT with varying cluster by simulating potential cluster-period sizes. As such, we do not repeat all of the methods here, but some will be reiterated for clarity. We begin with the analytical model used for a SW-CRT.

6.2.1 Analytical model used to analyse SW-CRTs

As highlighted in the previous chapter (section 5.2.1), the analysis of a SW-CRT typically involves the fitting of a generalised linear mixed model, as proposed by Hussey and Hughes (32), of the form:

$$Y_{ijk} = \mu + \beta_j + X_{ij}\delta + \alpha_i + \varepsilon_{ijk} \qquad 6.1$$

Where, Y_{ijk} is the outcome for patient *k* in cluster *i* at time *j*, μ is the mean outcome in the unexposed period in the first time-period, β_j is a fixed time effect for each time-period *j* = 2,…, *T* ($\beta_1 = 0$ for identifiability), δ is the treatment effect, α_i is a random effect for cluster *i*, ε_{ijk} is the residual error and X_{ij} is an indicator of treatment, where:

$$X_{ij} \begin{cases} 1 & \text{if cluster } i \text{ is exposed to the intervention at time } j \\ 0 & \text{if cluster } i \text{ is not exposed to the intervention at time } j \end{cases}$$

There are several assumptions being made when using this model framework (see section 5.2.1).

When following the above model (6.1), the power of a SW-CRT to detect a specified difference (δ) can be estimated using the design matrix (Z) and the variance-covariance matrix (V) (see section 5.2.2) as (32):

$$\phi\left(\left(\frac{\delta}{\sqrt{(Z'V^{-1}Z)^{-1}[1,1]}}\right) - Z_{1-\alpha/2}\right) \qquad 6.2$$

Where ϕ is the cumulative standard Normal distribution, [1,1] refers to the entry in the first column and first row, and $Z_{1-\alpha/2}$ is the $(1-\alpha/2)^{th}$ quantile of the standard Normal distribution function.

6.2.2 Estimating power in a SW-CRT with varying cluster sizes

In the design stage of a SW-CRT, the cluster sizes are either known or unknown – with existing methods for estimating the power able to handle either. Below, we use an example to illustrate the steps taken to estimate the power in a SW-CRT, which is conditional on whether the cluster sizes are known or unknown.

6.2.2.1 *Varying cluster size with known cluster sizes*

In this section, we show how the power can be estimated in a SW-CRT with known cluster sizes. Firstly, consider a cross-sectional SW-CRT in which there are 4 clusters, randomised over 4 steps, with one period for baseline measurements – resulting in 5 time-periods in which measurements are made. The primary outcome is BMI z-score, for which the mean in the unexposed period is 1.5 (SD = 1). A clinically important difference is determined as a lowering of BMI z-score of 0.25, and the ICC is 0.05.

We wish to estimate the power given that the cluster-period size (size of cluster at one time-period) for the four clusters A, B, C, and D are:

Cluster	**Cluster-period size**
A	10
B	50
C	100
D	500
Mean	165
Coefficient of variation	1.37

We are assuming that the clusters remain the same size at each time-period, and so there is no within-cluster variation in size. Since this is the cluster-period size, the total cluster size for a cluster (size of a cluster over whole study duration) will be 5 times its corresponding cluster-period size.

To estimate the power, the design matrix (Z), and the variance-covariance matrix of cluster means (V) are required (see section 5.2.2). Since this design consists of 4 clusters, randomised over 4 steps, the design pattern matrix (DPM) and the design matrix Z will be as follows:

Figure 6.1: Design pattern matrix (DPM) for four cluster example

$$\begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \end{array} \begin{array}{c} \begin{array}{ccccc} t_1 & t_2 & t_3 & t_4 & t_5 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Figure 6.2: Design matrix Z for four cluster example

$$Z = \begin{array}{c} \begin{array}{cccccc} I & t_1 & t_2 & t_3 & t_4 & t_5 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

The variance-covariance matrix of the cell means over time, matrix V, is composed of C matrices, given as V_i, where V_i is:

$$V_i = \begin{pmatrix} \frac{\sigma_w^2}{m_{i1}} + {\sigma_b}^2 & {\sigma_b}^2 & \cdots & {\sigma_b}^2 \\ {\sigma_b}^2 & \frac{\sigma_w^2}{m_{i2}} + {\sigma_b}^2 & \cdots & {\sigma_b}^2 \\ \vdots & \vdots & \ddots & \vdots \\ {\sigma_b}^2 & {\sigma_b}^2 & \cdots & \frac{\sigma_w^2}{m_{iT}} + {\sigma_b}^2 \end{pmatrix}$$

The term m_{ij} refers to the size of cluster *i* at time *j*, so that, for example, $m_{3\,4}$ refers to the size of cluster 3 at time-period 4. The variance components ${\sigma_b}^2$ and ${\sigma_w}^2$ are derived from the ICC and the variance of the outcome (σ^2) as:

$${\sigma_b}^2 = ICC \times \sigma^2$$

$${\sigma_w}^2 = \sigma^2 - {\sigma_b}^2 = \sigma^2(1 - ICC)$$

The power is then calculated using equation 6.2.

Typically, SW-CRTs assume equal sized clusters, and so all values of m_{ij} are identical. Therefore, the variance-covariance matrix *V* will be the same for each randomisation order, and so this order will not impact the power. When the cluster sizes vary, the matrices V_i will differ depending on the randomisation order (since the cluster-period sizes, m_{ij}, differ), and so will matrix *V*. Consequently, the power will also depend on the randomisation order. This is established in the previous chapter in terms of precision – whereby the precision can vary depending on the randomisation of clusters in a SW-CRT with varying cluster size. In a SW-CRT with 4 clusters, there are 24 (= 4 x 3 x 2 x 1) possible randomisation orders (Table 6.1).

Table 6.1: Possible randomisation orders for four clusters

A B C D	B A C D	C A B D	D A B C
A B D C	B A D C	C A D B	D A C B
A C B D	B C A D	C B A D	D B A C
A C D B	B C D A	C B D A	D B C A
A D B C	B D A C	C D A B	D C A B
A D C B	B D C A	C D B A	D C B A

Below, we show how the power can be determined for a given randomisation order, e.g. BACD, which indicates that:

$$m_{1j} = 50, \qquad m_{2j} = 10, \qquad m_{3j} = 100, \qquad m_{4j} = 500$$

Assuming that there is no within-cluster variation – that is, a particular cluster is the same size at each time-period – then these values of m_{ij} can then be inputted into each matrix V_i as follows:

$$V_1 = \begin{pmatrix} \frac{\sigma_w^2}{50} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & \frac{\sigma_w^2}{50} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{50} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{50} + \sigma_b & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{50} + \sigma_b \end{pmatrix} \qquad V_2 = \begin{pmatrix} \frac{\sigma_w^2}{10} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & \frac{\sigma_w^2}{10} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{10} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{10} + \sigma_b & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{10} + \sigma_b \end{pmatrix}$$

$$V_3 = \begin{pmatrix} \frac{\sigma_w^2}{100} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & \frac{\sigma_w^2}{100} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{100} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{100} + \sigma_b & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{100} + \sigma_b \end{pmatrix} \qquad V_4 = \begin{pmatrix} \frac{\sigma_w^2}{500} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & \frac{\sigma_w^2}{500} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{500} + \sigma_b & {\sigma_b}^2 & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{500} + \sigma_b & {\sigma_b}^2 \\ {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & {\sigma_b}^2 & \frac{\sigma_w^2}{500} + \sigma_b \end{pmatrix}$$

From this, we combine these four V_i matrices, into the variance-covariance matrix *V*:

$$V = \begin{pmatrix} V_1 & 0 & 0 & 0 \\ 0 & V_2 & 0 & 0 \\ 0 & 0 & V_3 & 0 \\ 0 & 0 & 0 & V_4 \end{pmatrix}$$

The matrix *V* can then be used with the design matrix, *Z*, to estimate the power using equation 6.2. This randomisation order (BACD) would provide 83.16% to detect a 0.25

difference in BMI z-scores. Below, we have repeated this for each of the 24 possible randomisation orders (Table 6.2).

Table 6.2: Estimates of the power for each randomisation order

A B C D *82.37%*	**B A C D** *83.16%*	**C A B D** *87.02%*	**D A B C** *89.04%*
A B D C *78.45%*	**B A D C** *82.45%*	**C A D B** *86.37%*	**D A C B** *88.98%*
A C B D *86.30%*	**B C A D** *88.98%*	**C B A D** *89.04%*	**D B A C** *87.02%*
A C D B *78.33%*	**B C D A** *81.52%*	**C B D A** *85.71%*	**D B C A** *86.30%*
A D B C *85.71%*	**B D A C** *86.37%*	**C D A B** *82.45%*	**D C A B** *83.16%*
A D C B *81.52%*	**B D C A** *78.33%*	**C D B A** *78.45%*	**D C B A** *82.37%*

The order of randomisation can impact the value of the power when clusters vary in size (Table 6.2), with over 10% difference in power between the lowest (ACDB = 78.33%) and the highest value (CBAD = 89.04%). Each permutation order contains clusters randomised in a different order. Since the clusters vary in size, many of these designs will contain a different number of observations contributing to the exposed and unexposed periods, resulting in the power varying between the orders. As such, it is necessary to distinguish between two types of power – the conditional power, and the marginal power, which we define below.

The conditional power refers to an estimate of the power that is conditional on the randomisation order – and so is most useful post-randomisation. The conditional power is simple to estimate. Once the clusters are randomised to a particular implementation order, the clusters-period sizes (m_{ij}) are inputted into each matrix V_i in this order. The power can then be estimated using equation 6.2. Since the randomisation order is fixed, the conditional precision is a single value. Each estimate of the power given on Table 6.2 is an estimate of the conditional power for its corresponding randomisation order.

In a pre-trial power calculation for a SW-CRT with varying cluster size, one value of the power – obtained from one particular randomisation order – may not be sufficient. Instead, the power should consider the full set of randomisation orders, since each order will have a unique estimate of the power. The marginal power is defined as the median of the conditional powers. The methodology used to estimate the conditional power is repeated for each randomisation order, and the values of the conditional power are then collated and averaged. Since each conditional power is unique, it is possible to consider the conditional powers as a distribution of the possible values of the power that considers every randomisation order of the clusters. As such, when reporting a marginal power – which is estimated from the distribution of conditional powers – an IQR is also reported to highlight the distribution. In our hypothetical example, the marginal power is 84.44% [IQR: 81.95% to 86.70%].

6.2.2.2 *Varying cluster size with unknown cluster sizes*

In Chapter 5, we described a method to estimate the precision and power, in a SW-CRT with unknown cluster sizes. For this an estimate of the true cluster-period size (μ) and the coefficient of variation of cluster sizes (cv) were required. The full methodology is not repeated here, but an overview of the methods used to simulate cluster sizes, and estimate the power is highlighted below. The methodology is split into SW-CRTs with between-cluster variation in cluster sizes only, and SW-CRTs with within-cluster and between-cluster variation in cluster sizes. The steps taken to estimate the power in a SW-CRT are summarised in Box 6.1 and Box 6.2.

Box 6.1: Outline of the steps taken to estimate the power of a stepped-wedge cluster randomised trial with varying cluster size

Background: **Define the SW-CRT design**

Define the design of the SW-CRT such as number of time-periods, number of clusters, number of steps, the cluster-period size, the ICC, the difference to be detected, the significance level, and the variance of the outcome.

Step 1: **Simulate cluster-period sizes**

Simulate for each cluster, its cluster-period size, m_{ij} such that $m_{ij} = m_{i`j} \; \forall \; i, i`$ using a gamma distribution, with mean μ and coefficient of variation, cv.

Step 2: **Apply a scaling factor**

- Calculate the average cluster-period size ($\bar{m}_{\bullet\bullet}$) as:

$$\bar{m}_{\bullet\bullet} = \frac{1}{CT} \sum_{i=1}^{C} \sum_{j=1}^{T} m_{ij}.$$

- Calculate ω which is the ratio of the total sample size in the unequal cluster size design (S_1) to the total sample size in an equal cluster size design (S_2) as:

$$\omega = \frac{S_1}{S_2} = \frac{\bar{m}_{\bullet\bullet} \times C \times T}{\mu \times C \times T} = \frac{\bar{m}_{\bullet\bullet}}{\mu}.$$

- Scale the cluster-period sizes m_{ij}, as $m_{ij} \times \omega$.

Box 6.1 continued…

Step 3: Estimate the power

- Calculate matrices V_i (i=1, ... , C) using the ωm_{ij} from step (2).
- Form matrix V using the block matrices V_i .
- Calculate the power using equation 6.2.

Step 4: Compile the power estimates

- Repeat steps (1) - (3) for a specified number of repetitions.
- Collate the estimates of the power to form the distribution (and average) of the power. The conditional power is then reported.

Box 6.2: Outline of the steps taken to estimate the power of a stepped-wedge cluster randomised trial with varying cluster size.

Background: Define the SW-CRT design

Define the design of the SW-CRT such as number of time-periods, number of clusters, number of steps, the cluster-period size, the ICC, the difference to be detected, the significance level, and the variance of the outcome.

Step 1: Simulate cluster-period sizes with no within-cluster variation in size

Simulate for each cluster, its cluster-period size, m_{ij} such that $m_{ij} = m_{ij`}\ \forall\, i, i`$ using a gamma distribution, with mean μ and coefficient of variation, cv.

Box 6.2 continued…

Step 2: **Apply a scaling factor**

- Calculate the average cluster-period size ($\bar{m}_{\bullet\bullet}$) as:

$$\bar{m}_{\bullet\bullet} = \frac{1}{CT}\sum_{i=1}^{C}\sum_{j=1}^{T} m_{ij}.$$

- Calculate ω which is the ratio of the total sample size in the unequal cluster size design (S_1) to the total sample size in an equal cluster size design (S_2) as:

$$\omega = \frac{S_1}{S_2} = \frac{\bar{m}_{\bullet\bullet} \times C \times T}{\mu \times C \times T} = \frac{\bar{m}_{\bullet\bullet}}{\mu}.$$

- Scale the cluster-period sizes m_{ij}, as $m_{ij} \times \omega$.

- Calculate for each cluster, the average cluster-period size $\overline{\omega m}_{i\bullet}$ as:

$$\overline{\omega m}_{i\bullet} = \frac{1}{T}\sum_{j=1}^{T} \omega m_{ij}$$

Step 3: **Simulate cluster-period sizes with within-cluster variation in size**

Simulate, for each cluster independently, a new cluster-period size (m_{ij}) using a Gamma distribution with mean $\overline{\omega m}_{i\bullet}$ and coefficient of variation cv_w.

Step 4: **Apply a scaling factor**

- For cluster i, calculate the average cluster-specific cluster-period size ($\bar{m}_{i\bullet}$), as:

$$\bar{m}_{i\bullet} = \frac{1}{T}\sum_{j=1}^{T} m_{ij}$$

Box 6.2 continued…

- For cluster i, calculate ω_i which is the ratio of the total cluster size in a design with within-cluster variation in size over time $\left(M_{2_i}\right)$ to the total cluster size in a design with no within-cluster variation in size over time $\left(M_{2_i}\right)$ as:

$$\omega_i = \frac{M_{2_i}}{M_{1_i}} = \frac{\overline{\omega m}_{\mathrm{i}\bullet} \times T}{\overline{m}_{\mathrm{i}\bullet} \times T} = \frac{\overline{\omega m}_{\mathrm{i}\bullet}}{\overline{m}_{i\bullet}}$$

- Scale the cluster-period sizes m_{ij}, as $\omega_i m_{ij}$.

Step 5: **Estimate the power**

- Calculate matrices V_i (i=1, … , C) using $\omega_i m_{ij}$ from step (4).
- Form matrix V using the block matrices V_i .
- Calculate the power using equation 6.2.

Step 6: **Compiling the power estimates**

- Repeat steps (1) - (5) for a specified number of repetitions.
- Collate the estimates of the power to form the distribution (and average) of the power. The conditional power is then reported.

Regardless of whether the variation in cluster sizes includes within-cluster variation or not, the methodology described above requires a large number of simulations to form a distribution of the potential values of the power. From this distribution, the median power can be reported, alongside the interquartile range to describe its likely dispersion.

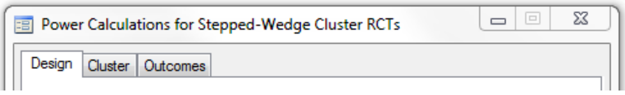
6.3 A function for estimating power in a SW-CRT

In this section, we describe the ***swpower*** Stata function that can be used to estimate the power in a SW-CRT with varying cluster size for a continuous outcome – using the methods described in section 6.2. The command has been developed in two parts – a command and a dialog box – and either can be used. Below, we discuss the dialog box that can be used, and highlight the output that is reported from the swpower function.

6.3.1 Dialog Box

The dialog box is made up of three tabs: Design, Clusters, and Outcomes (Figure 6.3).

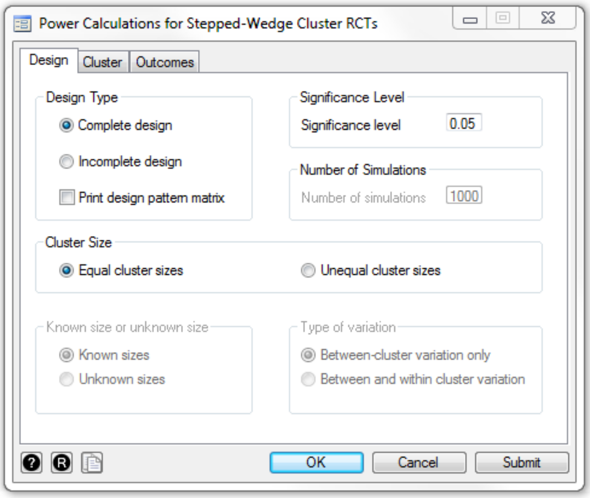
Figure 6.3: Tabs that form the *swpower* dialog box



6.3.1.1 *The Design tab*

The Design tab (given in Figure 6.4) allows the specification of the trial design, such as whether the design is complete or incomplete (see section 1.3.1). The DPM is required in the data editor for an incomplete design, and can, if selected, be printed alongside the results for confirmation. The Design tab requires a description of the cluster-period sizes. That is, whether the cluster-period sizes are equal or unequal and whether the sizes are known or unknown. Figure 6.5 shows a pathway that can be used to describe the cluster-period sizes.

Figure 6.4: Design tab from dialog box for *swpower* command



Power Calculations for Stepped-Wedge Cluster RCTs

Design | Cluster | Outcomes

Design Type
- ● Complete design
- ○ Incomplete design
- □ Print design pattern matrix

Significance Level
- Significance level: 0.05

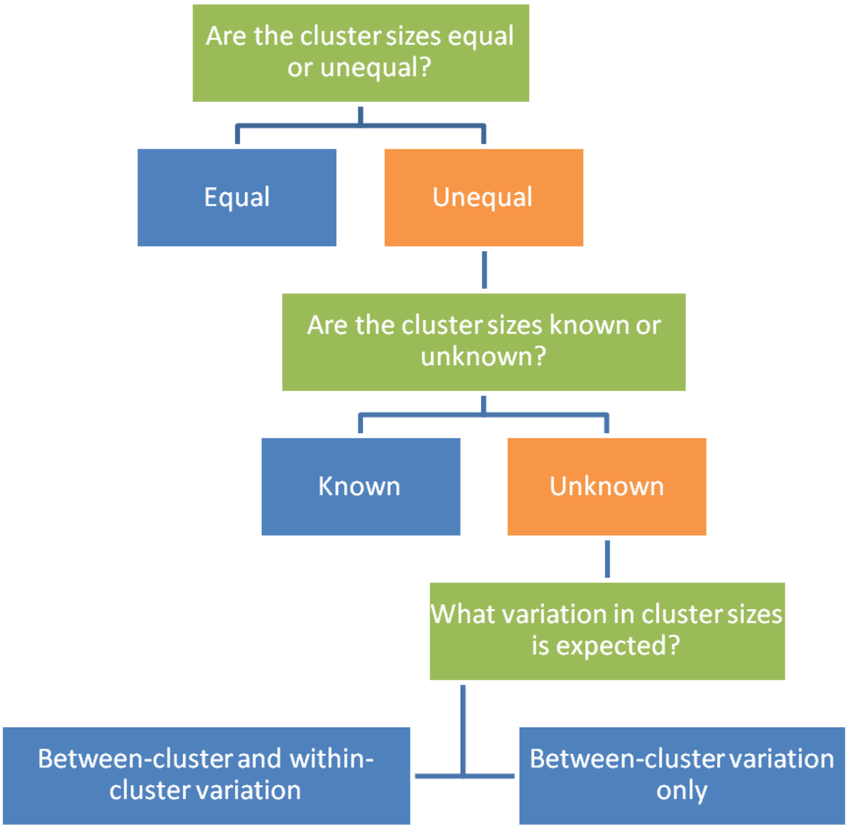
Number of Simulations
- Number of simulations: 1000

Cluster Size
- ● Equal cluster sizes
- ○ Unequal cluster sizes

Known size or unknown size
- ● Known sizes
- ○ Unknown sizes

Type of variation
- ● Between-cluster variation only
- ○ Between and within cluster variation

Figure 6.5: Pathway to describing cluster sizes


Are the cluster sizes equal or unequal?
Equal
Unequal
Are the cluster sizes known or unknown?
Known
Unknown
What variation in cluster sizes is expected?
Between-cluster and within-cluster variation
Between-cluster variation only

Green boxes are questions to be asked, orange boxes are partial answers, and blue boxes are the final description of cluster sizes.

Once it is known which description of the cluster sizes is appropriate, the necessary options can be selected on the swpower function. A summary of the cluster sizes, a description of them and how they are selected on the swpower dialog box is given in Table 6.3.

Table 6.3: Description of the cluster sizes are how they are selected on the swpower dialog box

Cluster sizes	Description	Selected on swpower function
Equal	The cluster are all of equal size, and are the same size during each time-period.	Select "equal cluster sizes" in the cluster size section.
Unequal clusters sizes with known cluster sizes	The cluster are unequal in size, but the size of each cluster is already known.	Select "unequal cluster size" in cluster size section and then "Known sizes".
Unequal cluster sizes, unknown cluster sizes, between-cluster variation in cluster sizes only	The clusters are unequal in size, but the size of each cluster is not known – but an average cluster-period size is known and a measure of the variation between cluster sizes. Each cluster remains the same size during each time-period.	Select "unequal cluster size" in cluster size section, then "Unknown sizes", and "Between-cluster variation only".
Unequal cluster sizes, unknown cluster sizes, between-cluster and within-cluster variation in cluster sizes	The clusters are unequal in size, but the size of each cluster is not known. Clusters vary in size between clusters but individual clusters also vary across time. An average cluster-period size, the between-cluster and within-cluster coefficient of variation of cluster sizes are required.	Select "unequal cluster size" in cluster size section, then "Unknown sizes", and "Between and within cluster variation".

The selections chosen on the Design tab influences the required input on the other tabs. The user can specify the significance level (default is 0.05) and, when appropriate, the number of simulations to be performed (default is 1000). The number of simulations is only required for an unequal cluster size scenario (either known or unknown). If the cluster-period sizes are known, then the number of simulations (N) specified will lead to either the conditional power being calculated (if N = 1) or the marginal power (if N>1).

6.3.1.2 *The Cluster tab*

The Cluster tab (Figure 6.6) is used to enter information regarding the clusters and is split into three sections: cluster information, cluster correlation, and the description of varying clusters (i.e. the specification of the between-cluster coefficient of variation and within-cluster coefficient of variation). The options available here are influenced by the selection on the Design tab, and a summary of them is given on Table 6.4. The ICC is required for all designs. For unequal cluster designs with unknown cluster sizes, the within-cluster coefficient of variation is only required if the user has specified between-cluster and within-cluster variation on the Design tab. A design with unequal cluster sizes and known sizes requires the user to specify: the number of time-periods (including baseline) (*T*), the number of clusters randomised per step (*S*), and the ICC. The number of clusters specified in the data editor must be equal to $(T-1) \times S$.

Figure 6.6: Cluster tab from dialog box for *swpower* command

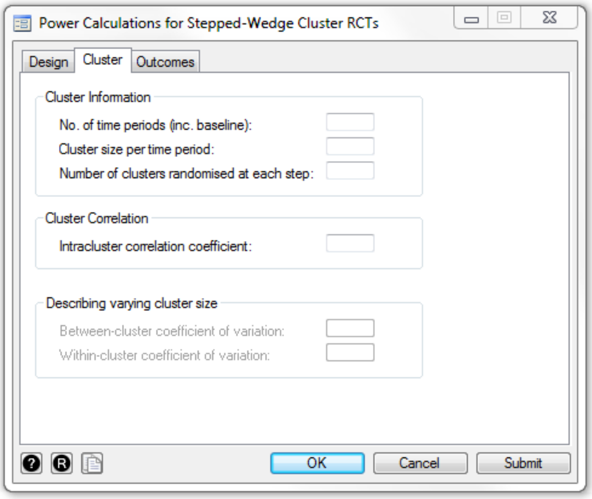

Power Calculations for Stepped-Wedge Cluster RCTs
Design
Cluster
Outcomes
Cluster Information
No. of time periods (inc. baseline):
Cluster size per time period:
Number of clusters randomised at each step:
Cluster Correlation
Intracluster correlation coefficient:
Describing varying cluster size
Between-cluster coefficient of variation:
Within-cluster coefficient of variation:
OK
Cancel
Submit

Table 6.4: Variables required for each design based on whether the cluster sizes are equal or unequal and whether the cluster sizes are known or unknown

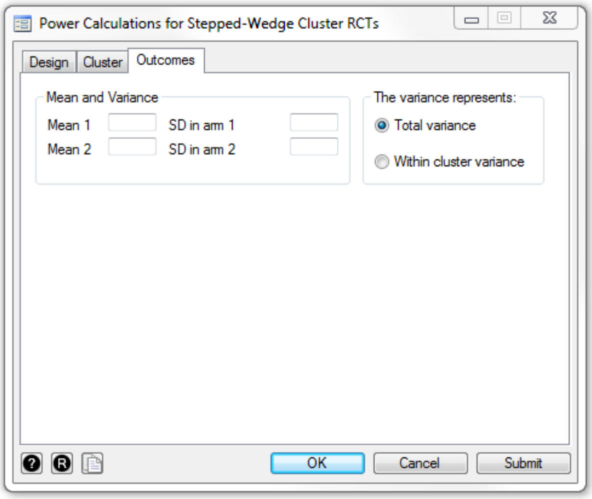
Equal or unequal cluster size	Known or unknown cluster sizes	Complete or incomplete design	Required variables
Equal	N/A	Complete	Time-periods, Cluster size, Number of clusters randomised per step, ICC.
Equal	N/A	Incomplete	Cluster size, Number of clusters randomised per step, ICC.
Unequal	Known	Complete	Time-periods, Number of clusters randomised per step, ICC.
Unequal	Known	Incomplete	ICC.
Unequal	Unknown	Complete	Time-periods, Cluster size, Number of clusters randomised per step, ICC, Between-cluster cv, Within-cluster cv (if appropriate).
Unequal	Unknown	Incomplete	Cluster size, Number of clusters randomised per step, ICC, Between-cluster cv, Within-cluster cv (if appropriate).

Note: an incomplete design is a design in which observations for one or more time-periods do not contribute towards the analysis

6.3.1.3 *The Outcomes tab*

The Outcomes tab (Figure 6.7) allows the user to specify the mean and standard deviation for the intervention and control conditions in the study. It is also possible to indicate whether the standard deviation corresponds to the total variance of the outcome, or the within cluster variance only. The default option is for the total variance to be selected.

Figure 6.7: Outcomes tab from dialog box for *swpower* command


Power Calculations for Stepped-Wedge Cluster RCTs

Mean and Variance

Mean 1 SD in arm 1

Mean 2 SD in arm 2

The variance represents:

● Total variance

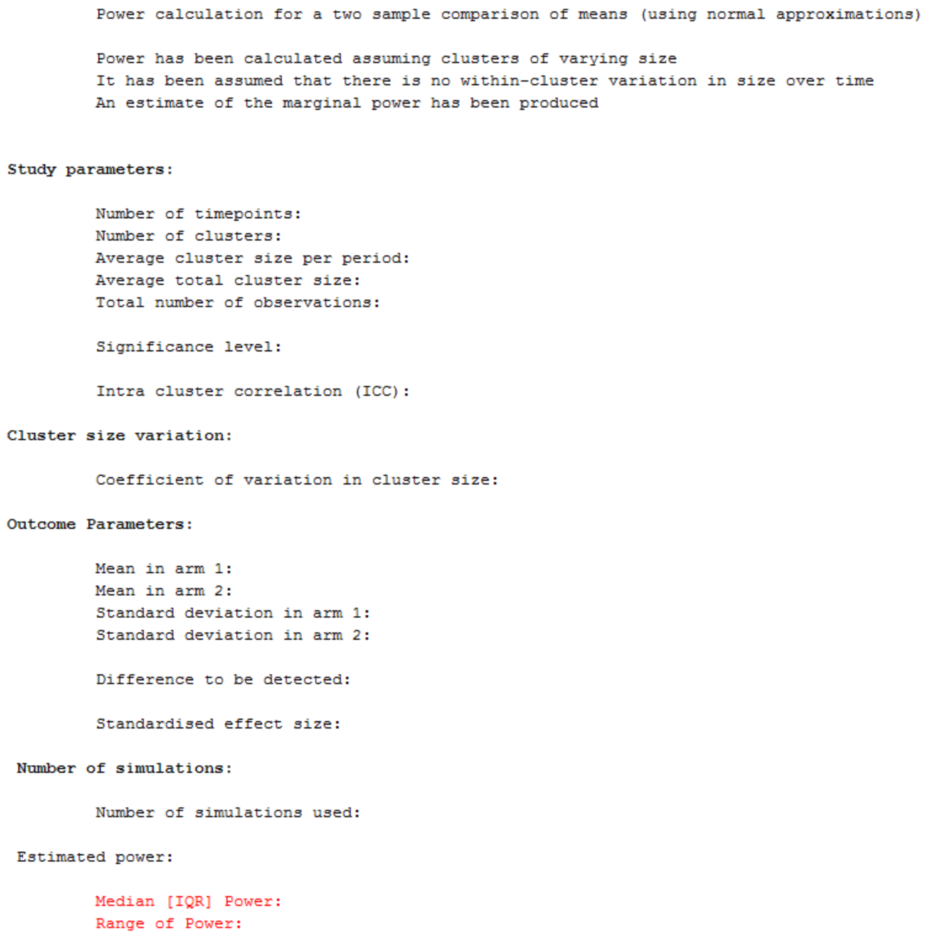
○ Within cluster variance

If the cluster sizes are known, then the (conditional or marginal) power is estimated following the methods described in section 6.2.2.1. If the cluster sizes are unknown, then the power is estimated using the methods described in section 6.2.2.2. Following this, a description of the study design and the power is presented in the window – which we discuss below.

6.3.2 Results output

Upon submitting the command, the results are shown in the Stata window (Figure 6.8). It is highlighted that the power calculation is for a two sample comparison of means using normal approximations and whether equal or unequal cluster sizes have been used. If selected, the DPM is then printed, though the printing of this matrix may not be appropriate for studies with a large number of clusters or time-periods. The main output is separated into five sections – study parameters, cluster size variation, outcome parameters, number of simulations, and the estimated power. The study parameters are printed for confirmation, and include: the number of time-periods, the number of clusters, the (average) cluster size per period, the (average) total cluster size, the (expected) total number of observations, the significance level, and the ICC. The cluster size variation reports, when applicable, the between-cluster, and within-cluster, coefficients of variation. If equal cluster sizes are assumed, then a message will be displayed at the top of results to report this, and the cluster size variation section will not be outputted. The outcome parameters are grouped together, and include: the mean in arms 1 and 2, the standard deviation in arms 1 and 2, the difference to be detected, and the standardised effect size. The number of simulations section reports the number of simulations used in the estimation of power and is only reported for designs with unequal cluster size. The final grouping reflects the estimated power, which is highlighted in red text. For designs with unequal cluster size, the median power is presented, alongside the inter-quartile range and the range. For designs with equal cluster size, or if the number of simulations is one, then the conditional power is presented as a single value.

Figure 6.8: Example of the results output for *swpower* command



```
         Power calculation for a two sample comparison of means (using normal approximations)

         Power has been calculated assuming clusters of varying size
         It has been assumed that there is no within-cluster variation in size over time
         An estimate of the marginal power has been produced


Study parameters:

         Number of timepoints:
         Number of clusters:
         Average cluster size per period:
         Average total cluster size:
         Total number of observations:

         Significance level:

         Intra cluster correlation (ICC):

Cluster size variation:

         Coefficient of variation in cluster size:

Outcome Parameters:

         Mean in arm 1:
         Mean in arm 2:
         Standard deviation in arm 1:
         Standard deviation in arm 2:

         Difference to be detected:

         Standardised effect size:

 Number of simulations:

         Number of simulations used:

 Estimated power:

         Median [IQR] Power:
         Range of Power:
```

6.4 Practical Examples of estimating power

In this section, we describe how the ***swpower*** function can be used to estimate the power for a SW-CRT by illustrating three examples: a SW-CRT with equal cluster size; a SW-CRT with unequal cluster sizes with unknown cluster sizes; and a SW-CRT with unequal cluster sizes with known cluster sizes. An example of the dialog box will be shown for the first example. All further examples will only show any necessary input into the data editor. The Stata output displaying the results will be shown for all examples.

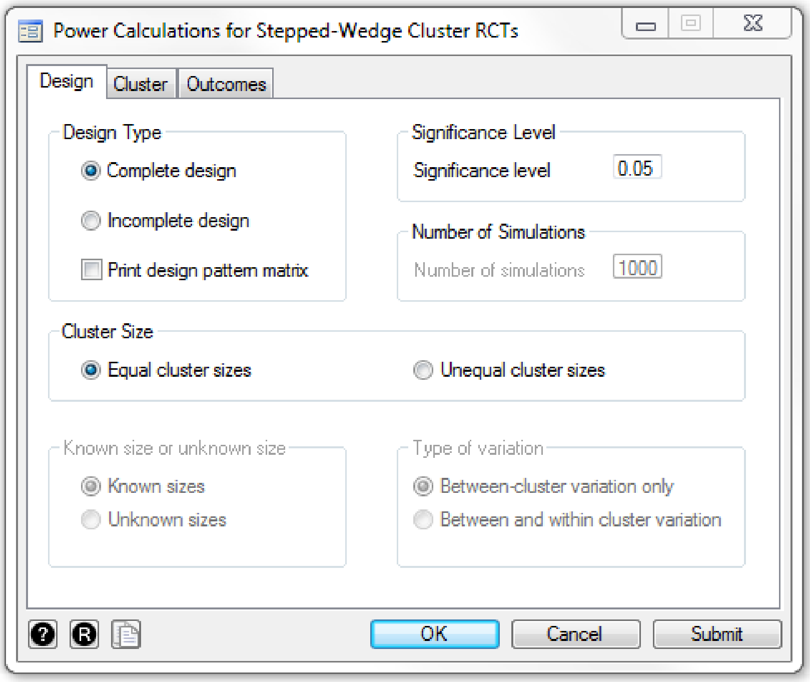
6.4.1 Equal sized clusters

The WOSLAD trial (see section 1.6.1) is an example of a SW-CRT in which the clusters are all equal in size (5). The study aimed to determine whether a hospital training programme increases women's satisfaction with doctor-patient relationship in labour and delivery rooms. The study recruited four hospitals (clusters), and expected the total cluster size to be 500 observations per cluster over the study duration, with observations made during 5 time-periods (100 observations per cluster per time-period).

The primary outcome is mean satisfaction score, with the mean (SD) in the control period expected to be 3.15 (0.75). A clinically important difference is a mean difference of 0.2. The power calculation was conducted using a 5% significance level. We present results with an ICC of 0.01.

For a SW-CRT design with equal cluster sizes, the Design (Figure 6.9), Cluster (Figure 6.10), and Outcomes (Figure 6.11) tabs are given below.

Figure 6.9: Design tab for a complete design with equal sized clusters


Power Calculations for Stepped-Wedge Cluster RCTs

Design Type

● Complete design

○ Incomplete design

□ Print design pattern matrix

Significance Level

Significance level: 0.05

Number of Simulations

Number of simulations: 1000

Cluster Size

● Equal cluster sizes ○ Unequal cluster sizes

Known size or unknown size

● Known sizes

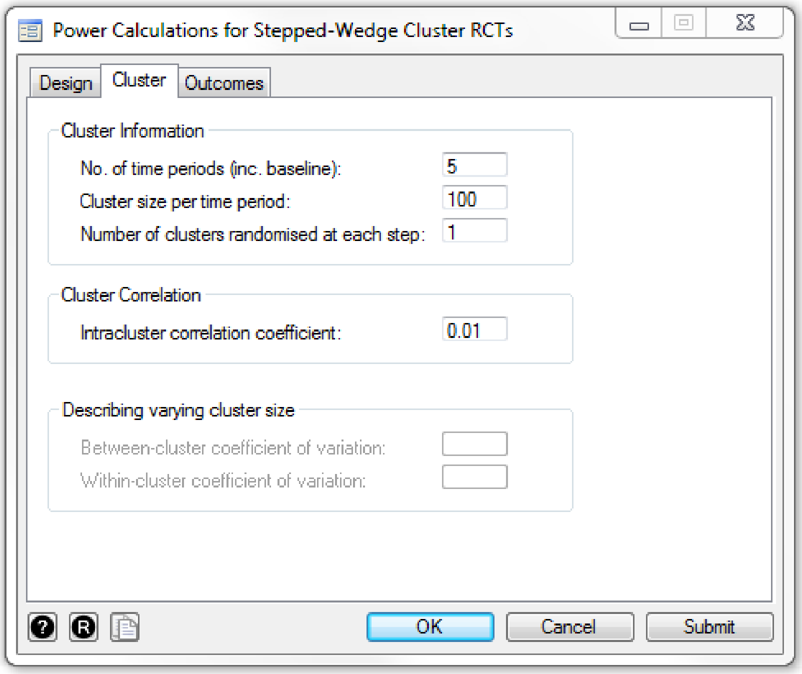
○ Unknown sizes

Type of variation

● Between-cluster variation only

○ Between and within cluster variation

Figure 6.10: Cluster tab for complete design with equal cluster sizes example


Power Calculations for Stepped-Wedge Cluster RCTs

Cluster Information

No. of time periods (inc. baseline): 5

Cluster size per time period: 100

Number of clusters randomised at each step: 1

Cluster Correlation

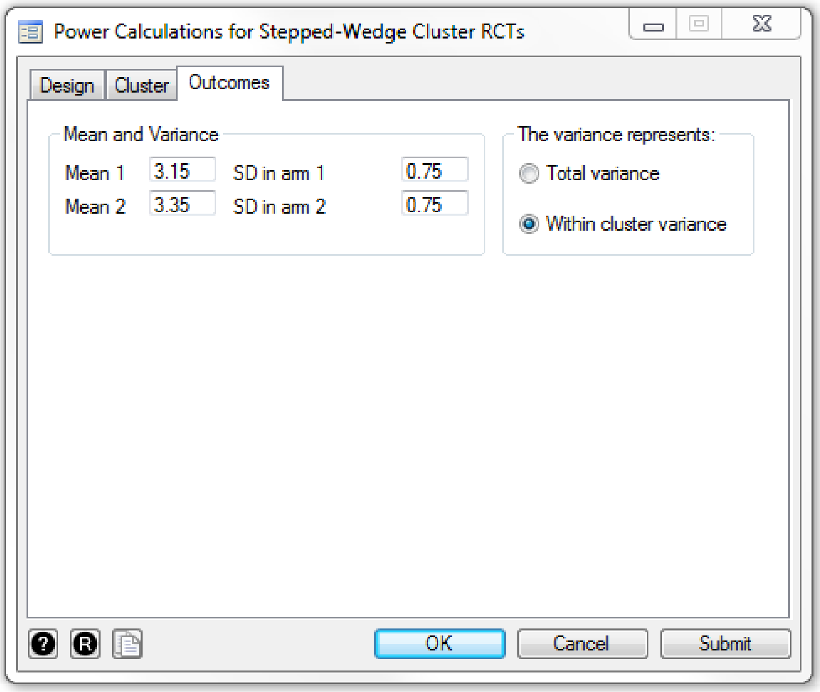
Intracluster correlation coefficient: 0.01

Describing varying cluster size

Between-cluster coefficient of variation:

Within-cluster coefficient of variation:

Figure 6.11: Outcomes tab for complete design with equal cluster sizes example



Mean and Variance

Mean 1	3.15	SD in arm 1	0.75
Mean 2	3.35	SD in arm 2	0.75

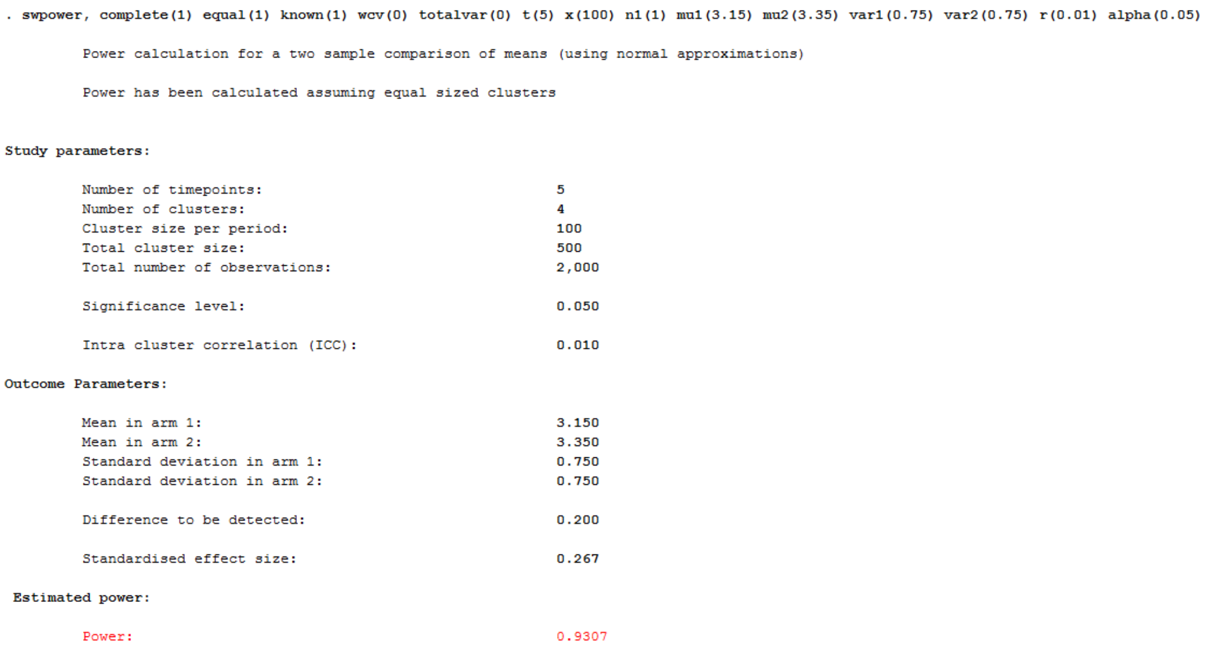
The variance represents:

○ Total variance

● Within cluster variance

The Stata output is given below (Figure 6.12). For the parameter values described, the SW-CRT would have 93% power to detect a 0.2 difference in means (standardised effect size = 0.27). Since the power is calculated using equal cluster size, the "Cluster size variation" and "Number of simulations" sections of the results are not displayed, since they are not applicable here.

Figure 6.12: Results output for complete design with equal cluster sizes example



```
. swpower, complete(1) equal(1) known(1) wcv(0) totalvar(0) t(5) x(100) n1(1) mu1(3.15) mu2(3.35) var1(0.75) var2(0.75) r(0.01) alpha(0.05)

          Power calculation for a two sample comparison of means (using normal approximations)

          Power has been calculated assuming equal sized clusters


Study parameters:

          Number of timepoints:                             5
          Number of clusters:                               4
          Cluster size per period:                          100
          Total cluster size:                               500
          Total number of observations:                     2,000

          Significance level:                               0.050

          Intra cluster correlation (ICC):                  0.010

Outcome Parameters:

          Mean in arm 1:                                    3.150
          Mean in arm 2:                                    3.350
          Standard deviation in arm 1:                      0.750
          Standard deviation in arm 2:                      0.750

          Difference to be detected:                        0.200

          Standardised effect size:                         0.267

 Estimated power:

          Power:                                            0.9307
```

6.4.2 Unequal sized clusters and unknown sizes

Our second example highlights how the swpower function can be used to estimate the power in a SW-CRT with clusters of unequal size, in which only the mean cluster size and a measure of the variation of cluster sizes are known.

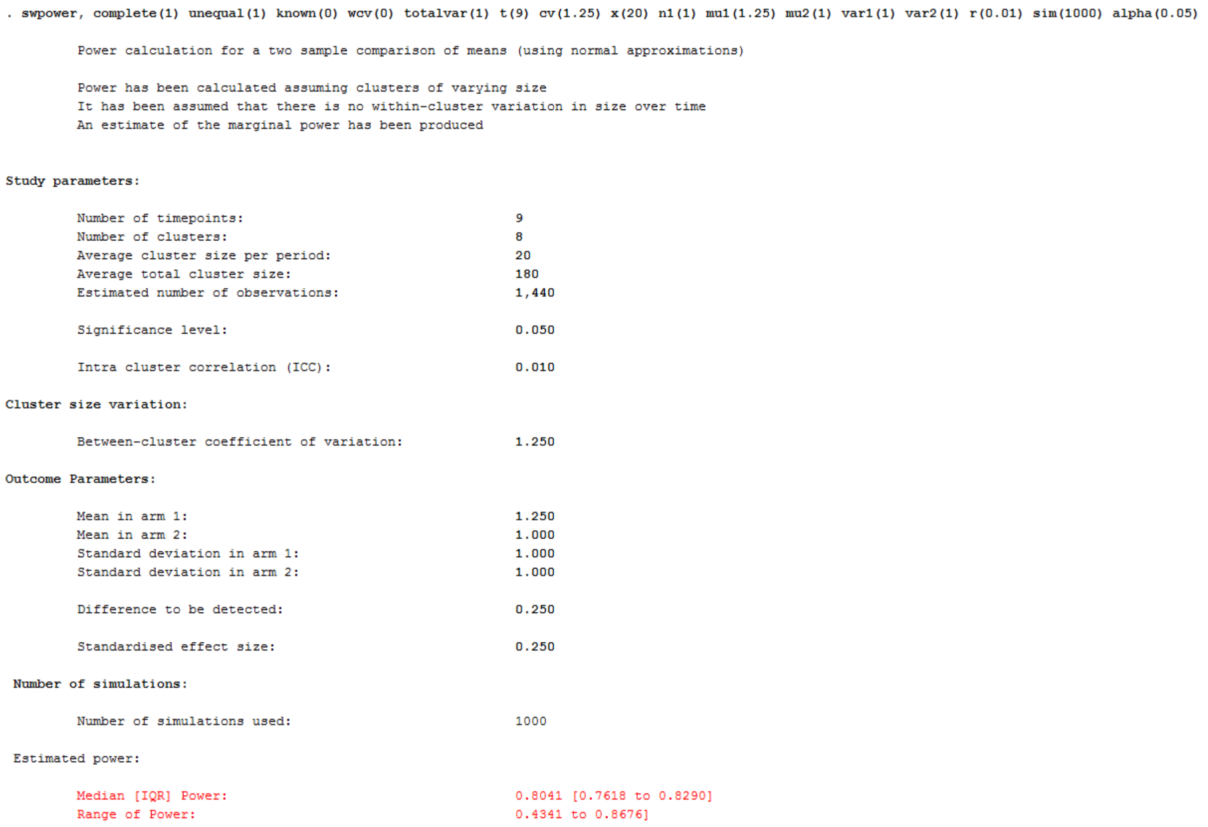
Consider a cross-sectional SW-CRT that plans to recruit 8 general practices (clusters) which will be randomised over 8 time-periods – so that the study contains 9 time-periods. Each practice contributes an average of 180 participants over the study duration, so that there are 20 participants per cluster per time-period on average. The study wants to lower the BMI Z-score from 1.25 (SD 1) in the control condition by 0.25. This corresponds to a standardised effect size of 0.25. The ICC used is 0.01, and the between-cluster coefficient of variation of cluster sizes is 1.25. The options selected on each tab to estimate the power is given by Table 6.5.

Table 6.5: Options selected on each tab to estimate the power for a SW-CRT with unequal cluster sizes that are unknown and fixed over time (no within-cluster variation in size)

Tab	Options Selected	Value inputted
Design	Complete design	
	Unequal cluster size	
	Unknown cluster sizes	
	Between-cluster variation only	
	Significance level	0.05
	Number of simulations	1000
Clusters	Time-periods	9
	Cluster size per time-period	20
	Clusters randomised per step	1
	ICC	0.01
	Between-cluster coefficient of variation	1.25
Outcomes	Mean 1	1.25
	Mean 2	1.00
	SD in arm 1	1.00
	SD in arm 2	1.00
	Within cluster variance	

In this example, the median power for the SW-CRT is 80.4% [IQR: 76.2% to 82.9%] (Figure 6.13). However, the full range of power is 45.4% to 86.8%. So whilst the average may be sufficient power for a study, the full range show how low the power may be, conditional on the randomisation order of the clusters.

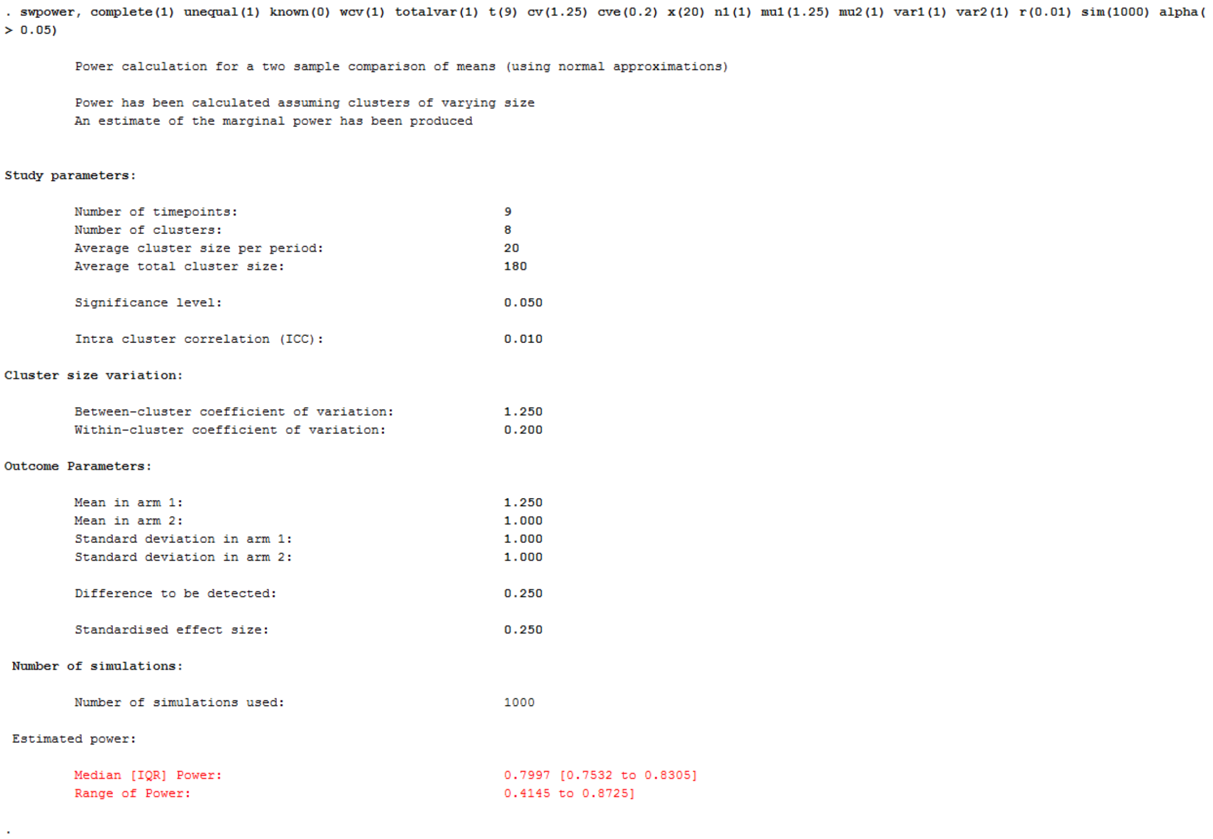
Figure 6.13: Results output for stepped-wedge design with unequal cluster sizes (unknown sizes) with no within-cluster variation in size over time



```
. swpower, complete(1) unequal(1) known(0) wcv(0) totalvar(1) t(9) cv(1.25) x(20) n1(1) mu1(1.25) mu2(1) var1(1) var2(1) r(0.01) sim(1000) alpha(0.05)

         Power calculation for a two sample comparison of means (using normal approximations)

         Power has been calculated assuming clusters of varying size
         It has been assumed that there is no within-cluster variation in size over time
         An estimate of the marginal power has been produced


Study parameters:

         Number of timepoints:                              9
         Number of clusters:                                8
         Average cluster size per period:                   20
         Average total cluster size:                        180
         Estimated number of observations:                  1,440

         Significance level:                                0.050

         Intra cluster correlation (ICC):                   0.010

Cluster size variation:

         Between-cluster coefficient of variation:          1.250

Outcome Parameters:

         Mean in arm 1:                                     1.250
         Mean in arm 2:                                     1.000
         Standard deviation in arm 1:                       1.000
         Standard deviation in arm 2:                       1.000

         Difference to be detected:                         0.250

         Standardised effect size:                          0.250

 Number of simulations:

         Number of simulations used:                        1000

 Estimated power:

         Median [IQR] Power:                                0.8041 [0.7618 to 0.8290]
         Range of Power:                                    0.4341 to 0.8676]
```

Suppose now that we wish to re-run the power calculation, but with the inclusion of a within-cluster coefficient of variation of 0.2. Now, “Between cluster and within cluster variation” is selected on the Design tab, and the within-cluster coefficient of variation in inputted on the Cluster tab. The average power is this SW-CRT with between-cluster and within-cluster variation in size is now 80.0% [IQR: 75.3% to 83.1%], and the range of power values is 41.5% to 87.3% (Figure 6.14). The inclusion of within-cluster variation has led to a slight decrease in the average power, and an increase in the possible variability of the power.

Figure 6.14: Results output for stepped-wedge design with unequal cluster sizes (unknown sizes) with within-cluster variation in size over time



```
. swpower, complete(1) unequal(1) known(0) wcv(1) totalvar(1) t(9) cv(1.25) cve(0.2) x(20) n1(1) mu1(1.25) mu2(1) var1(1) var2(1) r(0.01) sim(1000) alpha(
> 0.05)

         Power calculation for a two sample comparison of means (using normal approximations)

         Power has been calculated assuming clusters of varying size
         An estimate of the marginal power has been produced


Study parameters:

         Number of timepoints:                                9
         Number of clusters:                                  8
         Average cluster size per period:                     20
         Average total cluster size:                          180

         Significance level:                                  0.050

         Intra cluster correlation (ICC):                     0.010

Cluster size variation:

         Between-cluster coefficient of variation:            1.250
         Within-cluster coefficient of variation:             0.200

Outcome Parameters:

         Mean in arm 1:                                       1.250
         Mean in arm 2:                                       1.000
         Standard deviation in arm 1:                         1.000
         Standard deviation in arm 2:                         1.000

         Difference to be detected:                           0.250

         Standardised effect size:                            0.250

 Number of simulations:

         Number of simulations used:                          1000

 Estimated power:

         Median [IQR] Power:                                  0.7997 [0.7532 to 0.8305]
         Range of Power:                                      0.4145 to 0.8725]

.
```

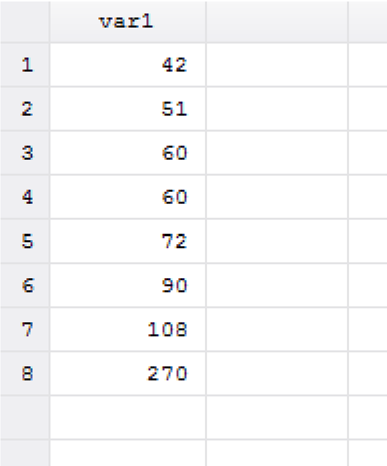
6.4.3 Unequal sized clusters and known sizes

In this section, we show how the marginal and conditional power can be calculated for a SW-CRT when the cluster sizes are known. This includes estimating the conditional power when the cluster sizes are known for every time-period and individual clusters vary in size over time.

6.4.3.1 *Estimating the marginal power*

Consider a SW-CRT that contains 8 hospitals (clusters) that will be randomised over 4 time-periods. The cluster sizes vary and are known – but each cluster remains fixed in size over time (no within-cluster variation). The study aims to decrease the waist Z-score in patients in a particular clinic from the current mean of 2 (SD 3.5) to 1.44 (SD 3.5). This corresponds to an effect size of 0.16. The power will be estimated, assuming an ICC of 0.002. The sizes of the hospitals in ascending order are: 42, 51, 60, 72, 90, 108, and 270, and these can be inputted into the Stata data editor, as shown by Figure 6.15.

Figure 6.15: Data editor for input of known cluster sizes



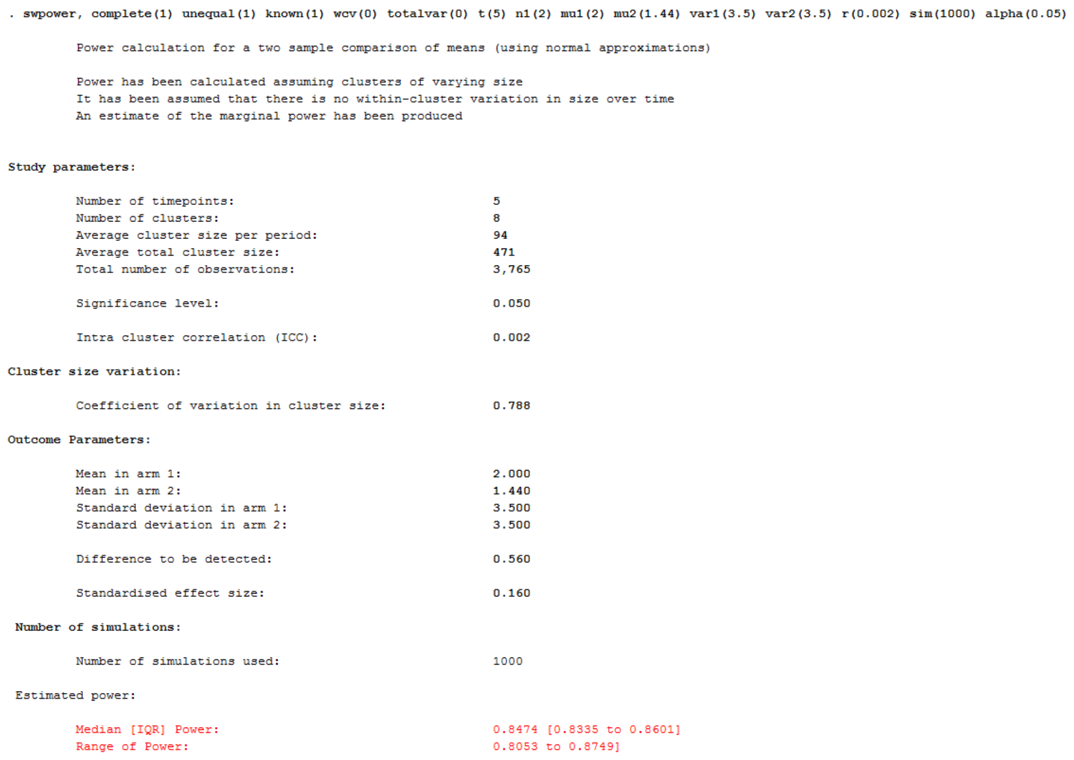
	var1
1	42
2	51
3	60
4	60
5	72
6	90
7	108
8	270

We plan to estimate the marginal power – to give an indication of the expected power for our study and the expected variation in this value. To estimate the marginal power, the options selected on the ***swpower*** tabs are given on Table 6.6. Although there are 8! (40,320) possible randomisation orders for the clusters, we consider a subset of them, and leave the number of simulations as its default value of 1000. The median power for this design would be 84.74% [IQR: 83.35% to 86.01%] (Figure 6.16), and the range of possible power values would be 80.5% to 87.5%.

Table 6.6: Options selected on each tab to estimate the marginal power for a SW-CRT with unequal cluster sizes that are known and fixed over time

Tab	Options Selected	Value inputted
Design	Complete design	
	Unequal cluster size	
	Known cluster sizes	
	Significance level	0.05
	Number of simulations	1000
Clusters	Time-periods	5
	Clusters randomised per step	2
	ICC	0.002
Outcomes	Mean 1	2.00
	Mean 2	1.44
	SD in arm 1	3.50
	SD in arm 2	3.50
	Within cluster variance	

Figure 6.16: Results output for complete design with unequal cluster sizes (marginal power) example



```
. swpower, complete(1) unequal(1) known(1) wcv(0) totalvar(0) t(5) n1(2) mu1(2) mu2(1.44) var1(3.5) var2(3.5) r(0.002) sim(1000) alpha(0.05)

         Power calculation for a two sample comparison of means (using normal approximations)

         Power has been calculated assuming clusters of varying size
         It has been assumed that there is no within-cluster variation in size over time
         An estimate of the marginal power has been produced


Study parameters:

         Number of timepoints:                            5
         Number of clusters:                              8
         Average cluster size per period:                 94
         Average total cluster size:                      471
         Total number of observations:                    3,765

         Significance level:                              0.050

         Intra cluster correlation (ICC):                 0.002

Cluster size variation:

         Coefficient of variation in cluster size:        0.788

Outcome Parameters:

         Mean in arm 1:                                   2.000
         Mean in arm 2:                                   1.440
         Standard deviation in arm 1:                     3.500
         Standard deviation in arm 2:                     3.500

         Difference to be detected:                       0.560

         Standardised effect size:                        0.160

 Number of simulations:

         Number of simulations used:                      1000

 Estimated power:

         Median [IQR] Power:                              0.8474 [0.8335 to 0.8601]
         Range of Power:                                  0.8053 to 0.8749]
```

6.4.3.2 *Estimating the conditional power*

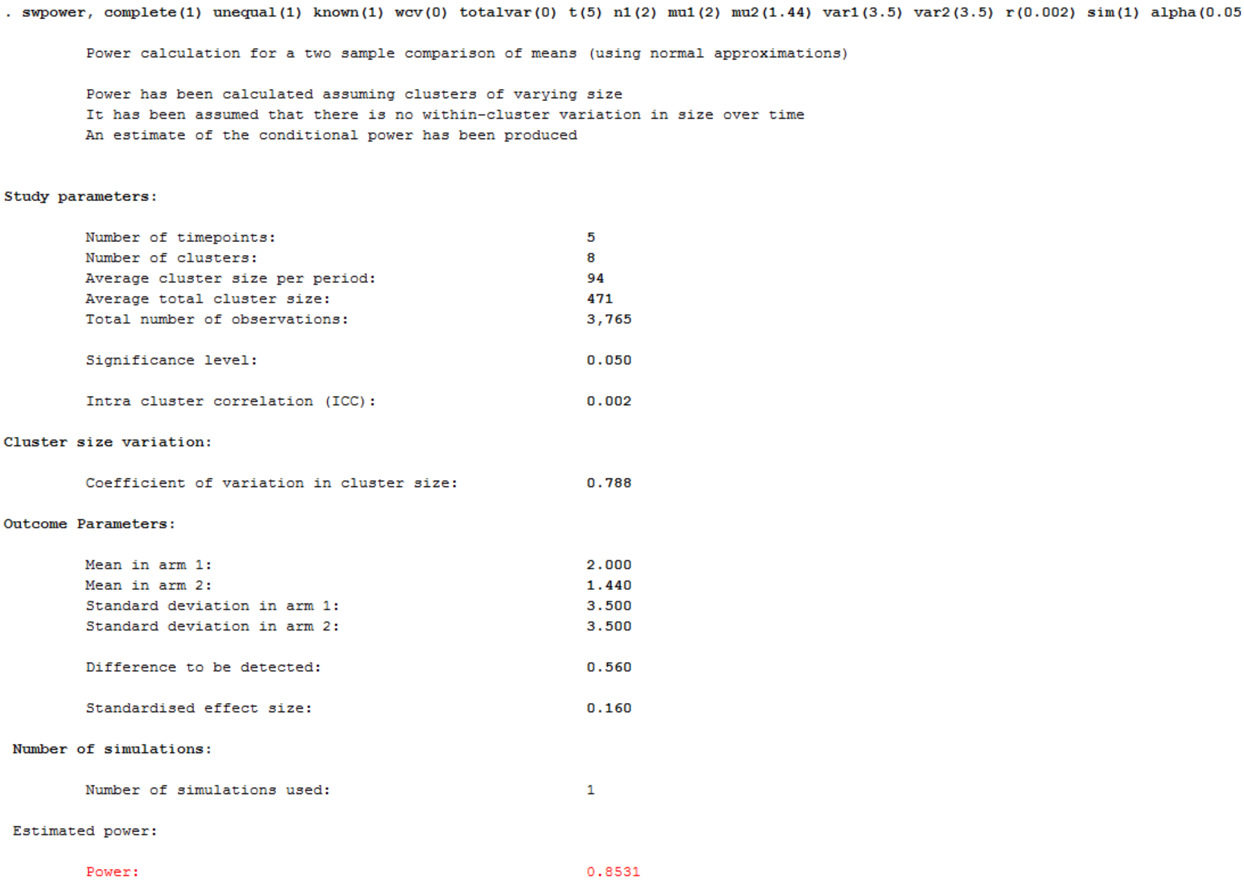
To estimate the power conditional on a randomisation order, there are two options within the ***swpower*** function – based on whether the cluster sizes vary in size over time or not. If there is no within-cluster variation in size over time, then the cluster sizes can be inputted into the first column of the data editor in the order they are randomised. See Figure 6.15 for example. Alternatively, if the cluster sizes vary over time – and these sizes are known – then the cluster sizes over every time-period can be inputted into the data editor. This can also be used post-trial to evaluate the power of a study.

If estimating the power conditional on the randomisation order given in Figure 6.15, assuming there is no within-cluster variation in size, then the options selected can be given by Table 6.7. This SW-CRT would have 85.31% power to detect the 0.16 effect size (Figure 6.17). Since this is a conditional power, there is no measure of variation around this value.

Table 6.7: Options selected on each tab to estimate the conditional power for a SW-CRT with unequal cluster sizes that are known and fixed over time

Tab	Options Selected	Value inputted
Design	Complete design	
	Unequal cluster size	
	Known cluster sizes	
	Significance level	0.05
	Number of simulations	1
Clusters	Time-periods	5
	Clusters randomised per step	2
	ICC	0.002
Outcomes	Mean 1	2.00
	Mean 2	1.44
	SD in arm 1	3.50
	SD in arm 2	3.50
	Within cluster variance	

Figure 6.17: Results output for complete design with unequal cluster sizes (conditional power) example with no variation in cluster size over time



```
. swpower, complete(1) unequal(1) known(1) wcv(0) totalvar(0) t(5) n1(2) mu1(2) mu2(1.44) var1(3.5) var2(3.5) r(0.002) sim(1) alpha(0.05

        Power calculation for a two sample comparison of means (using normal approximations)

        Power has been calculated assuming clusters of varying size
        It has been assumed that there is no within-cluster variation in size over time
        An estimate of the conditional power has been produced


Study parameters:

        Number of timepoints:                             5
        Number of clusters:                               8
        Average cluster size per period:                  94
        Average total cluster size:                       471
        Total number of observations:                     3,765

        Significance level:                               0.050

        Intra cluster correlation (ICC):                  0.002

Cluster size variation:

        Coefficient of variation in cluster size:         0.788

Outcome Parameters:

        Mean in arm 1:                                    2.000
        Mean in arm 2:                                    1.440
        Standard deviation in arm 1:                      3.500
        Standard deviation in arm 2:                      3.500

        Difference to be detected:                        0.560

        Standardised effect size:                         0.160

 Number of simulations:

        Number of simulations used:                       1

 Estimated power:

        Power:                                            0.8531
```

If the known cluster-period sizes vary in size over time, then the ***swpower*** function can be used to estimate the conditional power. In the above study with known cluster sizes, there are 8 clusters and 4 randomisation steps. We firstly consider the DPM for this study, given in Figure 6.18. We can also think of the size of each cluster over time as a matrix. Figure 6.19, for example, may describe the cluster-period size of each of cluster over time, in the order they are randomised in. That is, cluster C_1 and C_2 are randomised to receive the intervention in the second time-period, clusters C_3 and C_4 in the third time-period, and so on.

Figure 6.18: Design pattern matrix for an SW-CRT with 8 clusters and 4 steps

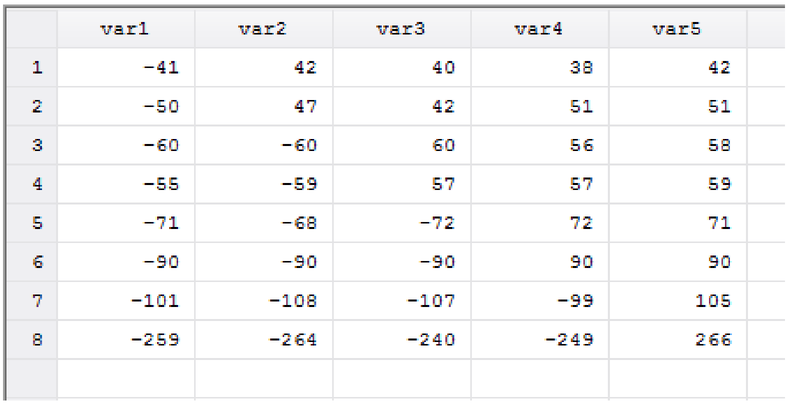
$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Figure 6.19: A matrix of cluster-period sizes for each cluster at each time-period

$$\begin{array}{cccccc} & T_1 & T_2 & T_3 & T_4 & T_5 \\ C_1 & 41 & 42 & 40 & 38 & 42 \\ C_2 & 50 & 47 & 42 & 51 & 51 \\ C_3 & 60 & 60 & 60 & 56 & 58 \\ C_4 & 55 & 59 & 57 & 57 & 59 \\ C_5 & 71 & 68 & 72 & 72 & 71 \\ C_6 & 90 & 90 & 90 & 90 & 90 \\ C_7 & 101 & 108 & 107 & 99 & 105 \\ C_8 & 259 & 264 & 240 & 249 & 266 \end{array}$$

In Figure 6.19, the individual clusters vary in size over time – though some clusters may not vary. To estimate the conditional power, the DPM, and the cluster-period sizes are combined to create a new matrix, which can be inputted into the data editor as shown in Figure 6.20. In this matrix, the “-“ (negative integer) indicates that the cluster is contribution to the control period at that time-period, whilst a “+” (positive integer) indicates that the cluster is contributing to the intervention period at that time-period. For example, a “-41” would indicate that the cluster is contribution 41 participants to the control, whilst a value of “(+)40” would indicate that the cluster is contributing 40 participants to the intervention. The positive and negative values are then used to generate the DPM, and in turn, the design matrix, Z, within the ***swpower*** function.

Figure 6.20: Data editor for input of known cluster sizes and incomplete design with variation in cluster size over time



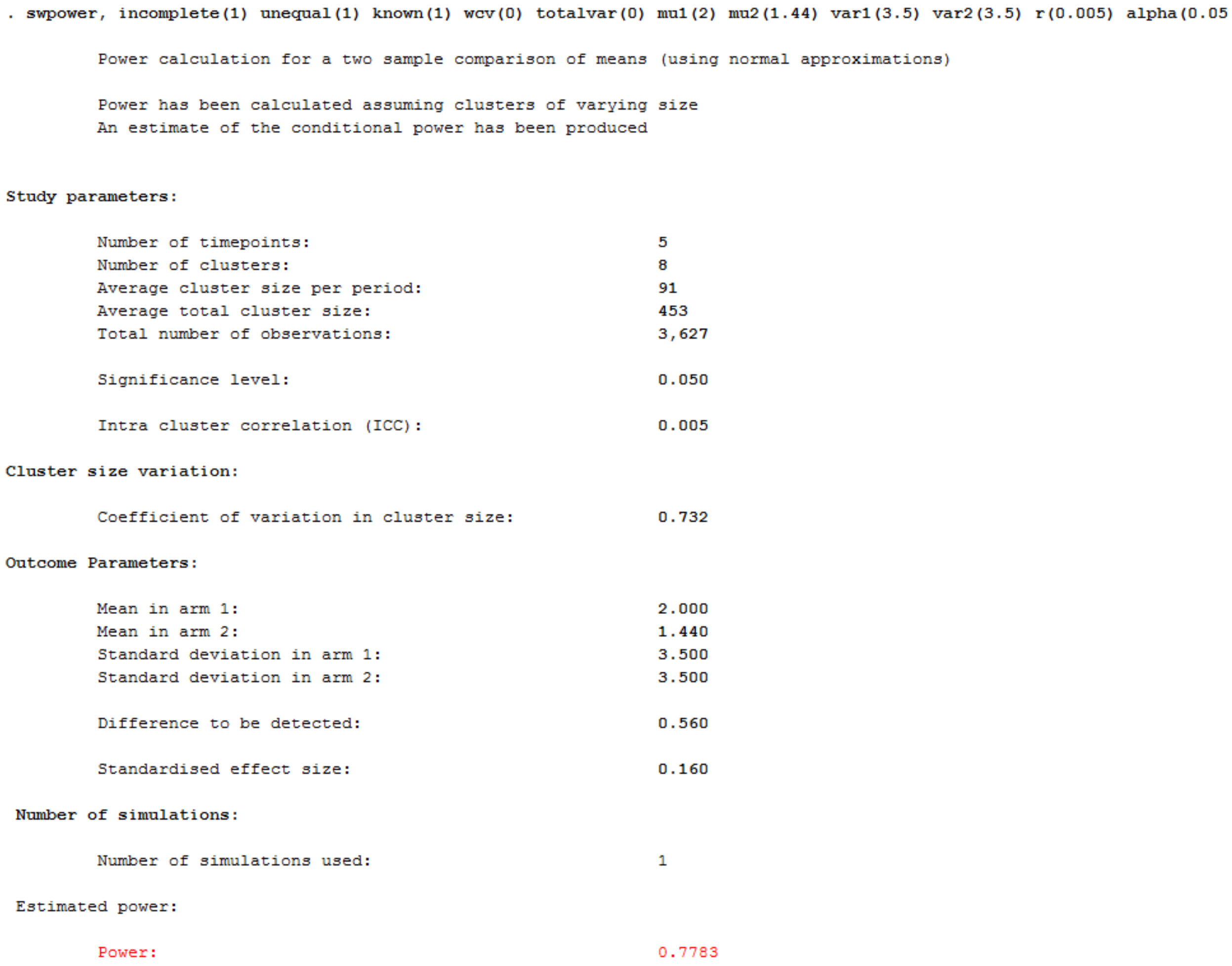
	var1	var2	var3	var4	var5
1	-41	42	40	38	42
2	-50	47	42	51	51
3	-60	-60	60	56	58
4	-55	-59	57	57	59
5	-71	-68	-72	72	71
6	-90	-90	-90	90	90
7	-101	-108	-107	-99	105
8	-259	-264	-240	-249	266

To estimate the conditional power for the known cluster-period sizes with within-cluster variation in size given in Figure 6.19, the options selected in the ***swpower*** function are given on Table 6.8. The incomplete design option is selected to inform the function that the cluster-period sizes for every time-period should be obtained from the data editor. This SW-CRT would have 77.83% power to detect the 0.16 effect size (Figure 6.21).

Table 6.8: Options selected on each tab to estimate the conditional power for a SW-CRT with unequal cluster sizes that are known and vary over time

Tab	Options Selected	Value inputted
Design	Incomplete design	
	Unequal cluster size	
	Known cluster sizes	
	Significance level	0.05
	Number of simulations	1
Clusters	Time-periods	5
	Clusters randomised per step	2
	ICC	0.002
Outcomes	Mean 1	2.00
	Mean 2	1.44
	SD in arm 1	3.50
	SD in arm 2	3.50
	Within cluster variance	

Figure 6.21: Results output for incomplete design with unequal cluster sizes (known cluster size) example with variation in cluster size over time



```
. swpower, incomplete(1) unequal(1) known(1) wcv(0) totalvar(0) mu1(2) mu2(1.44) var1(3.5) var2(3.5) r(0.005) alpha(0.05

         Power calculation for a two sample comparison of means (using normal approximations)

         Power has been calculated assuming clusters of varying size
         An estimate of the conditional power has been produced


Study parameters:

         Number of timepoints:                               5
         Number of clusters:                                 8
         Average cluster size per period:                    91
         Average total cluster size:                         453
         Total number of observations:                       3,627

         Significance level:                                 0.050

         Intra cluster correlation (ICC):                    0.005

Cluster size variation:

         Coefficient of variation in cluster size:           0.732

Outcome Parameters:

         Mean in arm 1:                                      2.000
         Mean in arm 2:                                      1.440
         Standard deviation in arm 1:                        3.500
         Standard deviation in arm 2:                        3.500

         Difference to be detected:                          0.560

         Standardised effect size:                           0.160

 Number of simulations:

         Number of simulations used:                         1

 Estimated power:

         Power:                                              0.7783
```

6.5 Discussion

In this chapter, we reiterated the method to estimate the power in a SW-CRT with varying cluster size – based on whether the cluster sizes were known or unknown. To this end, we presented the ***swpower*** Stata command, which has been developed to estimate the power in SW-CRTs with varying cluster size. Several examples were presented to show the practical use of the command for SW-CRT and P-CRTs with equal and unequal cluster sizes.

There is extensive research into the effect of varying cluster size in P-CRTs, which has led to numerous design effects that can be used to estimate the power (57, 58). However, the methodological research into SW-CRTs is still in its infancy, and so a much smaller pool of research has been published. Nevertheless, the number of SW-CRTs being conducted is steadily increasing (23, 28, 29). To aid power calculations, there is both a design effect (21) and Stata function (63) for SW-CRTs, though both assume that all clusters will be equal in size. Chapter 3 highlighted that varying cluster size is often not acknowledged in a SW-CRT, and there has been no evidence of a trial correctly acknowledging varying cluster size in a power calculation. In P-CRTs, it is known that a study with varying cluster sizes has less power than an identical study with equal sized clusters (57, 145).

In Chapter 5 it was shown that in relation to precision, the SW-CRT with varying cluster size may be less efficient, on average, than a SW-CRT with equal sized clusters. That is, the average power in a SW-CRT with varying cluster size will be lower than the power in a SW-CRT with equal cluster size. However, the variation in the potential power is also important. Here, we presented some examples to show how the swpower function can be used to estimate the power in a SW-CRT with varying cluster size. These examples showed that for

SW-CRTs with known and unknown cluster sizes that vary, the power can be extremely variable. When considering simply the average power and the inter-quartile range, a SW-CRT may be sufficiently powered. However, the full range of potential values of the power highlights its large variability, and emphasises the need to account for varying cluster size.

The methodological rigor of reported power calculations from SW-CRTs is sub-optimal, and many are conducted using inappropriate methods (Chapter 3) which could lead to an underpowered study. The Stata command ***swpower*** has been developed to ensure future power calculation use appropriate methodology and correctly acknowledge any variation in cluster sizes. The command is a useful tool that will allow the estimation of power in a SW-CRT with equal sized clusters, varying cluster sizes with unknown cluster sizes, and varying cluster sizes with known cluster sizes. To provide for greater usability, an accompanying dialog box has been produced to co-exist with the command. For greater flexibility, the function allows the input of the DPM, allowing trialists to indicate a design identical to their own study. The function also enables the power to be estimated for a P-CRT.

6.5.1 Limitations

This function only allows the estimation of power for studies, and does not allow for the computation of a sample size or detectable difference. As such, this command can aid researchers in estimating the power of a trial for a set of parameters, or allow the testing of the impact of cluster size variation on their trial, but will not provide an estimate of the necessary sample size.

Since the model used here is that presented by Hussey and Hughes (32), there is no random effect for the individual participants. As such, the model assumes that there is no correlation between the individuals at different time-periods and so is applicable only to a cross-sectional SW-CRT.

In Chapter 4, we introduced the concept of the within-period correlation and the inter-period correlation. They fundamentally rely on a model with a random cluster by time effect. Since the only random effect included in the model in the ***swpower*** function is for cluster, the within-period and inter-period correlations cannot be specified in this work. As such, it is assumed that the correlation between outcomes in a cluster is independent of the timing of the observations, so that the correlation does not decay over time.

6.5.2 Possible future adaptations

Currently, for designs in which the number of simulations exceeds one, the function reports the median power, alongside the inter-quartile range (IQR), and the range. The median is reported since it is known that the distribution of power is skewed (Chapter 5). However, it is possible that the distribution of power could be represented in a graphical format, in which the distributions are plotted on a graph, in addition to the reporting of the median, IQR and range. This would then include all estimates of the power from the number of specified simulations and highlight the likelihood of obtaining a specific power value.

When estimating the power for an incomplete design with varying cluster size and known cluster sizes, the current function is limited to estimating the conditional power. However, some studies may know the size of the clusters prior to the study, and be using an incomplete design, and so require the marginal power instead. As such, any future versions of the function will allow for the marginal power to be estimated for an incomplete design with known clusters of varying size.

6.6 Conclusions

This chapter establishes a Stata function for estimating power when clusters vary in size – for scenarios with known and unknown cluster size. This should lead to an improvement in the methodological quality of sample size in calculations reported in SW-CRTs, and allow for the inclusion of varying cluster size (if appropriate). This chapter has built on work presented in previous chapters, and adds an important tool to the SW-CRT literature. The next chapter pools together the key results from each of the preceding chapters, and highlights the implication on future research.

CHAPTER 7: DISCUSSION

7.1 Thesis summary

Stepped-wedge cluster randomised (SW-CRTs) are an increasingly used alternative to parallel cluster randomised trials (P-CRTs). However, there is a sparsity of research into the SW-CRT in comparison to the P-CRT and other design types. As such, the methodological literature is at a less advanced stage for a SW-CRT than a P-CRT. The work in this thesis aims to address some of the current methodological gaps by expanding the knowledge base of SW-CRTs and provide a platform for future work.

The overall aim of this thesis was to develop the understanding of several key design features found in a SW-CRT. Chapter 3 reviews the reporting quality and methodological quality of published SW-CRTs, which led to two key issues in the design stage being apparent – few studies allowed for the decay in correlation between observations within a cluster over time, and few allowed for varying cluster size. Chapter 4 presents a correlation structure for SW-CRTs that allows for the correlation between observations to be dependent on the timing of the observations and provides a resource of estimates of the inter-period correlation and the within-period correlation. The last two chapters (5 and 6) focus on varying cluster size in SW-CRTs, developing methods to estimate the power in a SW-CRT with varying cluster size and investigating how a SW-CRT with varying cluster size compares to a SW-CRT with equal cluster size. A short summary of the chapters is given below.

7.2 Chapter overview

Chapter 2 presented a review of the methodological literature for SW-CRTs – highlighting the sparsity compared to the literature for P-CRTs. There are a number of methodological issues that were yet to be addressed, with little research on varying cluster size in SW-CRTs , and only recent research on alternative correlation structures in SW-CRTs.

A methodological review of the reporting quality of published SW-CRTs was given in Chapter 3. The adherence to the CONSORT statement for randomised trials and the extension for cluster trials were identified for each study. Additionally, the methodological rigor of the sample size calculation was assessed and the analytical method used in full trial reports was evaluated.

In Chapter 4, we present a method to allow for the correlation between observations within a cluster to be dependent on the timing of the observations. We introduce the notion of the intra-cluster correlation (ICC) for studies in which the correlation is time-independent, and the inter-period correlation (IPC), and the within-period correlation (WPC) for studies in which the correlation is time-dependent. In addition, and as an illustration, we provide estimates of the ICC, IPC, and WPC for outcomes associated with type-2 diabetes using anonymised patient data from The Health Improvement Network. For dichotomous outcomes, we make a careful distinction between the latent ICC and the natural ICC, and clarify why the natural ICC should be used for sample size calculations.

The impact of varying cluster size in a SW-CRT was explored in Chapter 5, by comparing the precision of a SW-CRT with unequal cluster size to the precision of a SW-CRT with equal cluster size using a measure of a relative efficiency (RE). A simulation study was presented to show how the cluster mean correlation, the number of clusters, the number of steps, and the degree of variation in cluster sizes affect the RE. The RE of a P-CRT with varying cluster size compared to a P-CRT with equal cluster size was also reported – as a comparison. We explored both between-cluster variation in size and within-cluster variation in size.

Chapter 6 implemented the methodology developed in Chapter 5 to allow researchers and trialists to estimate power in a SW-CRT with varying cluster size. To this end, we presented a Stata function that could allow the methodology to be used in practice.

7.3 Areas of contribution to the field of SW-CRTs

This thesis has contributed to the methodological literature for SW-CRTs by addressing each of the following research aims:

1. Review the sample size calculations for SW-CRTs to assess whether they are sufficiently reported, and whether appropriate methodology has been used.
2. Evaluate the validity of the assumption that the correlation between observations within a cluster is independent of time.
3. Demonstrate the impact of varying cluster size in a SW-CRT, and make comparisons to the impact of varying cluster size in a P-CRT.
4. Propose a method for estimating power in a SW-CRT when clusters vary in size.

The key results emanating from this work in relation to these research aims are highlighted in the discussion section within each chapter. A summary of the main results are given below.

7.3.1 Methodological review of sample size calculations

Early systematic reviews had suggested that the justification of a sample size is often under reported for published SW-CRTs (28, 29). However, in those that justify a sample size, it was not known whether the details of the sample size calculation are well reported, or whether appropriate methodology is used. To address this research question, a methodological review of protocols and full trial reports was conducted to assess the quality of reporting by checking the adherence to the CONSORT recommendations for randomised trials and the extension for cluster randomised trials. A list of SW-CRT specific items was also extracted, alongside the methodology of the sample size calculations. For full trial reports, an evaluation of the analytical methods used was conducted.

It was identified that the quality of reporting in SW-CRTs is sub-optimal, and the methodology being used is often inadequate and inappropriate. Of the 9 items recommended by the CONSORT cluster extension, the median number of items reported was 5 [IQR: 2 to 6], with no trials reporting all 9 items – though there was some evidence of an improvement in reporting post 2012. Generally, the number of steps and the number of clusters randomised per step were well reported. However, the design type (cross-sectional or cohort) was often not explicitly reported, and there was often ambiguity about whether the cluster size referred to the total cluster size or the cluster-period size. In a SW-CRT, the lack of clarity of the design type is especially important – since the methods used to estimate the sample size and for post-trial analysis hinge on whether a cross-sectional or cohort design was used. Cohort designed SW-CRTs require additional random effects to indicate repeated measurements from participants over time – not including this random effect may lead to an over-precise confidence interval for a treatment effect.

Often, sample size calculations we identified not to be using appropriate the methodology, with over a quarter not reporting any adjustment for clustering in the sample size calculation. This is notable since there are published guidelines for CRTs that explicitly state that an adjustment for clustering in the sample size calculation should be reported. Additionally, almost two-thirds of the included studies did not report an adjustment for time effects in the sample size calculation. Since the treatment effect can be confounded with time, ignoring time in the sample size calculation may lead to an underpowered study. The lack of reporting of clustering and time in the sample size calculation may be a result of incorrect methods being used. Typically, a design effect appropriate for a P-CRT was reported, even though it ignores the effect of time. Nevertheless, almost a third of studies reported using the recommended Hussey and Hughes framework. However, this framework is appropriate for cross-sectional designs, and many of the studies were cohort designs. As such, without adaptations, the methods may still not be appropriate for the sample size calculation. There was some evidence of studies including a transition period in the design between the control and intervention conditions. Though, there was little reporting of this being addressed in the sample size calculation – which may have led to underpowered studies.

A systematic review of P-CRTs found two-thirds of full trials contains unequal sized clusters (146). This review found that only a small number of SW-CRTs had reported varying cluster size in the context of the sample size calculation or in their trial description. However, of the few SW-CRTs that acknowledged that the clusters may vary in size, they often did not report an adjustment for this in the sample size calculation, or simply did not report how or whether varying cluster size was accounted for.

The reporting of an extended correlation structure was extremely poor – though this may be expected since the research into extended correlation structures in a SW-CRT has only recently being in the forefront of the methodological literature. Indeed, only two studies reported extended correlation measures, and these related to a correlation coefficient of repeated measures and a within-patient correlation over time. Neither of these used an appropriate sample size method.

The post-trial analysis of studies was generally mixed. Whilst the majority of studies reported an inclusion for clustering in the analysis, only half reported an acknowledgment of the effect of time. A failure to include time in the analysis is likely to produce an over-precise treatment effect estimate. Additionally, although most of the studies were cohort or open cohort designs, repeated measures were often not reported or included in the analysis. This would also lead to an over-precise treatment effect estimate as the analysis would treat each observation as a new participant and ignore the within-participant correlation over time.

This work echoes the call for reporting guidelines for SW-CRTs. Currently, a CONSORT extension is being developed for SW-CRTs which may help to address some of these concerns (71).

7.3.2 Appropriate correlation structure in a SW-CRT

An estimate of the correlation between observations within a cluster (ρ) is necessary for pre-trial sample size calculations. In the absence of period effects, the ICC denotes the correlation between observations within a cluster (i.e. $ICC = \rho$). However, in longitudinal CRTs, there is a need to acknowledge the possibility of a decay in correlation over time within a cluster. This led to the concept of the WCC, IPC, and WPC. The WPC is defined as the correlation between any two observations from within the same cluster and the same time-period. The IPC is defined as the correlation between any two observations from within the same cluster during but from different time periods. The WCC is the correlation between any two random observations within a cluster, and is a function of the IPC and WPC. The WCC and the ICC are both equal to ρ, but are only defined in time-dependent models and time-independent models, respectively. If the IPC=WPC – which is often assumed (32) – then there are no period effects present, and the ICC can be used to replace ρ in a design effect for a P-CRT. However, if the IPC ≠ WPC, then a period effect is present, and the WCC should be used to replace ρ in a design effect for a P-CRT. Recently, research has suggested that the WPC and IPC should be included in the sample size calculation for a SW-CRT (27, 48). Recently, a DE was published for SW-CRTs that acknowledged the IPC and WPC (26) (equation 2.10). However, there is a dearth of likely values of the IPC and WPC to inform methodological research and sample size calculations. As such, there have been calls for estimates of the IPC and WPC to be published (26, 27). Previously, the same value had been used to represent the ratio of the IPC to the WPC (27, 48), but with little evidence to justify the value.

In Chapter 4, as an illustration, estimates of the ICC, IPC, and WPC were presented for typical outcomes associated with type-2 diabetes using anonymised patient data from The Health Improvement Network. In all of the results presented, the WPC and IPC differ – leading to the conclusion that period effects were present. Within this, it was shown how increasing the study length (with a fixed time-period length) or increasing the time-period length (with a fixed study length) may impact the IPC and WPC. When presenting the impact of study length and time-period length on the IPC and WPC, increasing the period length (with a fixed study length) tended to decrease both the IPC and WPC. This is perhaps expected as increasing the potential length of time between two observations is likely to lead to a decrease in the correlation between them. Increasing the study length (with a fixed time-period length), tended to increase the WPC but decrease the IPC. Whilst it is not surprising that the IPC decreases (since there is a longer duration between possible observations), it was unexpected that the WPC increased, since the time-periods length remained fixed.

It is recommended that in similar settings, the ICC should not be used as an estimate for ρ in the design effect for a P-CRT. Instead the WCC should be calculated – as a function of the IPC and WPC in a setting with an equal number of observations in each time-period – and used to replace ρ in the design effect. The presence of a period effect emphasises that a recent DE for SW-CRT that includes the IPC and WPC should be used for cross-sectional SW-CRTs.

Recently, the IPC and WPC have been recognised as important to SW-CRTs and other longitudinal designs. However, there has been little evidence of likely values of the IPC and WPC, which has led to ad-hoc values being used in the methodological literature. Chapter 4 is the first resource of possible IPC and WPC values, which may be useful for pre-trial sample size calculations and for methodological research.

In scenarios in which the WPC and IPC differ, the current assumption of correlation between observations being independent of the time in which the observations are made is invalid. Instead, a time-dependent model – that is, a model with a cluster-by-time random effect – should be fitted to the data in the analysis stage, and in the estimation of the correlation structure. This allows the WPC and the IPC to depict the correlation between participants over multiple time-periods. Furthermore, the WCC – estimated as a function of the IPC and WPC (equation 4.5) when the number of observations is equal in each time-period – should be used in the sample size calculation for longitudinal CRTs.

7.3.3 Explored the impact of varying cluster size in a SW-CRT

Typically, it is assumed that a CRT will contain clusters of equal size. However, in practice this is unlikely to be true (57, 146). In the context of P-CRTs, there has been a variety of research into the effect of varying cluster size on the precision and power – and it is known that cluster size variation decreases the power compared to a P-CRT with equal cluster size. In published SW-CRTs, there is little evidence of the reporting of an allowance for varying cluster size in the power calculation.

In Chapter 5, a set of simulation studies were presented to examine how the precision in a CRT with unequal cluster sizes compares to the precision of a CRT with equal cluster size – presented as a relative efficiency. Results were shown for: a SW-CRT with between-cluster variation in size; a P-CRT with between-cluster variation in size; and a SW-CRT with between-cluster and within-cluster variation in size.

For a SW-CRT with between-cluster variability in cluster sizes, the average amount of precision lost compared to a SW-CRT with equal cluster sizes may not be large in magnitude. However, when considering all possible values of the precision, there is a large degree of variability. That is, on average, a SW-CRT with varying cluster size may only have slightly less precision (and hence power) than a SW-CRT with equal cluster size; nevertheless, the large amount of variability in the potential precision could lead to a potential 20% increase or 80% decrease in precision, compared to a SW-CRT with equal cluster sizes. As such, the precision for a SW-CRT with varying cluster size is conditional on the randomisation order of the clusters. In a SW-CRT with varying cluster size, it is possible for the precision (and power) to be greater than a SW-CRT with equal cluster size. As the number of clusters in a SW-CRT increase, the design is less influenced by cluster size variation – and there is a decrease in the variability in the distribution of the RE.

In the context of P-CRTs, it is often assumed that there will be a balance in the number of observations in the intervention and control arms (60, 62). When this is true, the precision is a fixed value, and so the ratio of the precision in an unequal cluster size design to the precision in an equal cluster size design will always be a single fixed value. However, in practice, CRTs often contain a small number of clusters, and so there is inevitably an imbalance in terms of the number of observations in the intervention and control arms. In Chapter 4, we allow the number of observations to vary between arms, and show that the RE for a P-CRT is a distribution of RE values, and may contain a large degree of variation. In a P-CRT with varying cluster size, the precision (and power) can never be greater than a P-CRT with equal cluster size – and so an unequal cluster size design is always less efficient than an equal cluster size design. The average RE and the distribution of RE of a P-CRT with unequal cluster size is greatly affected by changes to the cluster mean correlation – which is particularly evidenced when the variation in cluster sizes is not small. That is, for a given trial, the power may be a lot less than the average power.

Since the SW-CRT is a longitudinal CRT and is split into time-periods, individual clusters can differ in size over time. Chapter 5 also presented results of a simulation study that considered both between-cluster and within-cluster variation in size. The inclusion of within-cluster variability leads to a greater variability in the RE. That is, for a given trial, the power may be a lot different to the average or expected power. Again, regardless of the number of steps, it is possible for a SW-CRT with unequal cluster size (with between-cluster and within-cluster variation) to offer more precision, and power, than a SW-CRT with equal cluster size.

The key result emanating from this work in relation to the research question is:

On average, a SW-CRT with varying cluster size will have less power than a SW-CRT with equal cluster size. However, the potential power is hugely variable, and so a given trial could potentially have a lot less power or more power than the expected power. Although not appreciated, a P-CRT suffers from the same phenomenon – though to a lesser degree. A SW-CRT is, on average, affected less by varying cluster size than a P-CRT. However, the variability in power in a SW-CRT with varying cluster size is much greater than in a P-CRT.

7.3.4 Estimating power in a SW-CRT with varying cluster size

The assessment of power is important when designing a study. For SW-CRTs, there exists a framework for estimation of power under the assumption of equal cluster sizes (45) which has been presented as a Stata function (63) to encourage future trialists to use the correct methodology. However, Chapter 5 established that the SW-CRT is highly affected by varying cluster size. As such, a method is necessary to estimate the power in a SW-CRT with varying cluster size.

In Chapter 5, we proposed a method to estimate the precision in a SW-CRT with varying cluster size, if the cluster sizes are not known. We then implement this method in Chapter 6. Additionally, in Chapter 6, we extend this method to consider the power of a SW-CRT with unequal cluster sizes, if the cluster sizes are not known, but estimates of the average cluster size and the between-cluster coefficient of variation (and within-cluster coefficient of variation – if applicable) are known. This is applicable for SW-CRTs with between-cluster variation in cluster sizes or between-cluster and within-cluster variation in cluster sizes. It is then shown how the power can be estimated in a SW-CRT with unequal cluster sizes, if the cluster sizes are known. When the cluster sizes are known, the power is affected by the randomisation order of the clusters. As such, we introduce the concept of the marginal power and the conditional power. The conditional power is the power for a SW-CRT conditional on the randomisation order – and is useful post-randomisation. The marginal power is the average power calculated as the median of all possible conditional powers.

In Chapter 6, we implement a method to estimate the power in a SW-CRT with unequal cluster sizes as a Stata function, with examples presented to show its use. The examples highlight that variability in power that can be expected when cluster sizes vary. The ***swpower*** function will ensure that the design of future SW-CRTs uses appropriate methods and, if appropriate, acknowledges varying cluster size in their sample size calculation.

We found that it is possible for some randomisation orders to have a relative efficiency greater than one, indicating that unequal cluster sizes could offer more precision, and power than a design with equal cluster sizes. Nevertheless, it is unlikely that an investigator could capitalise on unequal cluster sizes to improve power without undermining the randomisation process. Prior to randomising, all clusters should have an equal opportunity to be randomised to a given sequence – i.e. they have an equal chance to be initiated to the intervention condition at each time-period. Since a relative efficiency greater than one applies to only specific realisations, trying to pick one of these realisations would invalidate the randomisation process.

7.4 Limitations of this work

The limitations of each research question are reported in the discussion section of the appropriate chapter. Here, we report only an overall limitation of this thesis. All of the methodological work in this thesis stems from the Hussey and Hughes framework which is appropriate for continuous outcomes in a cross-sectional framework (32). However, Chapter 3 showed that many SW-CRTs are cohort designs and many have dichotomous outcomes. The use of a cohort design would require an additional random effect for participant to allow for a within-participant correlation over time – to allow for repeated measurement from the same participant over time. Whilst we extended the Hussey and Hughes framework to allow for a cluster-by-period interaction in Chapter 4, we did not consider the additional participant level random effect – as this would have added a further level of complexity.

7.5 Implications of research

Previous systematic reviews had identified that published SW-CRTs did often not justify the sample size included in the study. However, of those studies who had reported a sample size calculation, it was not known how well reported these calculations were with respect to published guidelines. Whilst CONSORT statements for RCTS and CRTs exist, it was unknown how much, or how little, previous SW-CRTs had adhered to the recommendations. Part of this thesis has shown that there is a severe under reporting of many sample size items in SW-CRT protocols and trial reports, and there is an urgent need for a published CONSORT extension for SW-CRTs – though it is known that one is currently in development. Noticeably, there seemed to be a slight improvement in reporting since the CONSORT extension for CRTs, so a CONSORT extension for SW-CRTs will hopefully lead to a vast improvement in reporting quality.

In P-CRTs and SW-CRTs, the ICC is used almost exclusively in pre-trial sample size calculations and to describe the correlation between observations within a cluster. However, the longitudinal nature of the SW-CRT means that this may not be sufficient to describe the correlation structure. Recent papers have introduced the concept of the correlation between observations being time-dependent in a SW-CRT through the IPC and WPC – though there is little evidence of likely values of these. If the IPC and WPC are equal, then there are no period effects present, and the ICC is sufficient to describe the correlation between observations within a cluster. However, if the IPC and WPC differ, then they, along with the WCC, should be used to describe the correlation between observations within a cluster. This thesis is the first known reporting of estimates of the IPC and WPC. This will allow future SW-CRTs to use these estimates in their power calculation, and acknowledge the decay of correlation over time, by using the DE published by Hooper et al. (equation 2.10) (26). As well as being used in future sample size calculations; the results establish that a time-dependent model is necessary for SW-CRTs to depict the correlation structure. The publication of possible values also enables other methodological researchers to consider the impact of the decay of correlation on the sample size of future SW-CRTs.

This thesis has established that the SW-CRT is affected less, on average, than the P-CRT when clusters vary in size. It is known that a large proportion of P-CRTs had varying cluster size (146), and that the P-CRT is heavily influenced by the degree of variation in cluster sizes (57). However, no current systematic review has indicated the proportion of published full-trial SW-CRTs that displayed varying cluster size. Currently, there may be a lack of acknowledgement that the clusters in a trial are likely to vary in size in pre-trial sample size calculations, since there has been no established methodology for estimating power in a SW-CRT with varying cluster size. However, the Stata function presented in this thesis may allow future sample size calculations to acknowledge unequal cluster sizes. Nevertheless, it is important that any sample size calculation for a SW-CRT or P-CRT that includes varying cluster size should report the expected power and a measure of the likely variation in the power. The results of this thesis have already seen a change in approach for some CRTs. The acute coronary syndrome quality improvement in Kerala (ACS QUIK) study is a SW-CRT that contains clusters of varying size. The results of this chapter showed that the ACS QUIK study would not be sufficiently powered if a P-CRT was used, but that since the SW-CRT was affected less than the P-CRT, it would still be sufficiently powered, despite a large variation in cluster sizes.

By developing a Stata function, a practical element to the work has been shown, so that we do not present merely theoretical work. The Stata function can allow the power to be estimated in a SW-CRT with varying cluster size, and will encourage trialists to use the correct methodology to estimate the power in a SW-CRT.

7.6 Future research

There are several areas of future research that has been identified from each of the chapters, which are discussed below.

The methodological review in Chapter 3 could be extended to assess whether there has been any improvement to the reporting standards of published SW-CRTs since the review was conducted (October 2014). In addition to the publication of the findings of the current review, the future publication of a CONSORT extension for SW-CRTs is likely to impact the quality of reporting, and so a future methodological review should be conducted to assess what impact is made on reporting standards.

The methodical review emphasised the need for guidelines on why a sample size calculation is important in a SW-CRT. This should include a lay term guide to how a sample size calculation should be conducted and reported. This could include details of the appropriate methodology, and how this methodology can be implemented in different statistical software packages. By highlighting how a sample size calculation is conducted and reported, it may increase the reporting quality and methodological rigor of future SW-CRT sample size calculations.

Chapter 4 highlighted that if period effects are present, the ICC may not be appropriate as an estimate of ρ for use in a design effect to estimate the sample size – since the ICC assumes that the correlation between observations is time-independent. Instead, the WCC should be used in a sample size calculation, since it is a time-dependent measure of correlation. However, the varying number of observations in each time-period meant that we could not provide an accurate estimate of the WCC from THIN data to compare to the ICC. Further research is required to provide estimates of the WCC for use in sample size calculations. IPCs and WPCs for outcomes associated with type-2 diabetes were presented in Chapter 4. Whilst useful as a platform for showing that correlation decays over time, IPCs and WPCs are necessary for other potential study populations, so that future trials can use correct values in a sample size calculation. Additionally, whilst a P-CRT is often treated as a single study-period, it may be that recruitment and observations can also be split into time-periods. As such, it may be that the ICC is not sufficient to describe the correlation in a P-CRT, and the IPC and WPC should also be used in a P-CRT.

Chapter 4 showed potential values of the IPC and WPC. Previously, there have been no published IPC or WPC estimates, and so it has often been assumed that they are equal. Since the results of Chapter 4 showed that the IPC and WPC differ – i.e. there are period effects present – recent literature surrounding the decay of correlation over time has become increasingly important. Previously, a single value had been used to represent the ratio of the IPC and WPC, which was much greater than found here. Future research is needed to increase the resource of IPC and WPC values.

Chapters 5 and 6 showed that it is possible for SW-CRTs with varying cluster size to have more precision (i.e. more power) than a SW-CRT with equal cluster sizes. Conversely, it is also possible to have SW-CRTs with varying cluster size have less power than an equal cluster size design, and less power than expected with unequal cluster sizes. That is, there may be a large variation in the possible power of a SW-CRT if the clusters are of unequal size. To minimise the probability of a SW-CRT having less power than expected, it may be that stratification is necessary in the randomisation process. Future research should investigate whether stratification by cluster sizes can guard against large losses in power in a SW-CRT. A second strand of future research should include the investigation of optimal SW-CRTs when clusters vary in size.

7.7 Conclusions

The literature for SW-CRTs is constantly evolving, and is a challenging research area that still requires more methodological research to improve the existing knowledge base. The aim of this thesis was to develop the understanding of some of the design issues that are faced when carrying out a SW-CRT. Though some of the issues remain, and many further questions have arose during this research, this thesis has contributed towards improving the methodological literature for SW-CRTs and improving the quality and reporting of sample size calculations for SW-CRTs. Without a rigorous sample size calculation, a SW-CRT is flawed from the outset, so improving the quality of a sample size calculation, through theoretical development and practical application, will hopefully see a development in the methodological rigor and quality of reporting of future sample size calculations.

CHAPTER 8: References

1. Campbell MJ. Challenges of cluster randomized trials. Journal of comparative effectiveness research. 2014;3(3):271-81.

2. Adab P, Pallan MJ, Lancashire ER, Hemming K, Frew E, Griffin T, et al. A cluster-randomised controlled trial to assess the effectiveness and cost-effectiveness of a childhood obesity prevention programme delivered through schools, targeting 6-7 year old children: the WAVES study protocol. BMC public health. 2015;15:488.

3. Chinbuah MA, Kager PA, Abbey M, Gyapong M, Awini E, Nonvignon J, et al. Impact of community management of fever (using antimalarials with or without antibiotics) on childhood mortality: A cluster-randomized controlled trial in Ghana. American Journal of Tropical Medicine and Hygiene. 2012;87(SUPPL.5):11-20.

4. Mills EJ, Chan AW, Wu P, Vail A, Guyatt GH, Altman DG. Design, analysis, and presentation of crossover trials. Trials. 2009;10:27.

5. Bashour HN, Kanaan M, Kharouf MH, Abdulsalam AA, Tabbaa MA, Cheikha SA. The effect of training doctors in communication skills on women's satisfaction with doctor-woman relationship during labour and delivery: A stepped wedge cluster randomised trial in Damascus. BMJ Open. 2013;3(8).

6. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ (Clinical research ed). 2015;350:h391.

7. Campbell DT, Cook TD. Quasi-experimentation: Design and analysis for field settings. Rand McNally College Publishing Company, Chicago. 1979.

8. The Gambia Hepatitis Intervention Study. The Gambia Hepatitis Study Group. Cancer Research. 1987;47(21):5782-7.

9. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. Statistics in medicine. 2015;34(2):181-96.

10. Mdege ND, Man MS, Taylor nee Brown CA, Torgerson DJ. There are some circumstances where the stepped-wedge cluster randomized trial is preferable to the alternative: no randomized trial at all. Response to the commentary by Kotz and colleagues. Journal of clinical epidemiology. 2012;65(12):1253-4.

11. Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W. Use of the stepped wedge design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. Journal of clinical epidemiology. 2012;65(12):1249-52.

12. Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W. Researchers should convince policy makers to perform a classic cluster randomized controlled trial instead of a stepped wedge design when an intervention is rolled out. Journal of clinical epidemiology. 2012;65(12):1255-6.

13. Mdege ND, Kanaan M. Response to Keriel-Gascou et al. Addressing assumptions on the stepped wedge randomized trial design. Journal of clinical epidemiology. 2014;67(7):833-4.

14. Lohaugen GCC, Beneventi H, Andersen GL, Sundberg C, Ostgard HF, Bakkan E, et al. Do children with cerebral palsy benefit from computerized working memory training? Study protocol for a randomized controlled trial. Trials. 2014;15(1).

15. Samuel-Hodge CD, Garcia BA, Johnston LF, Kraschnewski JL, Gustafson AA, Norwood AF, et al. Rationale, design, and sample characteristics of a practical randomized trial to assess a weight loss intervention for low-income women: The Weight-Wise II Program. Contemporary Clinical Trials. 2012;33(1):93-103.

16. Ratanawongsa N, Handley MA, Quan J, Sarkar U, Pfeifer K, Soria C, et al. Quasi-experimental trial of diabetes Self-Management Automated and Real-Time Telephonic Support (SMARTSteps) in a Medicaid managed care plan: study protocol. BMC health services research. 2012;12:22.
17. Grunewaldt KH, Lohaugen GCC, Austeng D, Brubakk AM, Skranes J. Working memory training improves cognitive function in VLBW preschoolers. Pediatrics. 2013;131(3):e747-e54.
18. Haines T, O'Brien L, McDermott F, Markham D, Mitchell D, Watterson D, et al. A novel research design can aid disinvestment from existing health technologies with uncertain effectiveness, cost-effectiveness, and/or safety. Journal of clinical epidemiology. 2014;67(2):144-51.
19. Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials. 2015;16:352.
20. Hooper R, Bourke L. The dog-leg: an alternative to a cross-over design for pragmatic clinical trials in relatively stable populations. International journal of epidemiology. 2014;43(3):930-6.
21. Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. Journal of clinical epidemiology. 2013;66(7):752-8.
22. Keriel-Gascou M, Buchet-Poyau K, Rabilloud M, Duclos A, Colin C. A stepped wedge cluster randomized trial is preferable for assessing complex health interventions. Journal of clinical epidemiology. 2014;67(7):831-3.
23. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC medical research methodology. 2006;6:54.
24. Priestley G, Watson W, Rashidian A, Mozley C, Russell D, Wilson J, et al. Introducing Critical Care Outreach: A ward-randomised trial of phased introduction in a general hospital. Intensive Care Medicine. 2004;30(7):1398-404.
25. Hill AM, Waldron N, Etherton-Beer C, McPhail SM, Ingram K, Flicker L, et al. A stepped-wedge cluster randomised controlled trial for evaluating rates of falls among inpatients in aged care rehabilitation units receiving tailored multimedia education in addition to usual care: A trial protocol. BMJ Open. 2014;4(1).
26. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Statistics in medicine. 2016;35(26):4718-28.
27. Taljaard M, Teerenstra S, Ivers NM, Fergusson DA. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. Clinical trials (London, England). 2016;13(4):459-63.
28. Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, et al. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. Trials. 2015;16:353.
29. Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. Journal of clinical epidemiology. 2011;64(9):936-48.
30. Grayling MJ, Wason JM, Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. Trials. 2017;18(1):33.
31. Barker D, McElduff P, D'Este C, Campbell MJ. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. BMC medical research methodology. 2016;16:69.
32. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemporary clinical trials. 2007;28(2):182-91.
33. Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W. The stepped wedge design does not inherently have more power than a cluster randomized controlled trial. Journal of clinical epidemiology. 2013;66(9):1059-60.
34. Viechtbauer W, Kotz D, Spigt M, Arts IC, Crutzen R. Response to Keriel-Gascou et al.: higher efficiency and other alleged advantages are not inherent to the stepped wedge design. Journal of clinical epidemiology. 2014;67(7):834-6.

35. Hemming K, Girling A, Martin J, Bond SJ. Stepped wedge cluster randomized trials are efficient and provide a method of evaluation without which some interventions would not be evaluated. Journal of clinical epidemiology. 2013;66(9):1058-9.
36. Prost A, Binik A, Abubakar I, Roy A, De Allegri M, Mouchoux C, et al. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. Trials. 2015;16:351.
37. Hargreaves JR, Copas AJ, Beard E, Osrin D, Lewis JJ, Davey C, et al. Five questions to consider before conducting a stepped wedge trial. Trials. 2015;16:350.
38. Zhan Z, van den Heuvel ER, Doornbos PM, Burger H, Verberne CJ, Wiggers T, et al. Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. Journal of clinical epidemiology. 2014;67(4):454-61.
39. de Hoop E, van der Tweel I, van der Graaf R, Moons KG, van Delden JJ, Reitsma JB, et al. The need to balance merits and limitations from different disciplines when considering the stepped wedge cluster randomized trial design. BMC medical research methodology. 2015;15:93.
40. Van den Heuvel ER, Zwanenburg RJ, Van Ravenswaaij-Arts CM. A stepped wedge design for testing an effect of intranasal insulin on cognitive development of children with Phelan-McDermid syndrome: A comparison of different designs. Statistical methods in medical research. 2014.
41. Heo M, Kim N, Rinke ML, Wylie-Rosett J. Sample size determinations for stepped-wedge clinical trials from a three-level data hierarchy perspective. Statistical methods in medical research. 2016.
42. Palmer VJ, Chondros P, Piper D, Callander R, Weavell W, Godbee K, et al. The CORE study protocol: a stepped wedge cluster randomised controlled trial to test a co-design technique to optimise psychosocial recovery outcomes for people affected by mental illness in the community mental health setting. BMJ open. 2015;5(3):e006688.
43. Freedman B. Equipoise and the ethics of clinical research. N Engl J Med. 1987;317(3):141-5.
44. Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ. Sample size calculation for a stepped wedge trial. Trials. 2015;16:354.
45. Hemming K, Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. Journal of clinical epidemiology. 2016;69:137-46.
46. Rhoda DA, Murray DM, Andridge RR, Pennell ML, Hade EM. Studies with staggered starts: multiple baseline designs and group-randomized trials. American journal of public health. 2011;101(11):2164-9.
47. de Hoop E, Woertman W, Teerenstra S. The stepped wedge cluster randomized trial always requires fewer clusters but not always fewer measurements, that is, participants than a parallel cluster randomized trial in a cross-sectional design. In reply. Journal of clinical epidemiology. 2013;66(12):1428.
48. Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. BMJ (Clinical research ed). 2015;350:h2925.
49. van der Tweel I, van der Graaf R. Issues in the use of stepped wedge cluster and alternative designs in the case of pandemics. The American journal of bioethics : AJOB. 2013;13(9):23-4.
50. Twisk JW, Hoogendijk EO, Zwijsen SA, de Boer MR. Different methods to analyze stepped wedge trial designs revealed different aspects of intervention effects. Journal of clinical epidemiology. 2016;72:75-83.
51. Heo M, Kim Y, Xue X, Kim MY. Sample size requirement to detect an intervention effect at the end of follow-up in a longitudinal cluster randomized trial. Statistics in medicine. 2010;29(3):382-90.
52. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. International journal of epidemiology. 2015;44(3):1051-67.
53. Hemming K, Girling AJ, Sitch AJ, Marsh J, Lilford RJ. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. BMC MedResMethodol. 2011;11:102.

54. Killip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. AnnFamMed. 2004;2(3):204-8.
55. Gao F, Earnest A, Matchar DB, Campbell MJ, Machin D. Sample size calculations for the design of cluster randomized trials: A summary of methodology. Contemporary clinical trials. 2015;42:41-50.
56. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. FamPract. 2000;17(2):192-6.
57. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. International journal of epidemiology. 2006;35(5):1292-300.
58. Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Statistics in medicine. 2001;20(3):377-90.
59. Pan W. Sample size and power calculations with correlated binary data. Control Clin Trials. 2001;22(3):211-27.
60. van Breukelen GJ, Candel MJ. Comments on 'Efficiency loss because of varying cluster size in cluster randomized trials is smaller than literature suggests'. Statistics in medicine. 2012;31(4):397-400.
61. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Statistics in medicine. 2016;35(13):2149-66.
62. van Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. Statistics in medicine. 2007;26(13):2589-603.
63. Hemming KG, A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. Stata Journal. 2014;14(2):363-80(18).
64. de Hoop E, Moerbeek M, Gerritsen D, Teerenstra S. Sample size estimation for cohort and cross-sectional cluster randomized stepped wedge designs In: Oomen-de Hoop E, Efficient designs for cluster randomized trials with small numbers of clusters: stepped wedge and other repeated measurements designs (doctoral thesis) Available from: http://repositoryubnrunl/bitstream/handle/2066/134179/134179pdf?sequence=1.
65. Hemming K. Sample size calculations for stepped wedge trials using design effects are only approximate in some circumstances. Trials. 2016;17(1):234.
66. Hemming K, Girling A. The efficiency of stepped wedge vs. cluster randomized trials: stepped wedge studies do not always require a smaller sample size. Journal of clinical epidemiology. 2013;66(12):1427-8.
67. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. Jama. 1996;276(8):637-9.
68. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Annals of internal medicine. 2010;152(11):726-32.
69. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Journal of the American Podiatric Medical Association. 2001;91(8):437-42.
70. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ (Clinical research ed). 2012;345:e5661.
71. Hemming K, Girling A, Haines T, Lilford R. Protocol: Consort extension to stepped wedge cluster randomised controlled trial. 2014.
72. Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. BMC medical research methodology. 2004;4:21.
73. Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. Bmj. 2011;343:d5886.

74. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. Journal of clinical epidemiology. 2015;68(6):716-23.
75. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. Statistics in medicine. 2007;26(1):2-19.
76. van Breukelen GJ, Candel MJ. Efficiency loss due to varying cluster sizes in cluster randomized trials and how to compensate for it: comment on You et al. (2011). Clinical trials (London, England). 2012;9(1):125; author reply 6-7.
77. van Holland BJ, de Boer MR, Brouwer S, Soer R, Reneman MF. Sustained employability of workers in a production environment: design of a stepped wedge trial to evaluate effectiveness and cost-benefit of the POSE program. BMC Public Health. 2012;12:1003.
78. Ni Mhurchu C, Gorton D, Turley M, Jiang Y, Michie J, Maddison R, et al. Effects of a free school breakfast programme on children's attendance, academic achievement, and short-term hunger: A stepped-wedge, cluster randomised controlled trial. Australasian Medical Journal. 2011;4 (12):805.
79. Bailet LL, Repper KK, Piasta SB, Murphy SP. Emergent literacy intervention for prekindergarteners at risk for reading failure. Journal of Learning Disabilities. 2009;42(4):336-55.
80. Muntinga ME, Hoogendijk EO, van Leeuwen KM, van Hout HP, Twisk JW, van der Horst HE, et al. Implementing the chronic care model for frail older adults in the Netherlands: study protocol of ACT (frail older adults: care in transition). BMC geriatrics. 2012;12:19.
81. Turner J, Kelly B, Clarke D, Yates P, Aranda S, Jolley D, et al. A randomised trial of a psychosocial intervention for cancer patients integrated into routine care: the PROMPT study (promoting optimal outcomes in mood through tailored psychosocial therapies). BMC Cancer. 2011;11:48.
82. Gozalo P, Prakash S, Qato DM, Sloane PD, Mor V. Effect of the bathing without a battle training intervention on bathing-associated physical and verbal outcomes in nursing home residents with dementia: A randomized crossover diffusion study. Journal of the American Geriatrics Society. 2014;62(5):797-804.
83. Durovni B, Saraceni V, Moulton LH, Pacheco AG, Cavalcante SC, King BS, et al. Effect of improved tuberculosis screening and isoniazid preventive therapy on incidence of tuberculosis and death in patients with HIV in clinics in Rio de Janeiro, Brazil: A stepped wedge, cluster-randomised trial. The Lancet Infectious Diseases. 2013;13(10):852-8.
84. Marshall T, Caley M, Hemming K, Gill P, Gale N, Jolly K. Mixed methods evaluation of targeted case finding for cardiovascular disease prevention using a stepped wedged cluster RCT. BMC public health. 2012;12:908.
85. Leontjevas R, Gerritsen DL, Smalbrugge M, Teerenstra S, Vernooij-Dassen MJFJ, Koopmans RTCM. A structural multidisciplinary approach to depression management in nursing-home residents: A multicentre, stepped-wedge cluster-randomised trial. The Lancet. 2013;381(9885):2255-64.
86. Clark T, Berger U, Mansmann U. Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees: review. BMJ (Clinical research ed). 2013;346:f1135.
87. Turner RM, White IR, Croudace T. Analysis of cluster randomized cross-over trial data: a comparison of methods. Statistics in medicine. 2007;26(2):274-89.
88. Donner A, Kong AP. Design and Analysis of Cluster Randomization Trials in Health Research: Arnold Publishers Limited, London, U.K.; 2000 2000.
89. Campbell MJ. Cluster randomized trials in general (family) practice research. Statistical methods in medical research. 2000;9(2):81-94.
90. Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research: John Wiley & Sons; 2012.

91. Puffer S, Torgerson DJ, Watson J. Cluster randomized controlled trials. Journal of Evaluation in Clinical Practice. 2005;11(5):479-83.
92. Campbell MJ, Walters SJ. How to design, analyse and report cluster randomised trials in medicine and health related research: John Wiley & Sons; 2014.
93. Lancaster GA, Campbell MJ, Eldridge S, Farrin A, Marchant M, Muller S, et al. Trials in primary care: statistical issues in the design, conduct and evaluation of complex interventions. Statistical methods in medical research. 2010;19(4):349-77.
94. Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. Journal of clinical epidemiology. 1996;49(4):435-9.
95. Kerry SM, Bland JM. Trials which randomize practices II: sample size. Family practice. 1998;15(1):84-7.
96. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. International journal of epidemiology. 1999;28(2):319-26.
97. Hayes R, Moulton L. Cluster randomised trials. Boca Raton: Chapman & Hall/CRC; 2009.
98. Chuang JH, Hripcsak G, Heitjan DF. Design and analysis of controlled trials in naturally clustered environments: implications for medical informatics. Journal of the American Medical Informatics Association : JAMIA. 2002;9(3):230-8.
99. Bell ML, McKenzie JE. Designing psycho-oncology randomised trials and cluster randomised trials: variance components and intra-cluster correlation of commonly used psychosocial measures. Psycho-oncology. 2013;22(8):1738-47.
100. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. Statistics in medicine. 2008;27(27):5578-85.
101. Parienti JJ, Kuss O. Cluster-crossover design: a method for limiting clusters level effect in community-intervention studies. Contemporary clinical trials. 2007;28(3):316-23.
102. Eldridge SM, Ukoumunne OC, Carlin JB. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. International Statistical Review 77(3):378-394. 2009.
103. Webb DR, Khunti K, Gray LJ, Srinivasan BT, Farooqi A, Wareham N, et al. Intensive multifactorial intervention improves modelled coronary heart disease risk in screen-detected Type 2 diabetes mellitus: a cluster randomized controlled trial. DiabetMed. 2012;29(4):531-40.
104. Murray DM, Blistein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. EvalRev. 2003;27(1):79-103.
105. Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. Journal of clinical epidemiology. 2005;58(3):246-51.
106. Taljaard M, Donner A, Villar J, Wojdyla D, Velazco A, Bataglia V, et al. Intracluster correlation coefficients from the 2005 WHO Global Survey on Maternal and Perinatal Health: implications for implementation research. Paediatric and perinatal epidemiology. 2008;22(2):117-25.
107. Ukoumunne OC, Thompson SG. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. Statistics in medicine. 2001;20(3):417-33.
108. Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. ContempClinTrials. 2012;33(5):869-80.
109. Yelland LN, Salter AB, Ryan P, Laurence CO. Adjusted intraclass correlation coefficients for binary data: methods and estimates from a cluster-randomized trial in primary care. Clinical trials (London, England). 2011;8(1):48-58.
110. Thompson DM, Fernald DH, Mold JW. Intraclass correlation coefficients typical of cluster-randomized studies: estimates from the Robert Wood Johnson Prescription for Health projects. Annals of family medicine. 2012;10(3):235-40.
111. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. Statistics in medicine. 2001;20(3):453-72.

112. Teerenstra S, Eldridge S, Graff M, de Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. Statistics in medicine. 2012;31(20):2169-78.
113. Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. Lancet. 2011;378(9785):31-40.
114. Aziz Z, Absetz P, Oldroyd J, Pronk NP, Oldenburg B. A systematic review of real-world diabetes prevention programs: learnings from the last 15 years. Implementation science : IS. 2015;10:172.
115. Sturt JA, Whitlock S, Fox C, Hearnshaw H, Farmer AJ, Wakelin M, et al. Effects of the Diabetes Manual 1:1 structured education in primary care. Diabetic medicine : a journal of the British Diabetic Association. 2008;25(6):722-31.
116. Khunti K, Gray LJ, Skinner T, Carey ME, Realf K, Dallosso H, et al. Effectiveness of a diabetes education and self management programme (DESMOND) for people with newly diagnosed type 2 diabetes mellitus: three year follow-up of a cluster randomised controlled trial in primary care. BMJ. 2012;344:e2333.
117. Foster GD, Linder B, Baranowski T, Cooper DM, Goldberg L, Harrell JS, et al. A school-based intervention for diabetes risk reduction. NEnglJMed. 2010;363(5):443-53.
118. Shahbazian H, Latifi SM, Jalali MT, Shahbazian H, Amani R, Nikhoo A, et al. Metabolic syndrome and its correlated factors in an urban population in South West of Iran. JDiabetes Metab Disord. 2013;12(1):11.
119. Heisler M, Hofer TP, Schmittdiel JA, Selby JV, Klamerus ML, Bosworth HB, et al. Improving blood pressure control through a clinical pharmacist outreach program in patients with diabetes mellitus in 2 high-performing health systems: the adherence and intensification of medications cluster randomized, controlled pragmatic trial. Circulation. 2012;125(23):2863-72.
120. Echouffo-Tcheugui J, Simmons R, Williams K, Barling R, Prevost AT, Kinmonth A, et al. The ADDITION-Cambridge trial protocol: a cluster - randomised controlled trial of screening for type 2 diabetes and intensive treatment for screen-detected patients. BMC Public Health. 2009;9(1):136.
121. Narayan KM, Gregg EW, Fagot-Campagna A, Engelgau MM, Vinicor F. Diabetes--a common, growing, serious, costly, and potentially preventable public health problem. Diabetes ResClinPract. 2000;50 Suppl 2:S77-S84.
122. Research CM. Our Data 2012 [20/02/2015]. Available from: http://csdmruk.cegedim.com/our-data/our-data.shtml.
123. Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. Pharmacoepidemiol Drug Saf. 2009;18(1):76-83.
124. Mukherjee M, Wyatt JC, Simpson CR, Sheikh A. Usage of allergy codes in primary care electronic health records: a national evaluation in Scotland. Allergy. 2016.
125. Harkness EF, Grant L, O'Brien SJ, Chew-Graham CA, Thompson DG. Using read codes to identify patients with irritable bowel syndrome in general practice: a database study. BMC family practice. 2013;14:183.
126. Diabetes UK - Facts and Stats 2015 [02/06/2016]. Available from: https://www.diabetes.org.uk/Documents/Position%20statements/Diabetes%20UK%20Facts%20and%20Stats_Dec%202015.pdf.
127. Bebb C, Kendrick D, Coupland C, Madeley R, Stewart J, Brown K, et al. A cluster randomised controlled trial of the effect of a treatment algorithm for hypertension in patients with type 2 diabetes. The British journal of general practice : the journal of the Royal College of General Practitioners. 2007;57(535):136-43.
128. Smith SM, Paul G, Kelly A, Whitford DL, O'Shea E, O'Dowd T. Peer support for patients with type 2 diabetes: cluster randomised controlled trial. Bmj. 2011;342:d715.

129. (NICE) NIfHaCE. Type 2 diabetes: The management of type 2 diabetes. NICE guidelines. 2009;[CG87].
130. Currie CJ, Peters JR, Tynan A, Evans M, Heine RJ, Bracco OL, et al. Survival as a function of HbA1c in people with type 2 diabetes: a retrospective cohort study. The Lancet.375(9713):481-9.
131. Kirby M. Achieving effective lipid management in diabetes. British Journal of Primary Care Nursing. 2009;6(2).
132. Razak F, Anand SS, Shannon H, Vuksan V, Davis B, Jacobs R, et al. Defining obesity cut points in a multiethnic population. Circulation. 2007;115(16):2111-8.
133. (NICE) NIfHaCE. Quality Outcomes Framework Indicators 2004. Available from: https://www.nice.org.uk/Standards-and-Indicators/QOFIndicators.
134. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. JClinEpidemiol. 2004;57(8):785-94.
135. Roudsari B, Fowler R, Nathens A. Intracluster correlation coefficient in multicenter childhood trauma studies. Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention. 2007;13(5):344-7.
136. Kul S, Vanhaecht K, Panella M. Intraclass correlation coefficients for cluster randomized trials in care pathways and usual care: hospital treatment for heart failure. BMC health services research. 2014;14:84.
137. Moineddin R, Matheson FI, Glazier RH. A simulation study of sample size for multilevel logistic regression models. BMC medical research methodology. 2007;7:34.
138. Pagel C, Prost A, Lewycka S, Das S, Colbourn T, Mahapatra R, et al. 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. 2011;12:151.
139. Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, et al. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. JEpidemiolCommunity Health. 2006;60(4):290-7.
140. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. ClinTrials. 2005;2(2):99-107.
141. Health Service Research Unit - Database of ICCs 2010 [updated 1/27/2010; cited 2013 9/5/2013]. Available from: http://www.abdn.ac.uk/hsru/research/delivery/behaviour/methodological-research/.
142. Shoukri MM, Donner A, El-Dali A. Covariate-adjusted confidence interval for the intraclass correlation coefficient. Contemporary clinical trials. 2013;36(1):244-53.
143. Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. Annals of internal medicine. 2001;135(2):112-23.
144. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. Clinical trials (London, England). 2005;2(2):152-62.
145. Lauer SA, Kleinman KP, Reich NG. The effect of cluster size variability on statistical power in cluster-randomized trials. PLoS One. 2015;10(4):e0119074.
146. Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clinical trials. 2004;1(1):80-90.
147. Murphy AW, Esterman A, Pilotto LS. Cluster randomized controlled trials in primary care: an introduction. Eur J Gen Pract. 2006;12(2):70-3.
148. van Breukelen GJ, Candel MJ. Calculating sample sizes for cluster randomized trials: we can keep it simple and efficient! Journal of clinical epidemiology. 2012;65(11):1212-8.

149. Stewart SB, Dahm P, Scales CD, Jr. How to appraise the effectiveness of treatment. Indian journal of urology : IJU : journal of the Urological Society of India. 2011;27(4):525-31.
150. Guittet L, Ravaud P, Giraudeau B. Planning a cluster randomized trial with unequal cluster sizes: practical issues involving continuous outcomes. BMC medical research methodology. 2006;6:17.
151. Kasza J, Hemming K, Hooper R, Matthews J, Forbes AB. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Statistical methods in medical research. 2017:962280217734981.
152. Hemming KM, J. A menu-driven facility for sample-size calculations in cluster randomized controlled trials. Stata Journal. 2013;13:114-35(22).

CHAPTER 9: Appendices

Appendix A

1. step$ wedge.ti,ab.

2. experimentally staged introduction.ti,ab.

3. delayed intervention.ti,ab.

4. (one directional cross over design or one directional crossover design).ti,ab.

5. ((incremental or phased or staggered or stepwise or step wise or delayed) adj1 (recruitment or introduction or implementation)).ti,ab.

6. or/1-5

7. limit 6 to english language

8. Remove duplicates from 7

9. limit 8 to randomized controlled trial

10. limit 9 to humans

Appendix B

Reviewers initials:	KH
Reviewers paper number:	29
Study title:	Does a multidisciplinary nutritional intervention prevent nutritional decline in hospital patients? A stepped wedge randomised cluster trial
Authors:	Timothy J. Schultz, Alison L. Kitson, Stijn Soenen, Leslye Long, Alison Shanks, Rick Wiechula a, Ian Chapman d, Kylie Lange d
Full trial or protocol?	Full Trial
Journal:	e-SPEN Journal
Journal impact factor:	Unknown
Year of publication:	2014
Paper ID:	SW080
Citation details of any linked protocol:	-

General study demographics			Additional comments
What is the country of study (or primary country of study)?	Australia	Canada	
	USA	UK or Ireland	
	Other:	....	
What is defined as a cluster? (Please capture the exact term used to describe the cluster and enter in the available space)	GP practices	Hospitals	
	Wards/specialities	Medical clinic/unit	
	Residential area	School	
	Other		
	....		
Is the study in a healthcare or non-healthcare setting?	Healthcare	Non-healthcare	
How many interventions are being compared?	1	2	
	3 or more		
What method of allocation is used?	Unrestricted	Stratified	
	Paired randomisation	Not reported	
	Other:	....	

General study demographics			Additional comments
What is the primary outcome data type?	Binary	Continuous	
	Categorical	Count	
	Ordinal	Rate	
	Time to event	Unclear	
	Not reported		
	Other:	……………………………………	

Reporting of items required under CONSORT RCT				Additional comments
1. (i)	Was there a justification of the sample size?	Yes	No	
1. (ii)	What was the basis of this sample size justification?	Pragmatic	Statistical	
		Unclear	N/A	
1. (iii)	Is the sample size or power calculation reported?	Yes	No	
		Unclear	N/A	
2.	Is the significance level (type I error rate) reported?	Yes	No	
		Unclear	N/A	
3.	Is the expected power of the trial reported?	Yes	No	
		Unclear	N/A	
4.	Is the outcome used in the sample size calculation consistent with the primary outcome of the trial?	Yes	No	
		Unclear	N/A	
5.	Was the estimated treatment effect sufficiently reported? (Details regarding sufficient reporting are given in supporting document)	Yes	No	
		Unclear	N/A	
6.	Was an allowance in the sample size made for attrition or non-compliance reported?	Yes	No	
		Unclear	N/A	

Reporting of items required under CONSORT extension for CRTs				Additional comments
7. (i)	Is the total sample size planned for the trial reported?	Explicitly reported	Deducible	
		Unclear	Not reported	
7. (ii)	Is the planned cluster size (over the whole trial duration) reported?	Explicitly reported	Deducible	
		Unclear	Not reported	
7. (iii)	Is the number of clusters planned for the trial reported?	Explicitly reported	Deducible	
		Unclear	Not reported	
8. (i)	Is an estimate of the clustering reported?	Yes	No	
		Unclear		
8. (ii)	How was the estimate of clustering reported?	ICC	CV	
		Design Effect	N/A	
		Other:	………………………………	
8. (iii)	Is an indication of the uncertainty surrounding the estimate of clustering (i.e. ICC) reported?	Yes	No	
		Unclear	N/A	
9.	Is the expected variation of the cluster sizes reported?	Yes	No	
		Unclear	N/A	

Reporting of items required under CONSORT extension for stepped wedge				Additional comments
10. (i)	Is the planned number of randomisation points reported?	Explicitly reported	Deducible	
		Unclear	Not reported	
10. (ii)	Is the planned number of measurement periods reported?	Explicitly reported	Deducible	
		Unclear	Not reported	
11.	Is the planned number of clusters that are to be randomised per step reported?	Explicitly reported	Deducible	
		Unclear	Not reported	

Reporting of items required under CONSORT extension for stepped wedge				Additional comments
12. (i)	Is it reported whether patients are followed up for all time periods (cohort design), whether new patients are recruited at each step (cross-sectional design), or whether the study is a combination of both (open cohort)?	Explicitly reported Unclear	Deducible Not reported	
12. (ii)	What type of design does the study take?	Cohort design Open cohort	Cross-sectional design Not reported/unclear	
13. (i)	Is there clarity between the cluster sample size per measurement period and total cluster sample size?	Yes Unclear	No	
13 (ii)	Which, of the total cluster sample size and the cluster sample size per measurement period, is reported?	Total cluster sample size Both Unclear	Cluster sample size per measurement period Neither	

Methodological assessment elements				Additional comments
14. (i)	What method was used to determine the sample size? (Note: The Woertman correction for cross-sectional design may be known as the Hemming correction)	Hussey & Hughes Woertman (without correcting for cross-sectional design) Simulation Individually randomised design Not stated Other (please specify)	Parallel CRT Woertman (with correcting for cross-sectional design) Before & after CRT N/A	
14 (ii)	If simulation methods were used, did they account for time effects?	Yes Unclear	No N/A	
14 (iii)	Was a time by treatment effect interaction included in the power calculation?	Yes Unclear	No N/A	

Methodological assessment elements				Additional comments
14. (iv)	Was an inter period correlation included in the power calculation?	Yes Unclear	No N/A	
14. (v)	Was a transition period allowed for in the power calculation?	Yes Unclear	No N/A	
14. (vi)	Were any other design variations allowed for in the power calculation?	Yes Unclear	No N/A	
14. (vii)	If yes, what variations were allowed for?	…………………………….	N/A	
14. (viii)	Should any other variation have been allowed for?	…………………………….	N/A	
14. (ix)	Was varying cluster size allowed for in the power calculation?	Yes Unclear	No N/A	
14. (x)	For a cohort, or open cohort design, were repeated measurements allowed for in the power calculation?	Yes Unclear	No N/A	
15.	What rationale was reported for choosing a cluster trial?	Avoid contamination Cluster level intervention Other: None reported	Practical reasons ………………………………	
16.	What rationale was reported for choosing the stepped wedge design?	Methodological Intention for all clusters to receive the treatment Evidence of effectiveness in other settings Evaluating a routine roll out Resource constraints Other: None reported	Ethical concerns Need for staggered implementation Expect that intervention will do more good than harm Evaluation as a natural experiment Social acceptability ………………………………	
17.	Is there a schematic representation of the design of the study?	Yes Unclear	No	

All further questions relate to full trials only.

Realised design characteristics				Additional comments
18. (i)	What is the total number of clusters?		Not reported	
18. (ii)	What is the (average) total cluster size?		Not reported	
18. (iii)	What is the (average) cluster size per measurement period?		Not reported	
18. (iv)	How many randomisation points does the study use?		Not reported	
18. (v)	How many measurement points does the study use? (Defined as a point in time in which data are collected		Not reported	
18. (vi)	How many clusters are randomised per step? (If this number varies, what is the average?)		Not reported	
18 (vii)	Does the design of the study differ from a "complete" stepped wedge design in any way? (i.e. Is the design non-	Contains extended or reduced pre & post periods No period with all clusters unexposed Unclear Other:	Contains transition periods Not all clusters receive the intervention No	
19	Is the outcome routinely collected data?	Yes Unclear	No	
20 (i)	What was the total study length? (including units) (Real study duration)		Not reported	
20 (ii)	What was the (average) step length? (including units)		Not reported	

Methods of analysis				Additional comments
22. (i)	Was clustering accounted for in the analysis?	Yes	No	
		Unclear		
22. (ii)	If clustering was accounted for in the analysis, how was this done?	GEE	GLMM	
		Fixed cluster effects	Adjustment of standard errors (i.e. robust)	
		Unclear	N/A	
		Other:	……………………………	
22. (iii)	Were time effects accounted for in the analysis?	Yes – fixed effect	Yes – random effect	
		Yes – other	Yes - linear	
		Yes - but not specified how accounted for	No – but argued not significant	
		Unclear	No	
22. (iv)	Was an inter period correlation included in the analysis?	Yes	No	
		Unclear		
22. (v)	Was a time by treatment effect interaction included in the analysis?	Yes	No	
		Unclear		
22. (vi)	For a cohort or open cohort design, did the analysis account for repeated measurements?	Yes	No	
		Unclear	N/A	

Appendix C

Sample size reporting – full trials vs protocols

Table 9.1 compares the reporting of the CONSORT statement and the extension for the CONSORT statement for cluster trial for full trials reports and published protocols. For many items, there is a great disparity between protocols and full trials. A large proportion of protocols (93%) reported a sample size calculation, compared to 59% of full trials (p=0.003). The reporting of the power (p=0.003) and significance level (p=0.039) were significantly higher in protocols than full trials. The treatment effect was sufficiently reported in 22 (79%) of protocols compared to only 11 (34%) of full trials. The reporting of attrition and of the number of clusters was similar in both groups. A measure of the variation of outcomes within a cluster was poorly reported in full trials (34%), but much higher in protocols (79%). No full trial reports considered an uncertainty in this estimate. The median number of items reported in full trials was 4 (IQR 1 – 6), and 6 (5 – 7) for protocols.

Table 9.1: Sample size reporting - full trials vs protocols

	Protocols N = 28	Full reports N = 32	Absolute difference (95% confidence interval)	P value
Sample size justification				
Reported	26 (93)	19 (59)	33.5 (14.0 to 53.0)	0.003
ITEM 1:				
Level of significance	22 (79)	17 (53)	25.4 (2.4 to 48.5)	0.039
ITEM 2:				
Power	26 (93)	19 (59)	33.5 (14.0 to 53.0)	0.003
ITEM 3:				
Treatment effect [1]	22 (79)	11 (34)	44.2 (21.8 to 66.6)	0.001
ITEM 4:				
Consistency with primary outcome	22 (79)	16 (50)	28.6 (5.5 to 51.6)	0.022
ITEM 5:				
Allowance for attrition	11 (39)	7 (22)	17.4 (-5.7 to 40.5)	0.142
ITEM 6:				
Number of clusters	27 (96)	31 (97)	-0.4 (-9.6 to 8.7)	0.923
Median cluster size	24 (86)	15 (47)	38.8 (17.2 to 60.4)	0.002
ITEM 7:				
Variation in cluster size*	2 (7)	4 (13)	-5.4 (-20.3 to 9.6)	0.675
ITEM 8				
Variation in outcomes across clusters (i.e. ICC)	22 (79)	11 (34)	44.2 (21.8 to 66.6)	0.001
ITEM 9:				
Uncertainty of ICC (or equivalent)*	8 (29)	0 (0)	28.6 (11.8 to 45.3)	0.001
All ITEMS				
Number items reported Median [IQR]	6 [5 to 7]	4 [1 to 6]	2.20 (1.16 to 3.24)	<0.001
Reporting all 9 items	0 (0)	0 (0)		

[1]: a sufficient reporting of the treatment effect consists of either a standardised effect size; a mean difference and standard deviation; means in both arms and standard deviation; proportions in both arms; proportion in one arm and a difference. IQR: Inter-quartile range. ICC: Intra Cluster Correlation. P-value is for the comparison of full trial publications and protocols using a chi-squared test for proportions (categorical outcomes) or Mann-Whitney U test (where medians are reported), or (*) using Fisher's exact test.

A comparison of the reporting of items relating to a SW-CRT between full trials and protocols is given in Table 9.2. The number of randomisation points (steps) was reported explicitly in all protocols, with 81% of full trials explicitly stating the number, with it deducible in 97% of full trial reports. The reporting of the number of clusters randomised per step was high for both report types (89% vs 97%). A schematic illustration was included in 25 (89%) of the protocols, compared to 66% of full trial reports (p = 0.031). The reporting of the study design type was poor for both report types. A high number of full trials reports did not report (or did not report clearly) the cluster size (59%). There was a larger degree of clarity of cluster size in protocols, with only 32% failing to clearly report the cluster size.

Table 9.2: Sample size reporting – stepped-wedge Items - Full trials vs Protocols

	Protocols N = 28	Full reports N = 32	Absolute difference (95% confidence interval)	P value
Number of steps				
Explicitly reported	28 (100)	26 (81)	18.8 (5.2 to 32.3)	0.016
Reported or deducible	28 (100)	31 (97)	3.1 (-2.9 to 9.2)	0.346
Number clusters randomised per step				
Reported	25 (89)	31 (97)	-7.6 (-20.5 to 5.4)	0.240
Schematic representation				
Reported	25 (89)	21 (66)	23.7 (3.6 to 43.7)	0.031
Design type (i.e. cross-sectional/cohort)				
Explicitly reported	8 (29)	8 (25)	3.6 (-18.9 to 26.0)	0.755
Reported or deducible	22 (79)	21 (66)	12.9 (-9.5 to 35.3)	0.267
Clarity of cluster size[1]				
Total cluster size reported	10 (36)	7 (22)	13.8 (-9.0 to 36.6)	0.235
Cluster size per measurement period	14 (50)	11(34)	15.6 (-9.2 to 40.4)	0.221
Unclear/not reported	9 (32)	19 (59)	-30.4 (-54.5 to -6.3)	0.035

[1]: some studies reported both total cluster size and cluster size per measurement period; P-value is for the comparison of protocols to full trial reports using a chi-squared test for proportions.

Appendix D

Derivation of the intra-cluster correlation coefficient

The intra-cluster correlation coefficient (ICC), is the correlation between two participants in the same cluster independent of time. For two participant's *k* and *k'* in cluster *i*, the correlation between them can be given as:

$$ICC = Corr(Y_{ik}, Y_{ik'})$$

Now, the correlation of two outcomes can be estimated using a function of the covariance of the two outcomes and their corresponding variances. As such, the ICC can be written as:

$$ICC = Corr(Y_{ik}, Y_{ik'}) = \frac{Cov(Y_{ik}, Y_{ik'})}{\sqrt{Var(Y_{ik}) \times Var(Y_{ik'})}}$$

Now, the covariance between Y_{ik} and $Y_{ik'}$ can be calculated using the model framework as:

$$Cov(Y_{ik}, Y_{ik'}) = Cov(\alpha_i + \varepsilon_{ik}, \alpha_i + \varepsilon_{ik'})$$

$$= Cov(\alpha_i, \alpha_i) + \ Cov(\varepsilon_{ik}, \varepsilon_{ik'}) + Cov(\alpha_i, \varepsilon_{ik'}) + Cov(\varepsilon_{ik},)$$

$$= Var(\alpha_i) \ = {\sigma_b}^2$$

The variance of Y_{ik} and $Y_{ik'}$ can also be calculated as:

$$\sqrt{Var(Y_{ik}) \times Var(Y_{ik'})} = \sqrt{({\sigma_b}^2 + {\sigma_w}^2) \times ({\sigma_b}^2 + {\sigma_w}^2)} = {\sigma_b}^2 + {\sigma_w}^2$$

Now, the ICC can be written as:

$$ICC = Corr(Y_{ik}, Y_{ik'}) = \frac{{\sigma_b}^2}{{\sigma_b}^2 + {\sigma_w}^2}$$

Appendix E

Derivation of the within-cluster correlation

Consider a stepped-wedge cluster randomised trial with T time-periods.

Let X and Y be randomly chosen observations from within the same cluster.

We now define J = 1, 0 which indicates whether observations are, or are not, made at the same time-period, so that:

$$J\begin{cases} 1 & \text{if X and Y are made at the same time-period} \\ 0 & \text{if X and Y are not made at the same time-period} \end{cases}$$

Now, let us assume that Var(X) = σ^2 = 1

When considering the conditional covariance formula:

$$WCC = cov(X,Y) = E_j cov(X,Y|J) + cov_J\{E(X|J), E(Y|J)\}$$

Since $E(X|J)$ is a constant, $cov_J\{E(X|J), E(Y|J)\} = 0$, and so

$$WCC = cov(X,Y) = E_j cov(X,Y|J)$$

Now, the term $cov(X,Y|J)$ can be broken down into the J = 0 and J = 1 components, so that:

$$cov(X,Y|J=0) = IPC \qquad\qquad cov(X,Y|J=1) = WPC$$

Now, $Pr\{J=1\} = \frac{1}{T} = 1 - Pr\{J=0\}$

As such,

$$WCC = cov(X,Y) = \left(1-\frac{1}{T}\right) IPC + \frac{1}{T} WPC \; = \; IPC + \frac{1}{T}(WPC - IPC)$$

Appendix F

Derivation of the within-period correlation

The within-period correlation (WPC), is the correlation between participants in the same cluster during the same time-period. For two participant's *k* and *k'* in cluster *i* at time *j*, the correlation between them can be given as:

$$WPC = Corr\left(Y_{ijk}, Y_{ijk'}\right)$$

Now, the correlation of two outcomes can be estimated using a function of the covariance of the two outcomes and their corresponding variances. As such, the WPC can be written as:

$$WPC = Corr\left(Y_{ijk}, Y_{ijk'}\right) = \frac{Cov\left(Y_{ijk}, Y_{ijk'}\right)}{\sqrt{Var\left(Y_{ijk}\right) \times Var\left(Y_{ijk'}\right)}}$$

Now, the covariance between Y_{ijk} and $Y_{ijk'}$ can be calculated using the model framework as:

$$Cov\left(Y_{ijk}, Y_{ijk'}\right) = Cov\left(\alpha_i + \omega_{ij} + \varepsilon_{ijk}, \alpha_i + \omega_{ij} + \varepsilon_{ijk'}\right)$$

$$= Cov(\alpha_i, \alpha_i) + \ Cov\left(\alpha_i, \omega_{ij}\right) + \ Cov\left(\alpha_i, \varepsilon_{ijk'}\right) + Cov\left(\omega_{ij}, \alpha_i\right) + \ Cov\left(\omega_{ij}, \omega_{ij}\right)$$

$$+ \ Cov\left(\alpha_i, \varepsilon_{ijk'}\right) + Cov\left(\varepsilon_{ijk}, \alpha_i\right) + \ Cov\left(\varepsilon_{ijk}, \omega_{ij}\right) + \ Cov\left(\varepsilon_{ijk}, \varepsilon_{ijk'}\right)$$

$$= Var(\alpha_i) + Var\left(\omega_{ij}\right) = \tau^2 + {\sigma_t}^2$$

The variance of Y_{ijk} and $Y_{ijk'}$ can also be calculated as:

$$\sqrt{Var\left(Y_{ijk}\right) \times Var\left(Y_{ijk'}\right)} = \sqrt{\left(\tau^2 + {\sigma_t}^2 + {\sigma_p}^2\right) \times \left(\tau^2 + {\sigma_t}^2 + {\sigma_p}^2\right)} = \tau^2 + {\sigma_t}^2 + {\sigma_p}^2$$

Now, the WPC can be written as:

$$WPC = Corr\left(Y_{ijk}, Y_{ijk'}\right) = \frac{\tau^2 + {\sigma_t}^2}{\tau^2 + {\sigma_t}^2 + {\sigma_p}^2}$$

Appendix G

Derivation of the inter-period correlation

The inter-period correlation (IPC), indicates the correlation between participants in the same cluster at different time-periods. As such, it provides the correlation between participant's *k* and *k'* from cluster *i* at times *j* and *j'*, and can be given as:

$$IPC = Corr\left(Y_{ijk}, Y_{ij'k'}\right)$$

The correlation between two participants *k* and *k'* from cluster *i* at different time-periods *j* and *j'* can be described using their covariance and variance as follows:

$$WPC = Corr\left(Y_{ijk}, Y_{ij'k'}\right) = \frac{Cov\left(Y_{ijk}, Y_{ij'k'}\right)}{\sqrt{Var\left(Y_{ijk}\right) \times Var\left(Y_{ij'k'}\right)}}$$

Where:

$$Cov\left(Y_{ijk}, Y_{ij'k'}\right) = Cov\left(\alpha_i + \omega_{ij} + \varepsilon_{ijk}, \alpha_i + \omega_{ij} + \varepsilon_{ij'k'}\right)$$

$$= Cov(\alpha_i, \alpha_i) + Cov\left(\alpha_i, \omega_{ij'}\right) + Cov\left(\alpha_i, \varepsilon_{ij'k'}\right) + Cov\left(\omega_{ij}, \alpha_i\right) + Cov\left(\omega_{ij}, \omega_{ij'}\right)$$
$$+ Cov\left(\alpha_i, \varepsilon_{ij'k'}\right) + Cov\left(\varepsilon_{ijk}, \alpha_i\right) + Cov\left(\varepsilon_{ijk}, \omega_{ij'}\right) + Cov\left(\varepsilon_{ijk}, \varepsilon_{ij'k'}\right)$$

$$= Var(\alpha_i) = \tau^2$$

And

$$\sqrt{Var\left(Y_{ijk}\right) \times Var\left(Y_{ij'k'}\right)} = \sqrt{\left(\tau^2 + {\sigma_t}^2 + {\sigma_p}^2\right) \times \left(\tau^2 + {\sigma_t}^2 + {\sigma_p}^2\right)} = \tau^2 + {\sigma_t}^2 + {\sigma_p}^2$$

Now, the IPC can be written as:

$$IPC = Corr\left(Y_{ijk}, Y_{ij'k'}\right) = \frac{\tau^2}{\tau^2 + {\sigma_t}^2 + {\sigma_p}^2}$$

Appendix H

Table 9.3: Dates used to define periods to estimate the impact of study length on the IPC and WPC

Number of periods	Dates for each period
2	**P1:** 1st January 2007 to 31st December 2007
	P2: 1st January 2008 to 31st December 2008
3	**P1:** 1st January 2006 to 31st December 2006
	P2: 1st January 2007 to 31st December 2007
	P3: 1st January 2008 to 31st December 2008
4	**P1:** 1st January 2005 to 31st December 2005
	P2: 1st January 2006 to 31st December 2006
	P3: 1st January 2007 to 31st December 2007
	P4: 1st January 2008 to 31st December 2008
5	**P1:** 1st January 2004 to 31st December 2004
	P2: 1st January 2005 to 31st December 2005
	P3: 1st January 2006 to 31st December 2006
	P4: 1st January 2007 to 31st December 2007
	P5: 1st January 2008 to 31st December 2008
6	**P1:** 1st January 2003 to 31st December 2003
	P2: 1st January 2004 to 31st December 2004
	P3: 1st January 2005 to 31st December 2005
	P4: 1st January 2006 to 31st December 2006
	P5: 1st January 2007 to 31st December 2007
	P6: 1st January 2008 to 31st December 2008

Table 9.4: Dates used to define periods to estimate the impact of period length on the IPC and WPC

Number of periods	Dates for each period
2	**P1:** 1st February 2009 to 15th September 2009 **P2:** 16th September 2009 to 30th April 2010
3	**P1:** 1st February 2009 to 30th June 2009 **P2:** 1st July 2009 to 30th November 2009 **P3:** 1st December 2009 to 30th April 2010
4	**P1:** 1st February 2009 to 24th May 2009 **P2:** 25th May 2009 to 15th September 2009 **P3:** 16th September 2009 to 6th January 2010 **P4:** 7th January 2010 to 30th April 2010
5	**P1:** 1st February 2009 to 30th April 2009 **P2:** 1st May 2009 to 31st July 2009 **P3:** 1st August 2009 to 31st October 2009 **P4:** 1st November 2009 to 31st January 2009 **P5:** 1st February 2010 to 30th April 2010

Appendix I

Table 9.5: Intra-cluster correlation coefficients (ICCs) for binary outcomes (dichotomised clinical measures) associated with type-2 diabetes from THIN database for study period 01/02/2009 to 30/04/2010 from an adjusted model

Outcome	Prevalence of outcome	Latent ICC (95% CI)	Natural ICC (95% CI)
Clinical Measures			
HbA1c (>7.5)	0.34240	0.037 (0.031 to 0.043)	0.027 (0.023 to 0.031)
Systolic blood pressure (>140)	0.27024	0.0595 (0.051 to 0.069)	0.035 (0.030 to 0.041)
Systolic blood pressure (>130)	0.58416	0.043 (0.037 to 0.050)	0.035 (0.030 to 0.040)
Diastolic blood pressure (>80)	0.24068	0.084 (0.073 to 0.097)	0.045 (0.039 to 0.051)
BMI (>30)	0.49166	0.021 (0.017 to 0.025)	0.015 (0.013 to 0.019)
BMI (>25)	0.83543	0.021 (0.017 to 0.027)	0.010 (0.008 to 0.012)
Total cholesterol (>4)	0.50256	0.026 (0.022 to 0.030)	0.020 (0.017 to 0.024)
HDL cholesterol (<1.2)	0.50389	0.036 (0.031 to 0.043)	0.027 (0.023 to 0.032)
Medication			
Taking of Insulin	0.59059	0.108 (0.095 to 0.123)	0.076 (0.067 to 0.087)
Clinical Outcomes			
Atrial fibrillation	0.01034	0.010 (0.002 to 0.052)	0.000 (0.000 to 0.002)
Chronic kidney disease	0.00332	0.132 (0.088 to 0.193)	0.002 (0.002 to 0.003)
Chronic obstructive pulmonary disease	0.00804	0.053 (0.032 to 0.087)	0.002 (0.001 to 0.003)
Ischaemic heart disease	0.01061	0.026 (0.011 to 0.060)	0.001 (0.000 to 0.002)
Peripheral vascular disease	0.00537	0.117 (0.084 to 0.162)	0.004 (0.003 to 0.005)
Stroke	0.00410	0.258 (0.200 to 0.325)	0.010 (0.008 to 0.012)

[1]: Adjusted for age, sex, location, and deprivation quintiles

Table 9.6: Estimates of the within-period correlation and inter-period correlation for continuous outcomes associated with type-2 diabetes from THIN database for study period 01/11/2007 to 30/04/2010 from an adjusted model

Outcome	**WPC** **(95% CI)**	**IPC** **(95% CI)**	**CA**
Clinical measures			
HbA1c (%) [2]	0.0348 (0.0298 to 0.0405)	0.0186 (0.0135 to 0.0255)	0.5334
Systolic blood pressure (mmHg) [3]	0.0286 (0.0245 to 0.0332)	0.0174 (0.0127 to 0.0239)	0.6107
Diastolic blood pressure (mmHg) [3]	0.0392 (0.0339 to 0.0452)	0.0287 (0.0227 to 0.0364)	0.7341
BMI (kg/m^2) [3]	0.0219 (0.0186 to 0.0258)	0.017 (0.0134 to 0.0216)	0.775
Total cholesterol (mmol/L) [3]	0.021 (0.0179 to 0.0247)	0.0068 (0.0035 to 0.0133)	0.3245
HDL cholesterol (mmol/L) [3]	0.0206 (0.0175 to 0.0243)	0.0175 (0.0139 to 0.0220)	0.8482

WPC: Within-period correlation. IPC: Inter-period correlation. CA: Cluster autocorrelation. CI: Confidence interval. BMI: Body mass index. [1]: Adjusted for age, sex, location, and deprivation quintiles. [2]: Two consecutive 12-month periods were used (01/01/2007 – 31/12/2007 & 01/01/2008 – 31/12/2008). [3]: Two consecutive 15-month periods were used.

Table 9.7: Impact of increasing the study length on the within-period correlation and inter-period correlation for HbA1c from included participants from THIN database when maintaining a one-year period length from an adjusted model

Study length (years)	Number of periods	WPC (95% CI)	IPC (95% CI)	CA
2	2	0.0348 (0.0298 to 0.0405)	0.0186 (0.0135 to 0.0255)	0.5334
3	3	0.0382 (0.0325 to 0.0448)	0.0173 (0.0124 to 0.0240)	0.4532
4	4	0.0396 (0.0337 to 0.0466)	0.0176 (0.0128 to 0.0241)	0.4436
5	5	0.0402 (0.0341 to 0.0473)	0.0166 (0.0119 to 0.0232)	0.4136
6	6	0.0412 (0.0349 to 0.0486)	0.0166 (0.0119 to 0.0231)	0.4023

WPC: Within-period correlation. IPC: Inter-period correlation. CA: Cluster autocorrelation. CI: Confidence interval. [1]*: Adjusted for age, sex, location, and deprivation quintiles.*

Table 9.8: Impact of period length on the within-period correlation and inter-period correlation for continuous outcomes associated with type-2 diabetes from THIN database for study period 01/02/2009 to 30/04/2010 from an adjusted model

Period Length	WPC[1] (95% CI)	IPC[1] (95% CI)	CA[1]
7.5 month period length			
Systolic blood pressure (mmHg)	0.040 (0.032 to 0.042)	0.014 (0.010 to 0.020)	0.387
Diastolic blood pressure (mmHg)	0.040 (0.034 to 0.045)	0.034 (0.028 to 0.040)	0.867
BMI (kg/m^2)	0.022 (0.020 to 0.027)	0.020 (0.017 to 0.024)	0.872
Total cholesterol (mmol/L)	0.021 (0.017 to 0.024)	0.017 (0.014 to 0.021)	0.869
HDL cholesterol (mmol/L)	0.023 (0.019 to 0.026)	0.019 (0.015 to 0.023)	0.839
Median			
5 month period length			
Systolic blood pressure (mmHg)	0.042 (0.035 to 0.044)	0.019 (0.015 to 0.024)	0.492
Diastolic blood pressure (mmHg)	0.041 (0.034 to 0.046)	0.036 (0.030 to 0.042)	0.894
BMI (kg/m^2)	0.024 (0.021 to 0.029)	0.020 (0.017 to 0.024)	0.809
Total cholesterol (mmol/L)	0.021 (0.017 to 0.024)	0.018 (0.015 to 0.022)	0.875
HDL cholesterol (mmol/L)	0.024 (0.020 to 0.027)	0.019 (0.016 to 0.023)	0.825
Median			
3.75 month period length			
Systolic blood pressure (mmHg)	0.043 (0.035 to 0.045)	0.021 (0.017 to 0.026)	0.529
Diastolic blood pressure (mmHg)	0.040 (0.034 to 0.046)	0.036 (0.031 to 0.042)	0.912
BMI (kg/m^2)	0.023 (0.021 to 0.029)	0.021 (0.017 to 0.025)	0.840
Total cholesterol (mmol/L)	0.022 (0.017 to 0.024)	0.018 (0.015 to 0.022)	0.873
HDL cholesterol (mmol/L)	0.025 (0.020 to 0.028)	0.019 (0.016 to 0.023)	0.811
Median			
3 month period length			
Systolic blood pressure (mmHg)	0.044 (0.037 to 0.047)	0.022 (0.018 to 0.027)	0.529
Diastolic blood pressure (mmHg)	0.041 (0.035 to 0.046)	0.037 (0.031 to 0.043)	0.915
BMI (kg/m^2)	0.025 (0.022 to 0.030)	0.021 (0.017 to 0.025)	0.807
Total cholesterol (mmol/L)	0.022 (0.018 to 0.025)	0.018 (0.015 to 0.022)	0.875
HDL cholesterol (mmol/L)	0.026 (0.021 to 0.029)	0.020 (0.016 to 0.023)	0.789
Median			
1 month period length			
Systolic blood pressure (mmHg)	0.046 (0.039 to 0.049)	0.025 (0.021 to 0.030)	0.578
Diastolic blood pressure (mmHg)	0.043 (0.037 to 0.048)	0.037 (0.032 to 0.043)	0.883
BMI (kg/m^2)	0.027 (0.023 to 0.032)	0.021 (0.018 to 0.025)	0.784
Total cholesterol (mmol/L)	0.022 (0.018 to 0.025)	0.019 (0.016 to 0.022)	0.884
HDL cholesterol (mmol/L)	0.026 (0.021 to 0.029)	0.020 (0.017 to 0.024)	0.835
Median			

WPC: Within-period correlation. IPC: Inter-period correlation. CA: Cluster autocorrelation. CI: Confidence interval. [1]: Adjusted for age, sex, location, and deprivation quintiles.

Appendix J

The Gamma distribution has a scaling property, so that, if $X \sim \Gamma(\alpha,\beta)$, the distribution of $Y = cX$ (where c is some scalar) also follows the Gamma distribution.

Firstly, let X follow a Gamma distribution so that $X \sim \Gamma(\alpha,\beta)$. As such, the moment generating function of X can be given as:
$M_x(X) = (1 - \beta X)^{-\alpha}$

Now, let c be a positive, real constant. Now, we consider the transformation $Y = g(X) = cX$. From this , the moment generating function of Y can be given as:
$M_Y(Y) = M_{cX}(X) = M_X(cX) = (1 - c\beta X)^{-\alpha}$

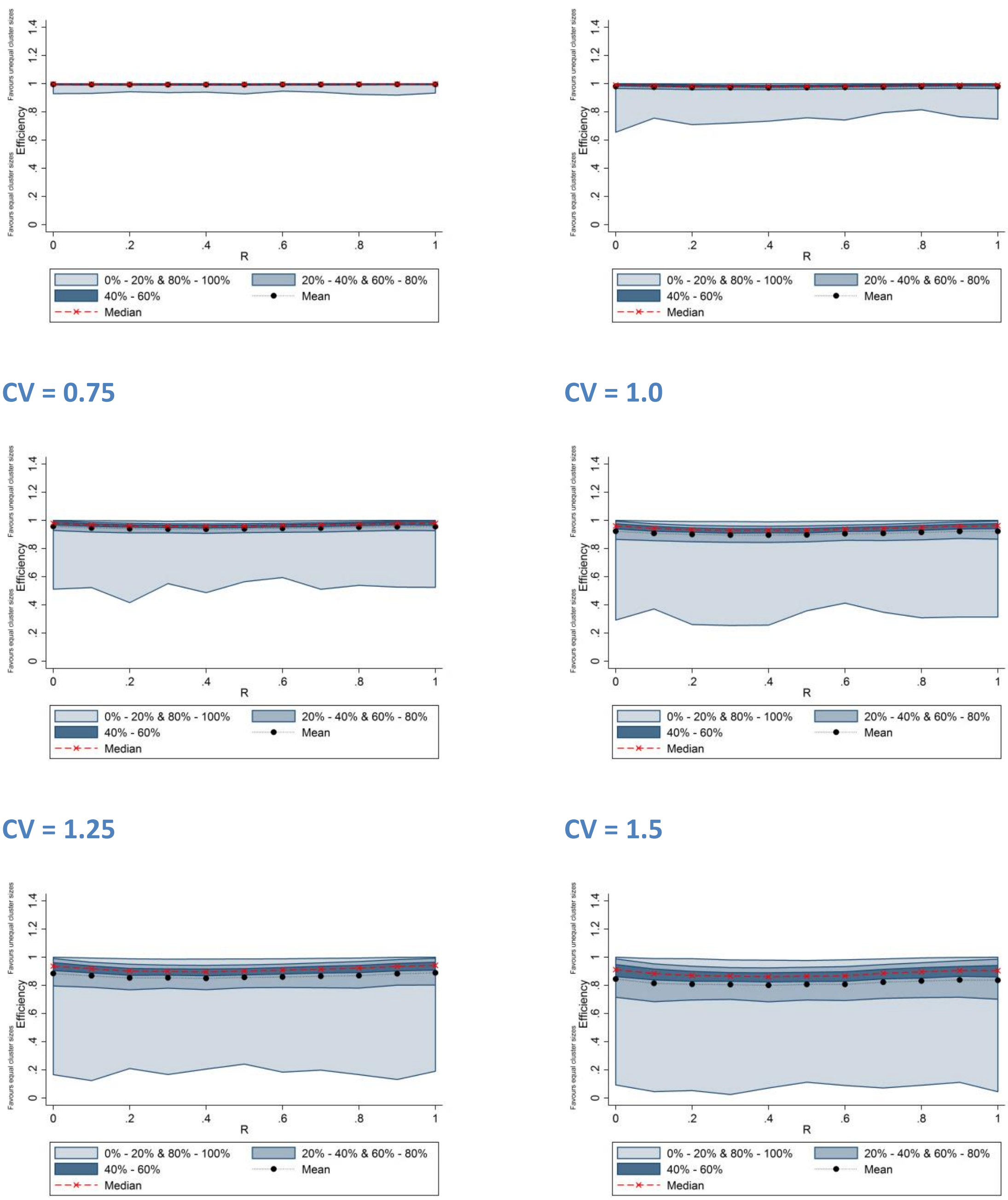
This indicates that Y has a Gamma density given as $\Gamma(\alpha, c\beta)$.

Appendix K

Below, we present the efficiency curves for the simulation study to assess the impact of between-cluster variation in a SW-CRT. Results are presented grouped by the number of clusters in the SW-CRT.

Efficiency of a SW-CRT with 12 clusters

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 12 clusters and 2 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

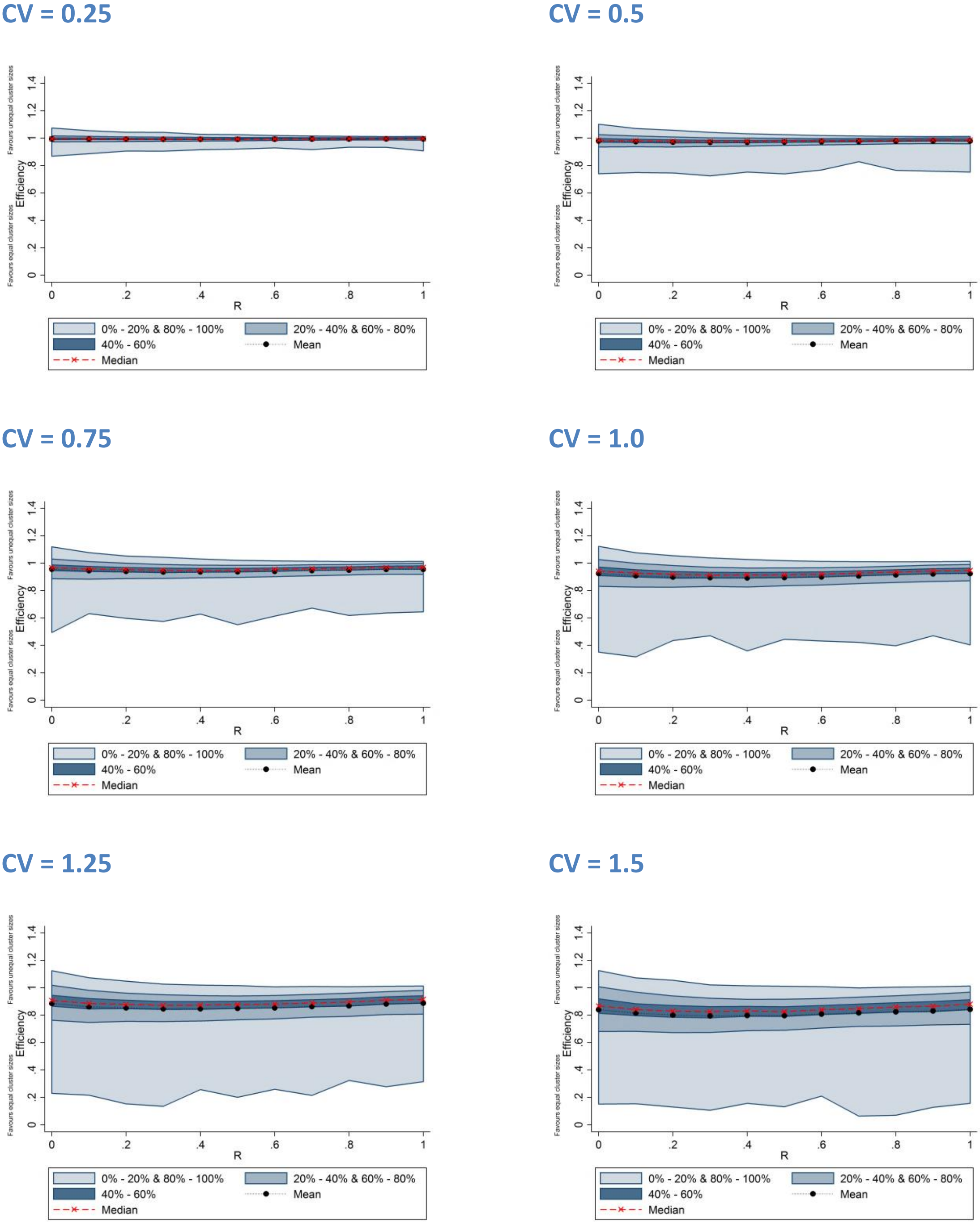
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 12 clusters and 3 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

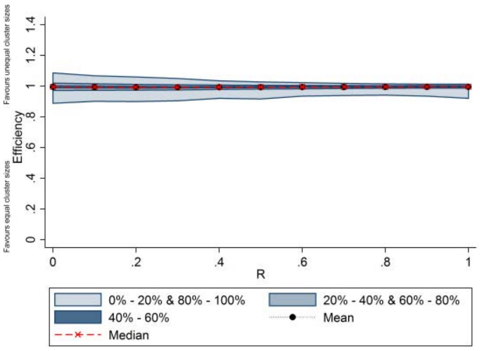
CV = 1.0

CV = 1.25

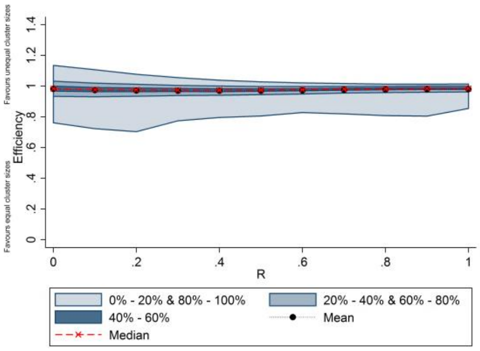
CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

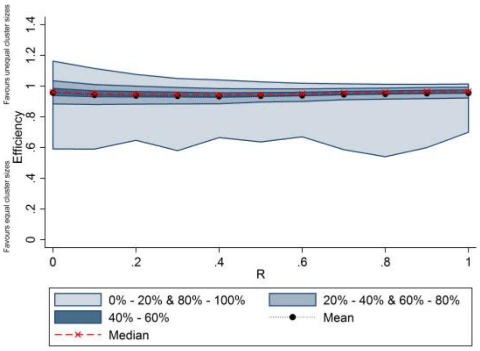
CV = 0.25



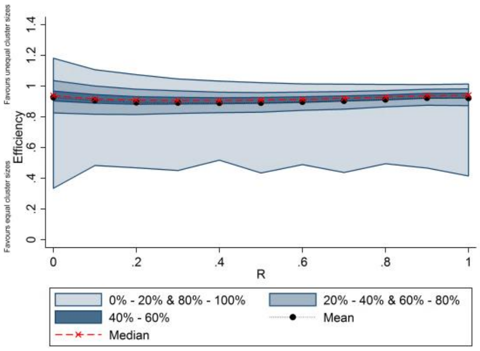
CV = 0.5



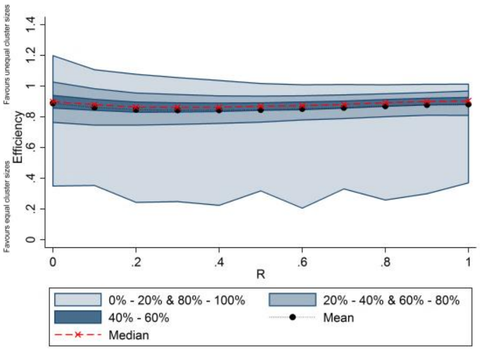
CV = 0.75



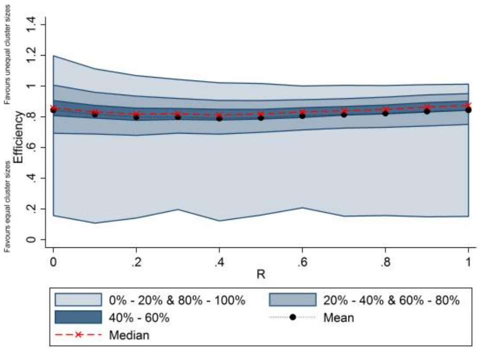
CV = 1.0



CV = 1.25

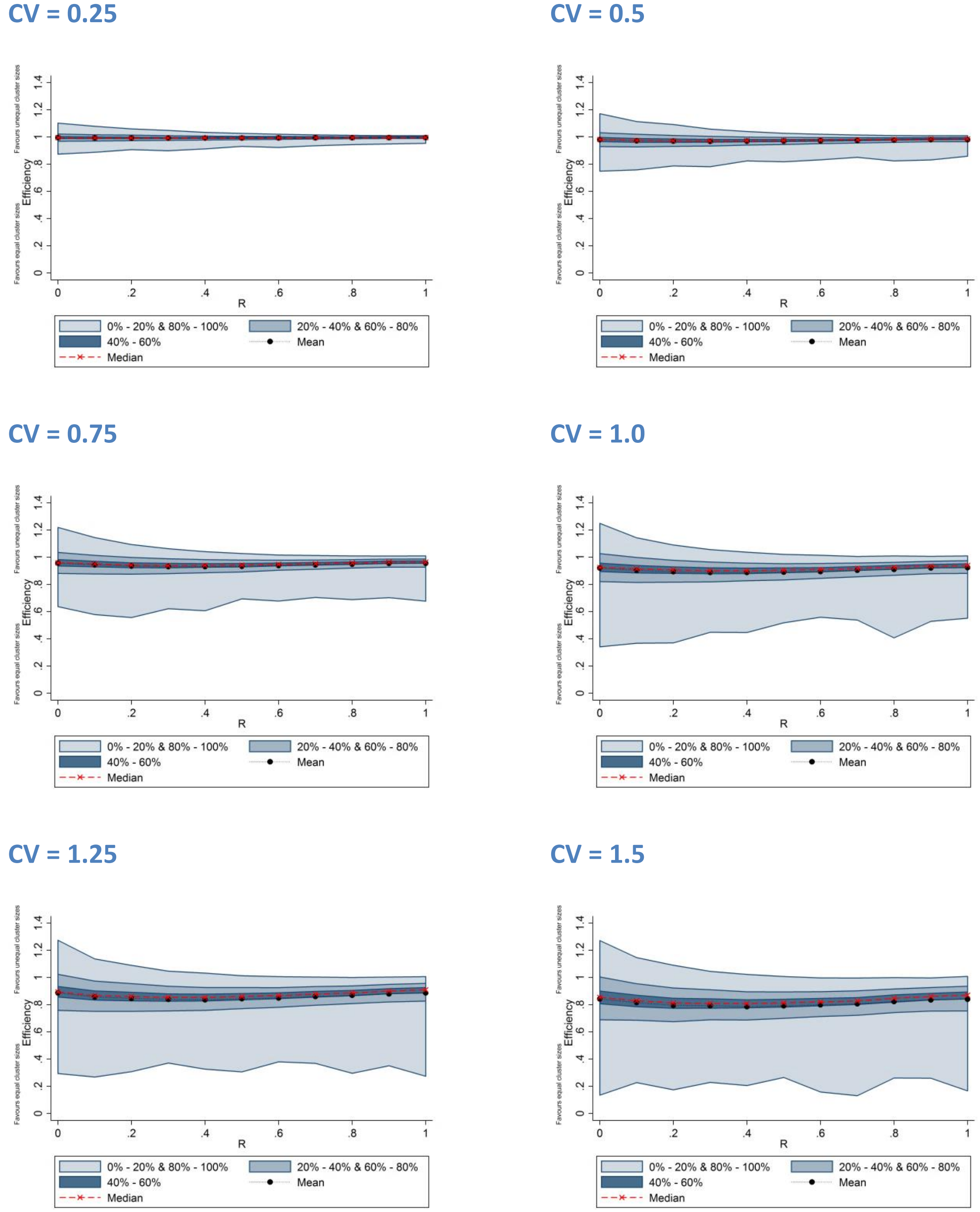


CV = 1.5



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 12 clusters and 6 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

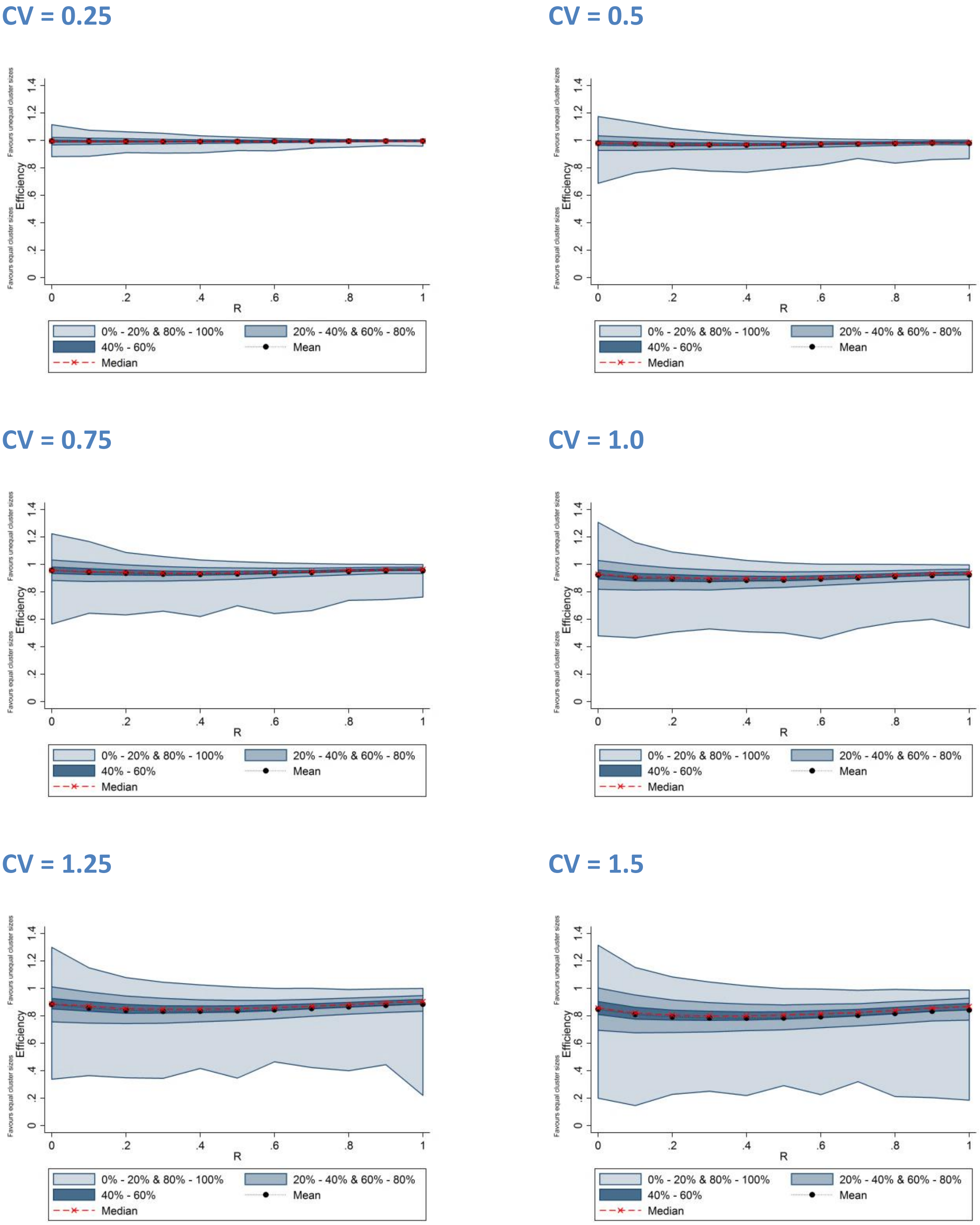
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 12 clusters and 12 randomisation steps. Cluster size = 72.


CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

CV = 1.25

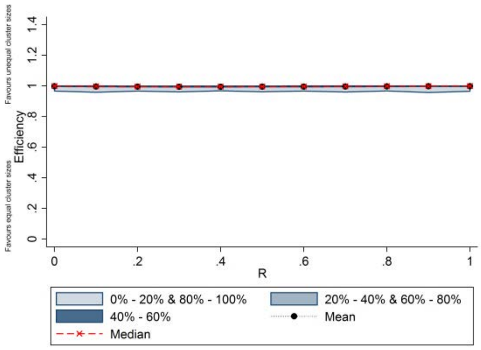
CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

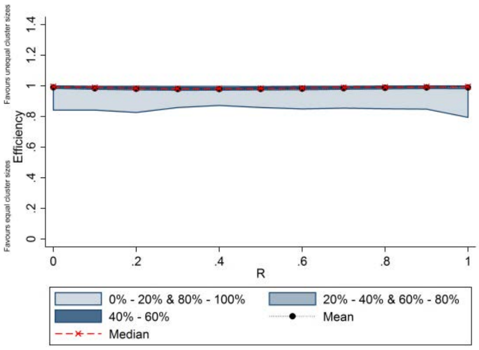
Efficiency of a SW-CRT with 24 clusters

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 2 randomisation steps. Cluster size = 72.

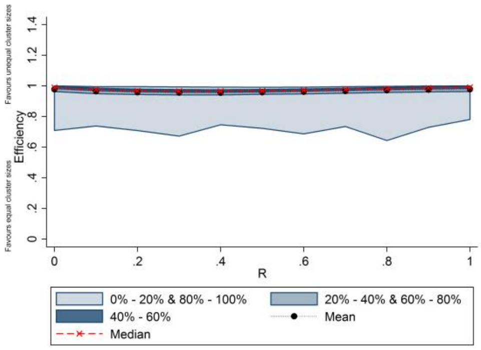
CV = 0.25



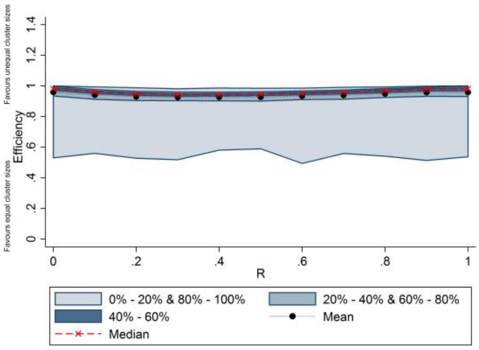
CV = 0.5



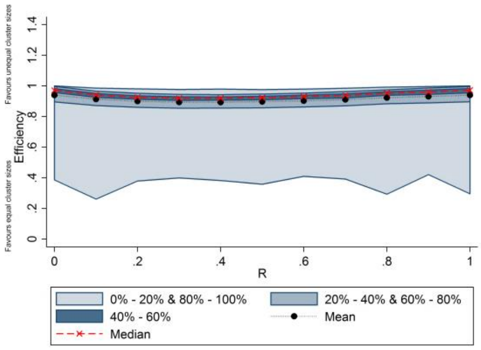
CV = 0.75



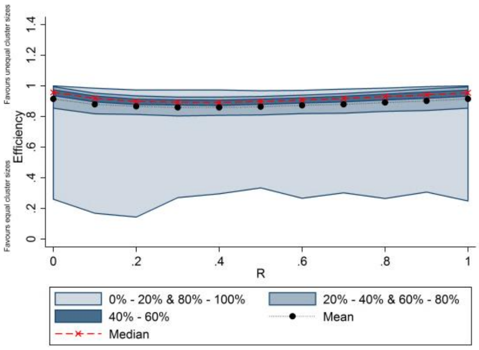
CV = 1.0



CV = 1.25

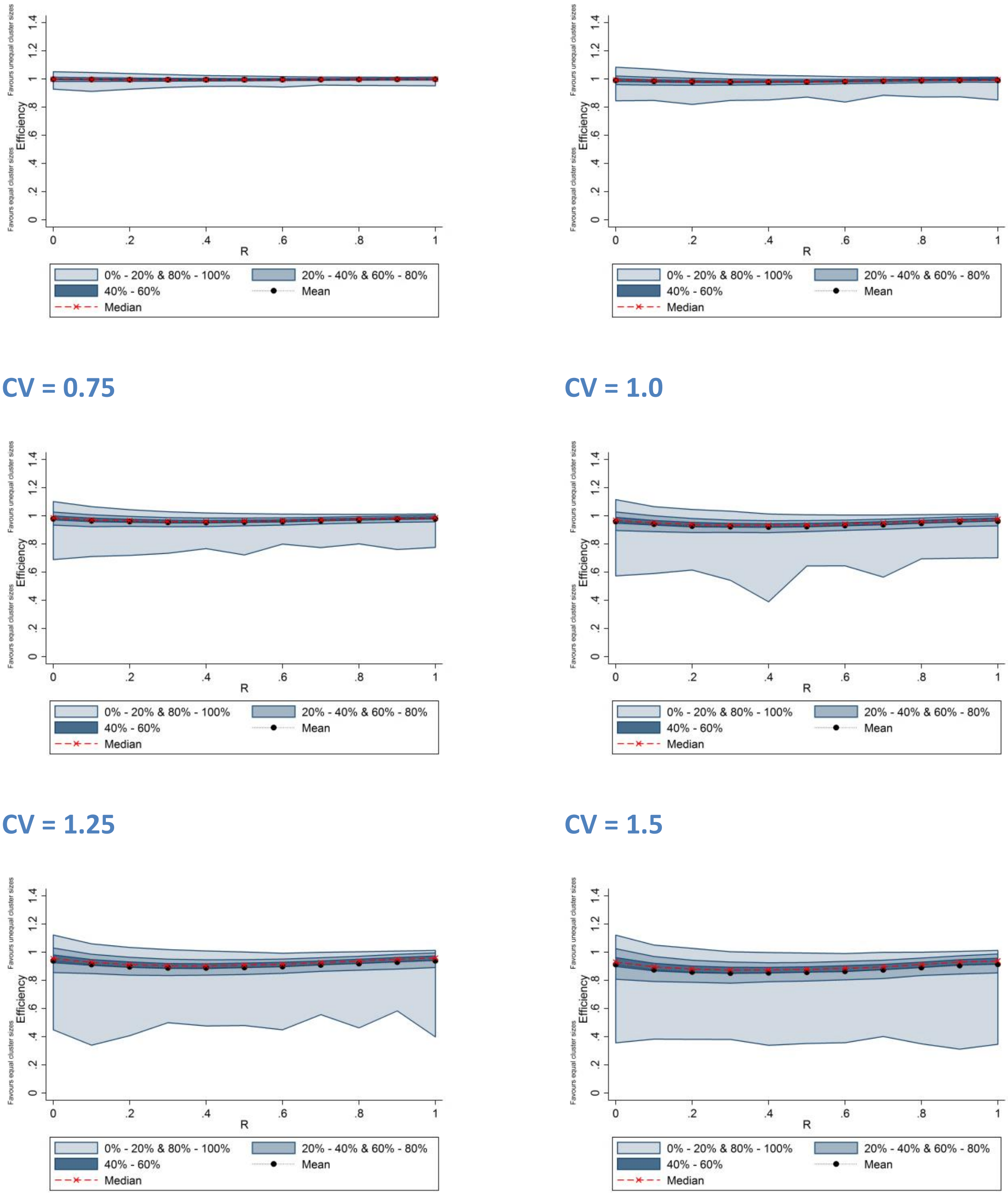


CV = 1.5



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 3 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

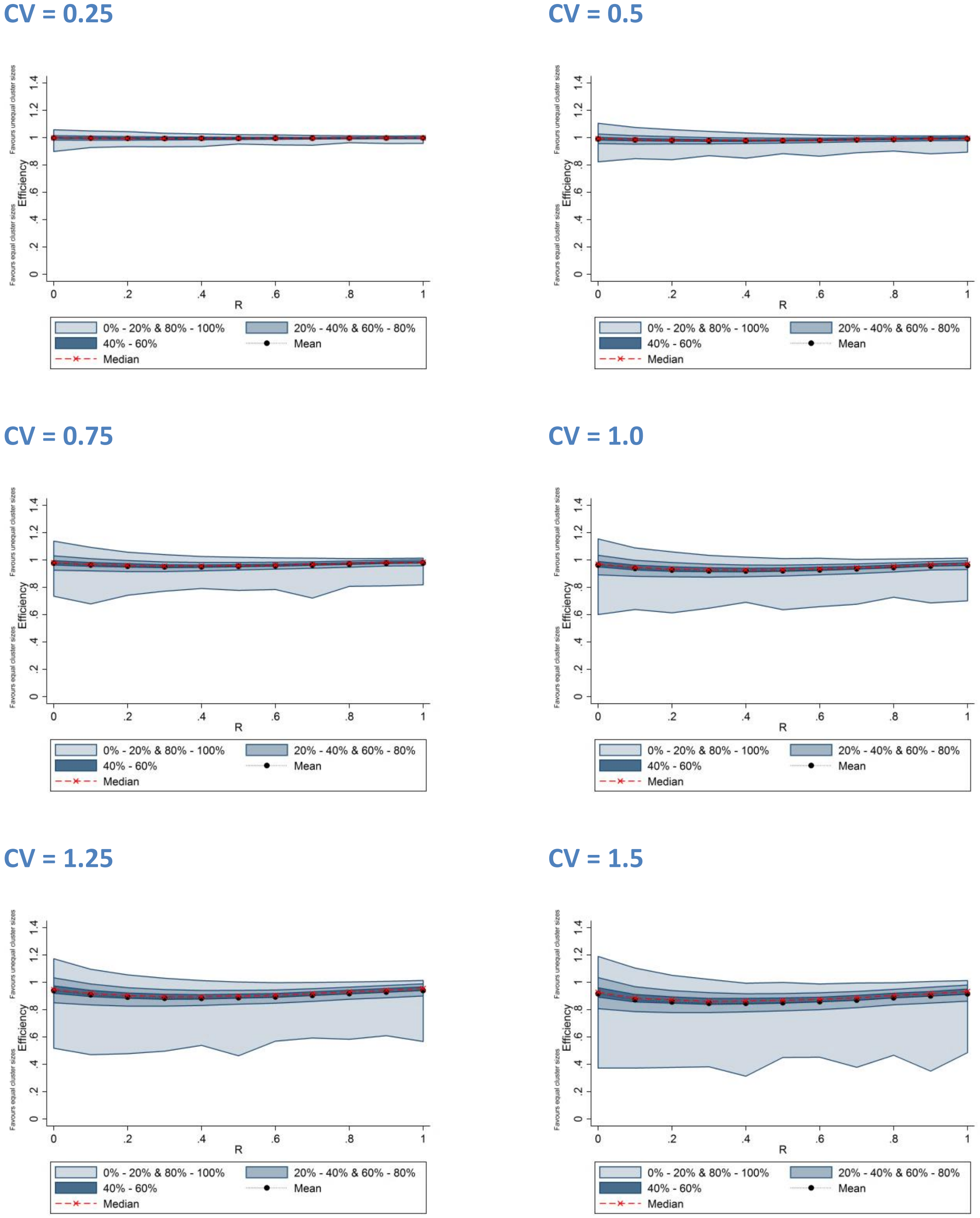
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 4 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

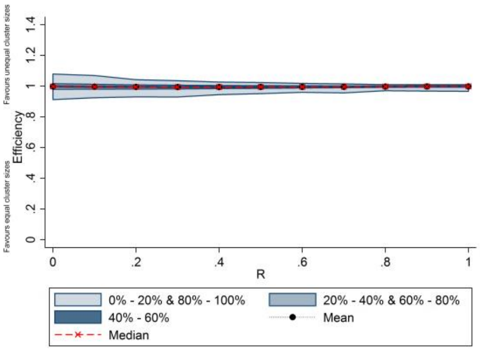
CV = 1.25

CV = 1.5

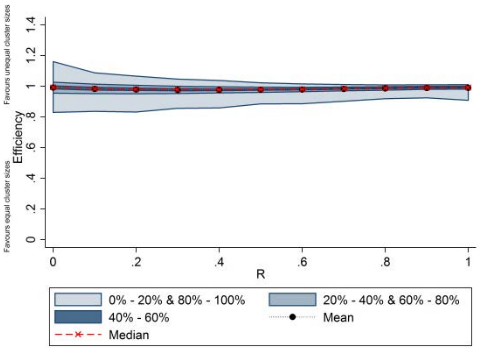
Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 6 randomisation steps. Cluster size = 72.

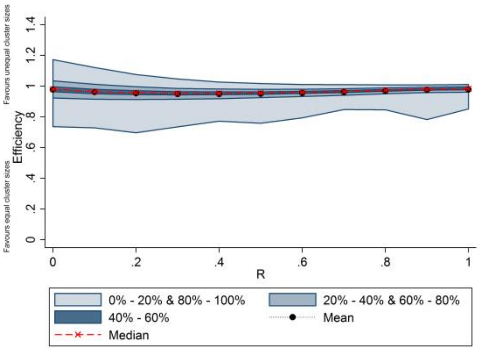
CV = 0.25



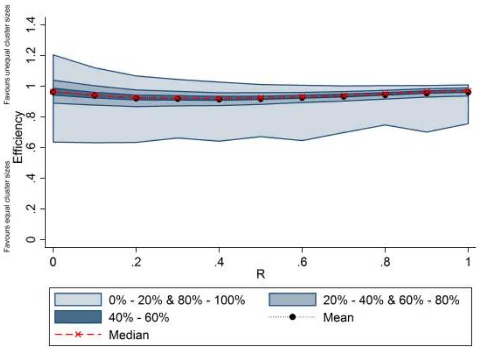
CV = 0.5



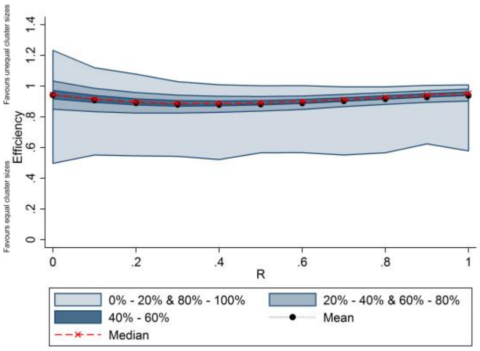
CV = 0.75



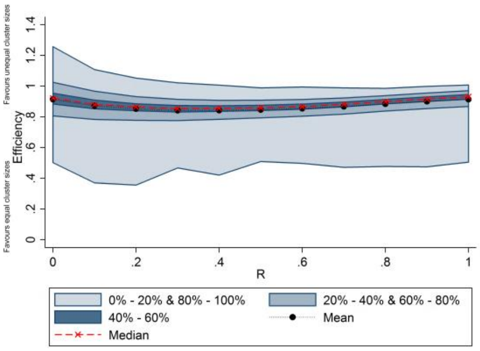
CV = 1.0



CV = 1.25

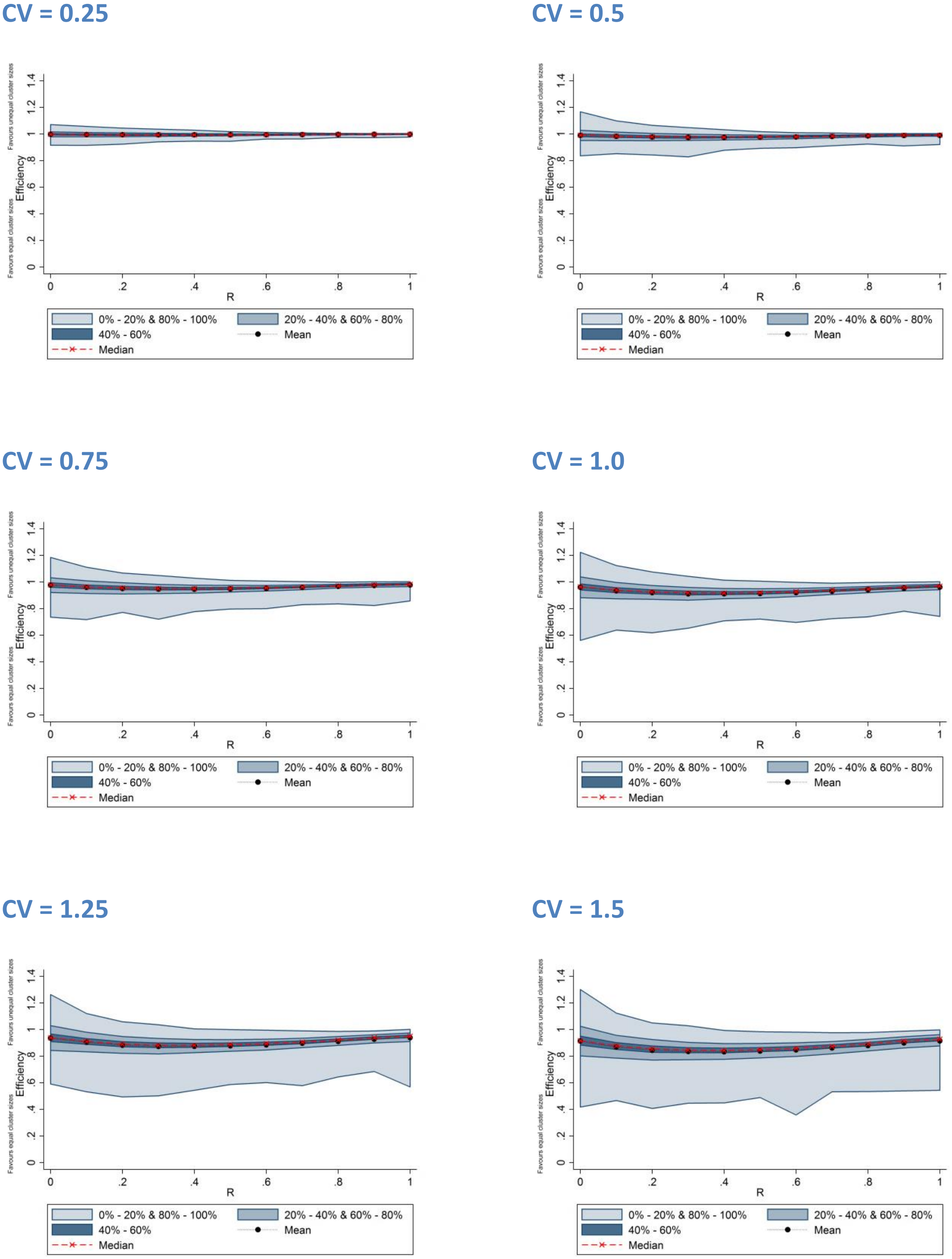


CV = 1.5



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 12 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

CV = 1.25

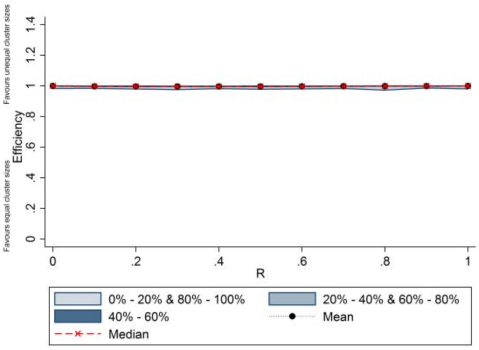
CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

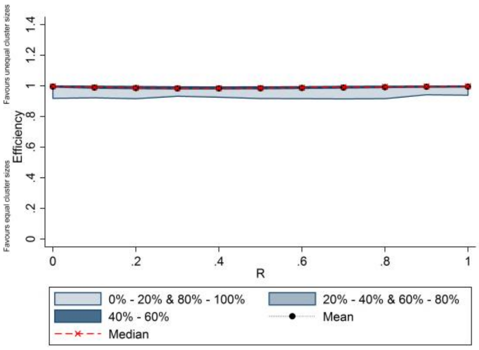
Efficiency of a SW-CRT with 48 clusters

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 48 clusters and 2 randomisation steps. Cluster size = 72.

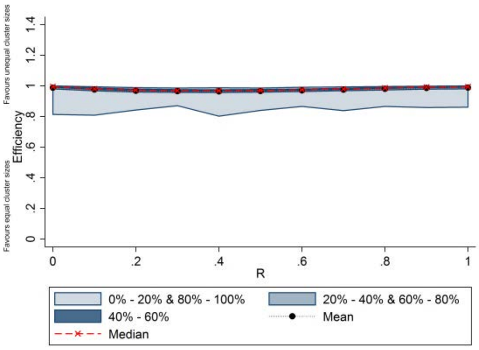
CV = 0.25



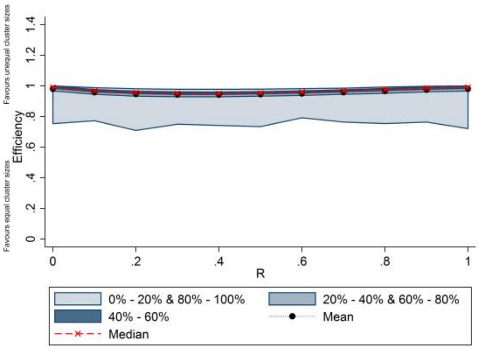
CV = 0.5



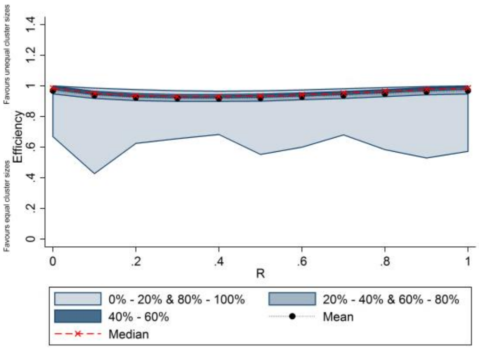
CV = 0.75



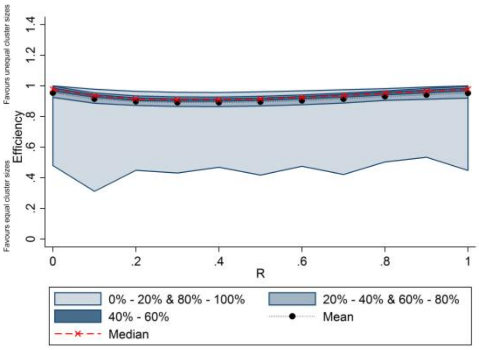
CV = 1.0



CV = 1.25

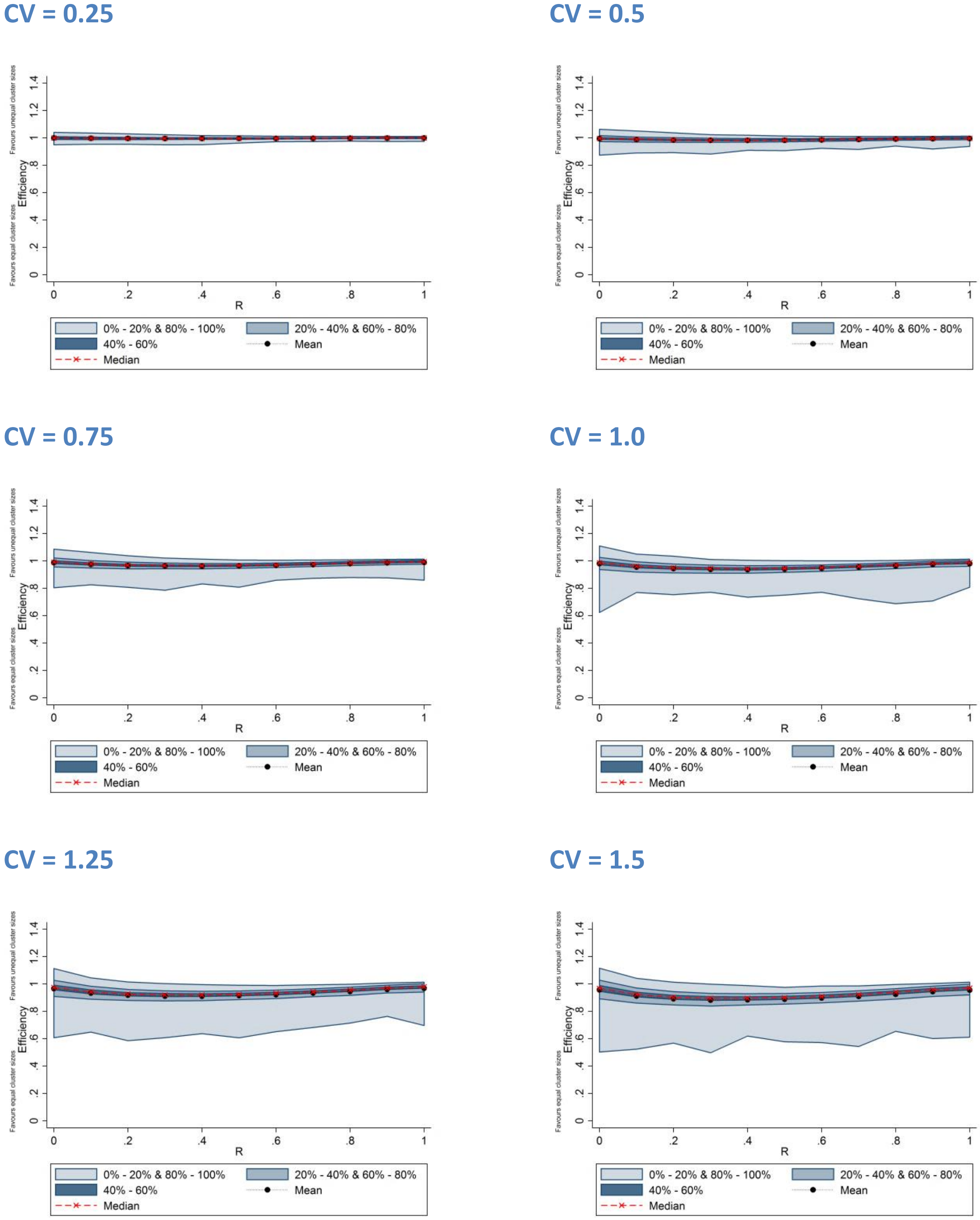


CV = 1.5



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 48 clusters and 3 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

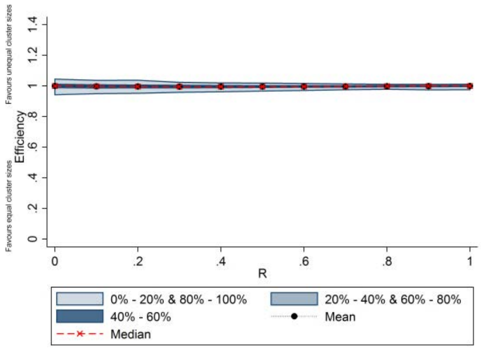
CV = 1.25

CV = 1.5

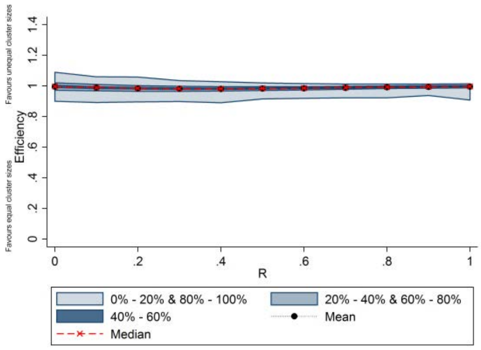
Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 48 clusters and 4 randomisation steps. Cluster size = 72.

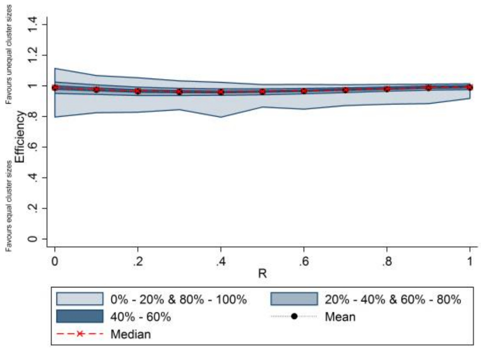
CV = 0.25



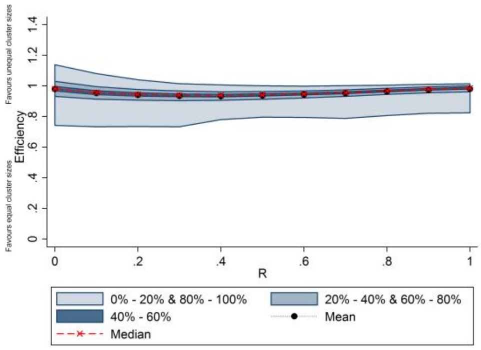
CV = 0.5



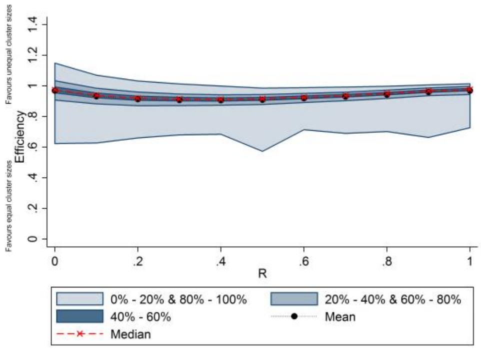
CV = 0.75



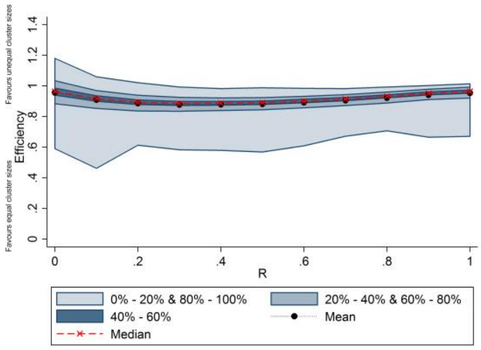
CV = 1.0



CV = 1.25

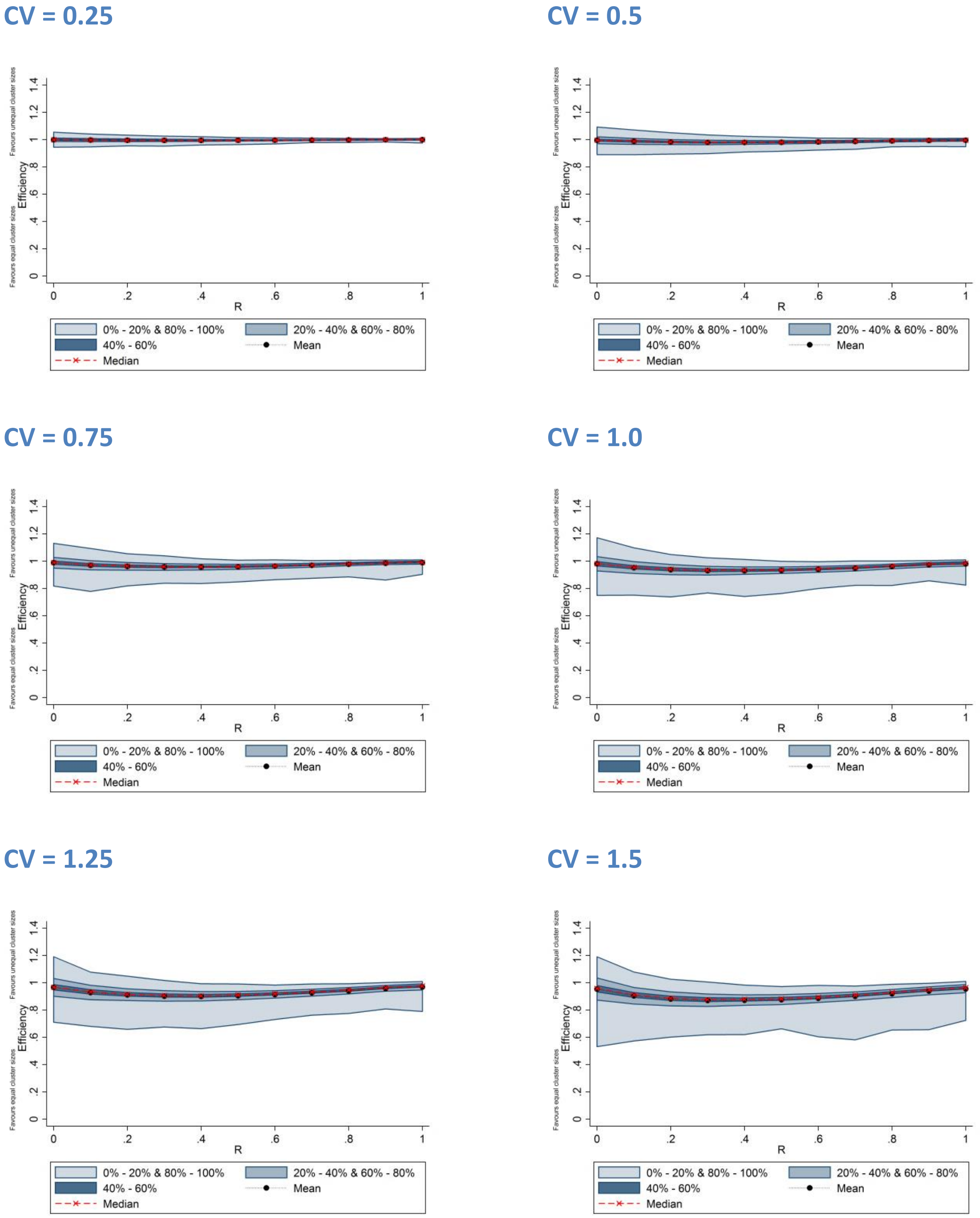


CV = 1.5



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 48 clusters and 6 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

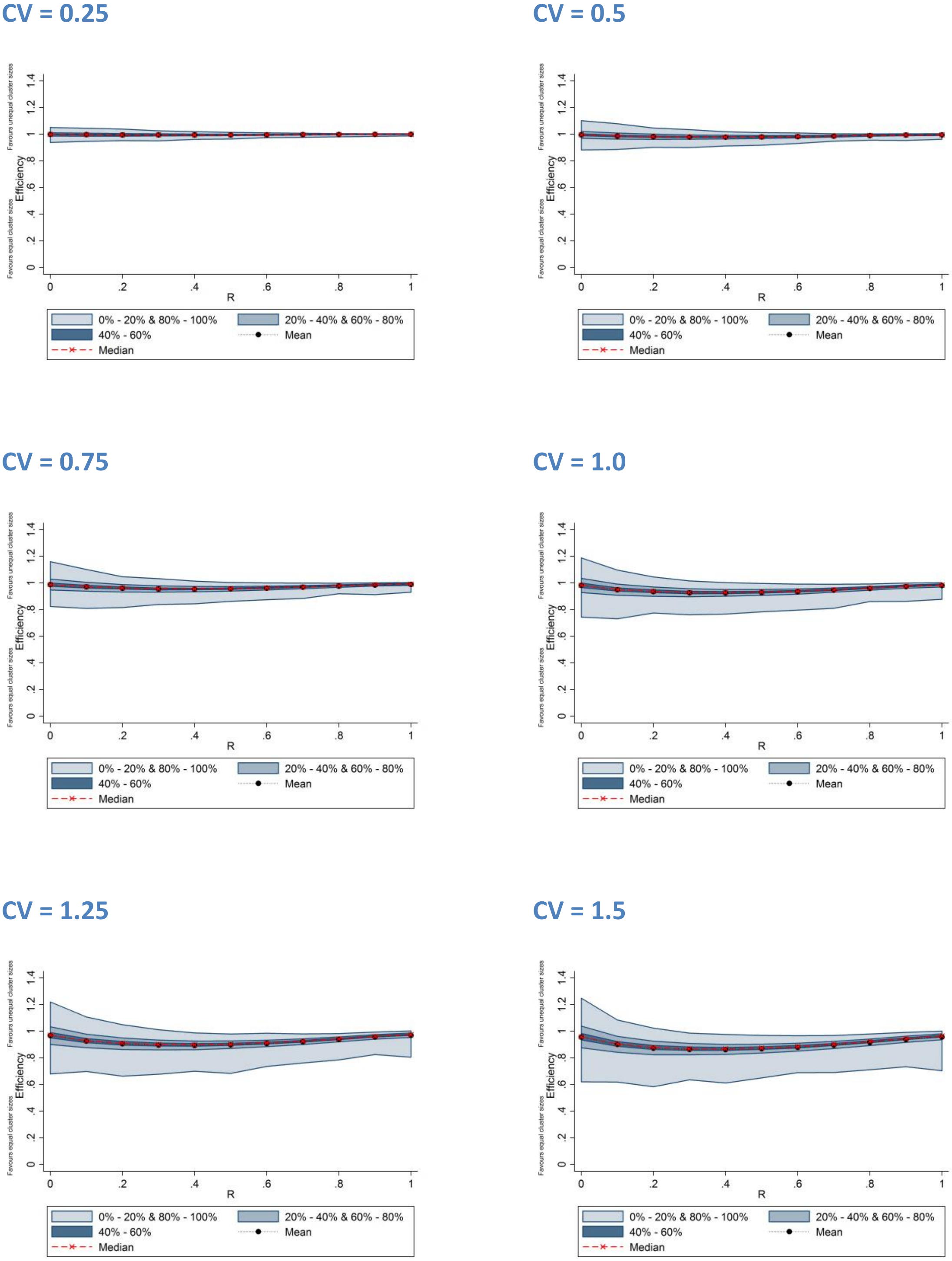
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 48 clusters and 12 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

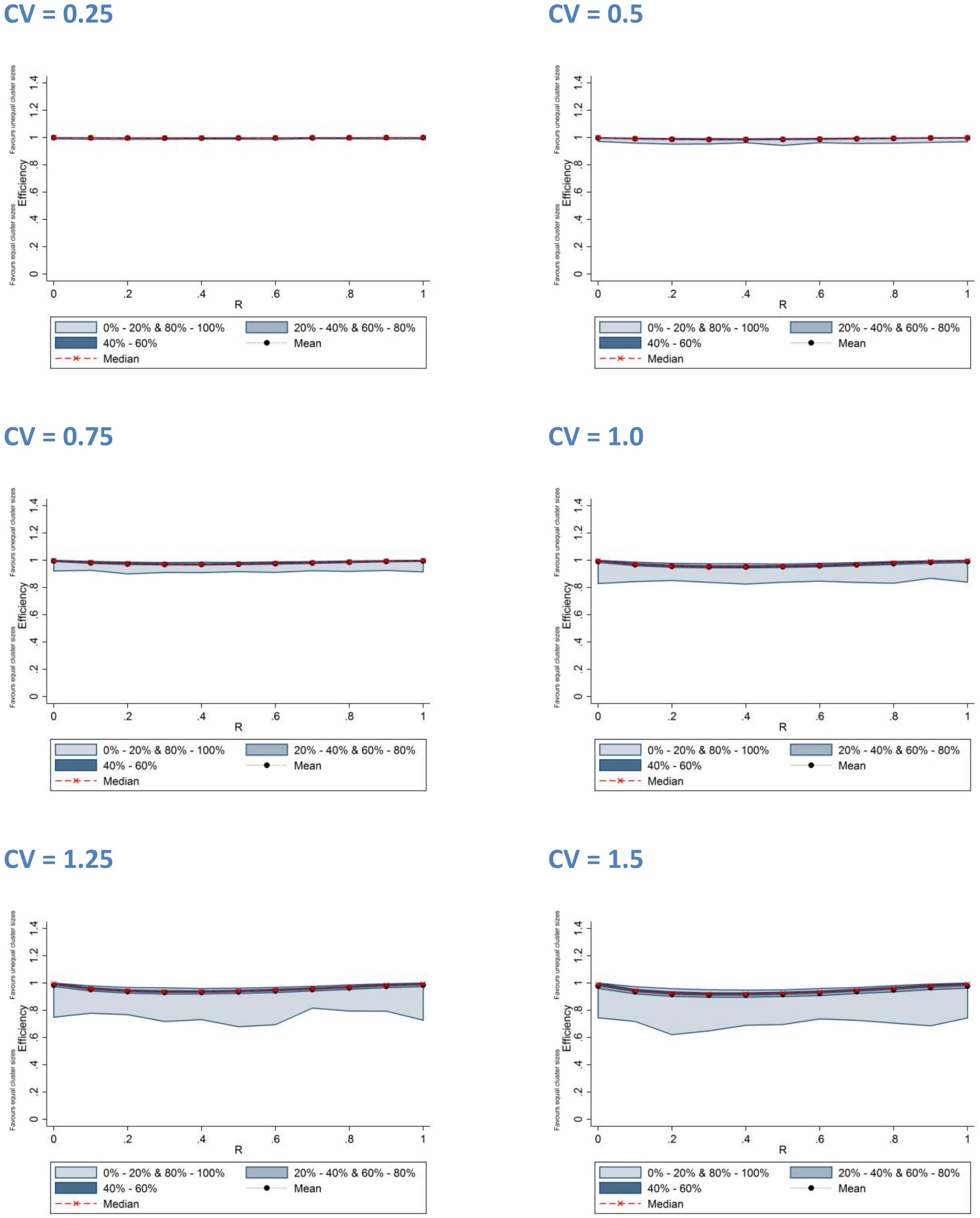
CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency of a SW-CRT with 96 clusters

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 96 clusters and 2 randomisation steps. Cluster size = 72.


CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

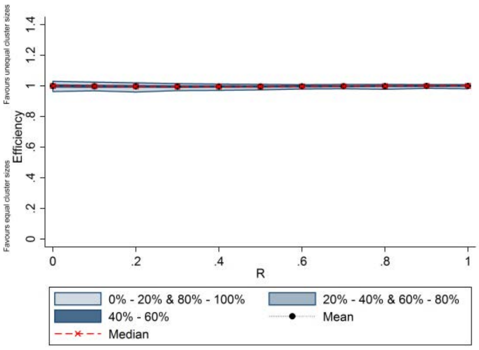
CV = 1.25

CV = 1.5

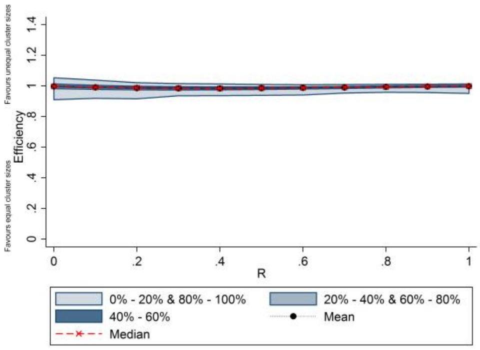
Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 96 clusters and 3 randomisation steps. Cluster size = 72.

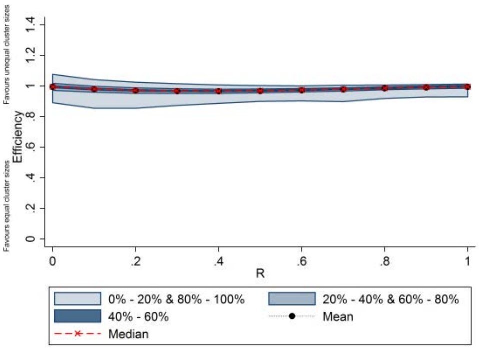
CV = 0.25



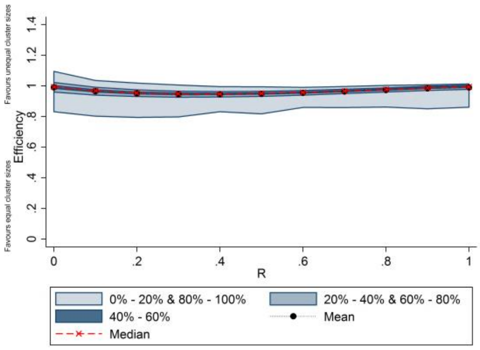
CV = 0.5



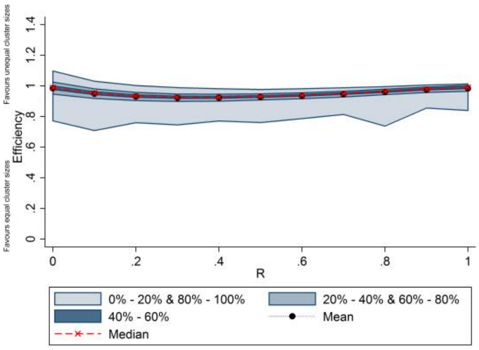
CV = 0.75



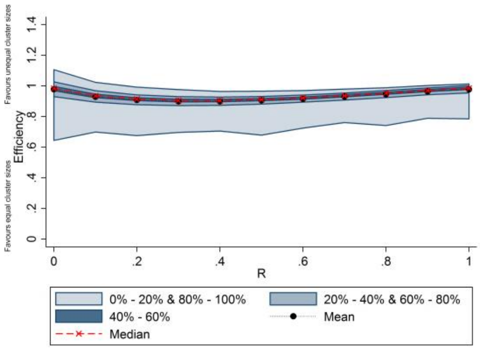
CV = 1.0



CV = 1.25

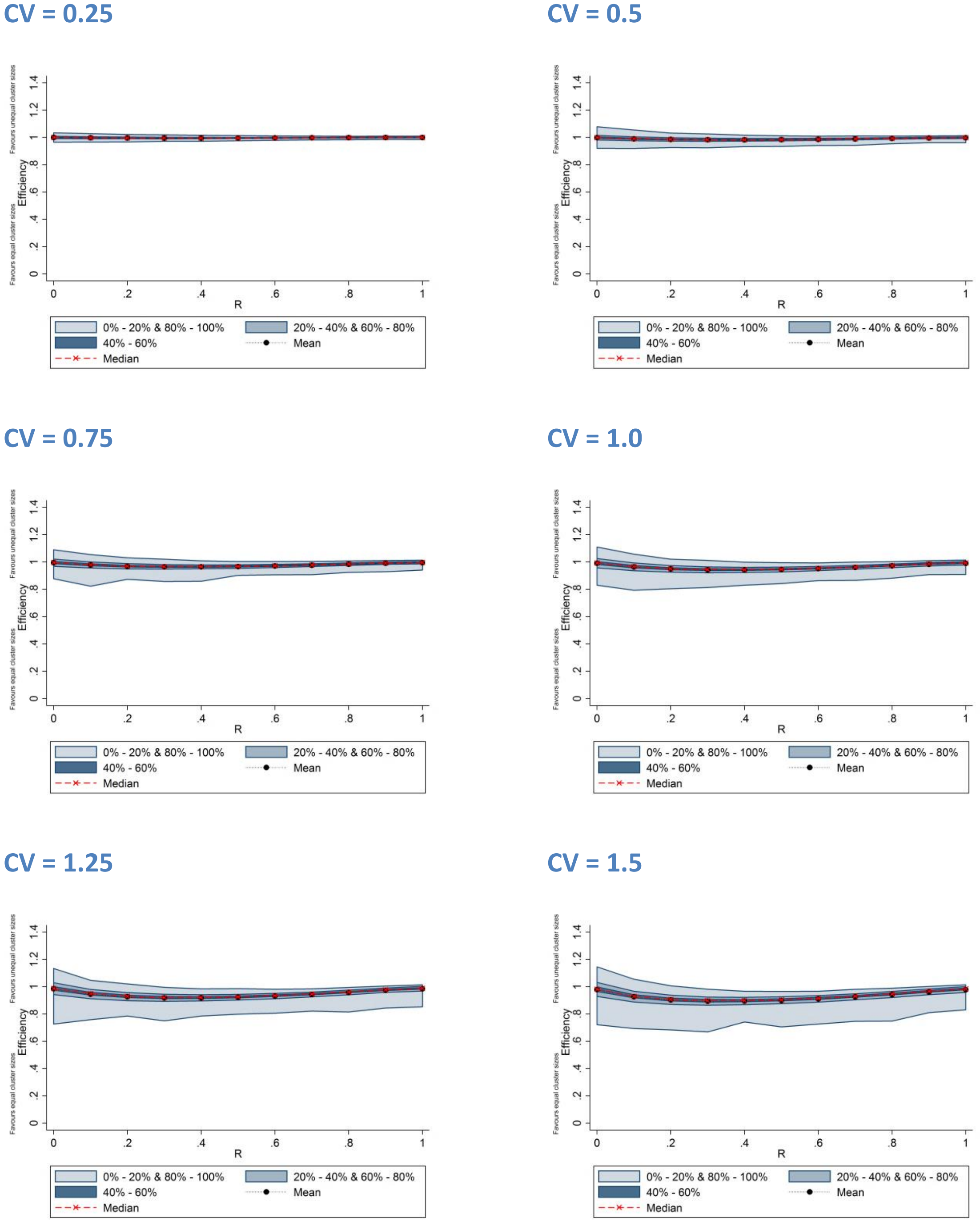


CV = 1.5



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 96 clusters and 4 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

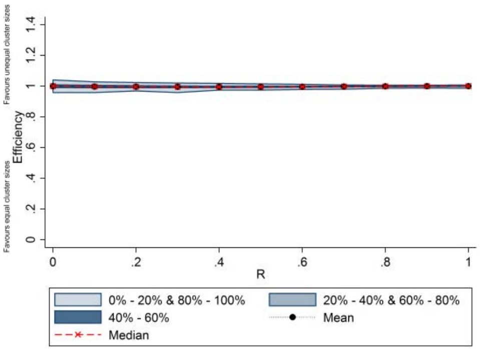
CV = 1.25

CV = 1.5

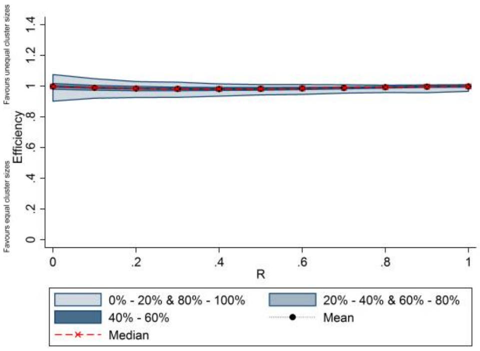
Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 96 clusters and 6 randomisation steps. Cluster size = 72.

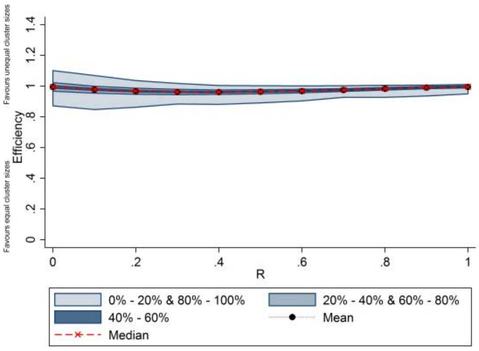
CV = 0.25



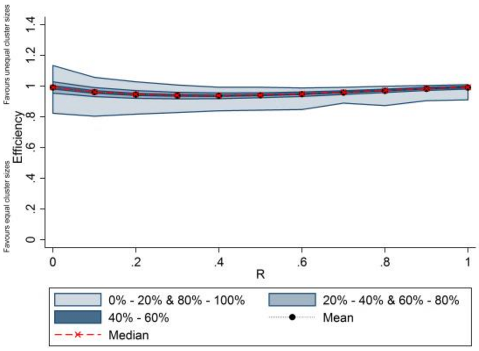
CV = 0.5



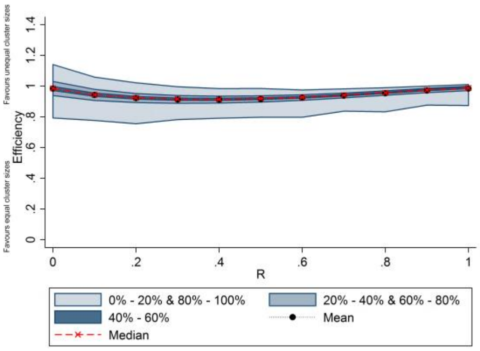
CV = 0.75



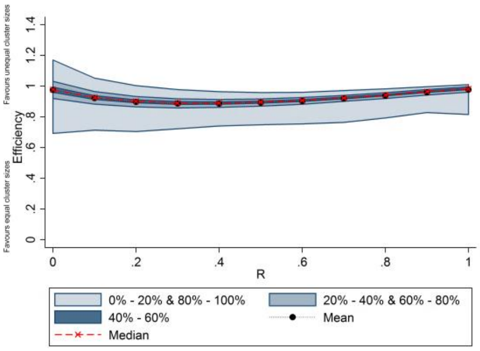
CV = 1.0



CV = 1.25

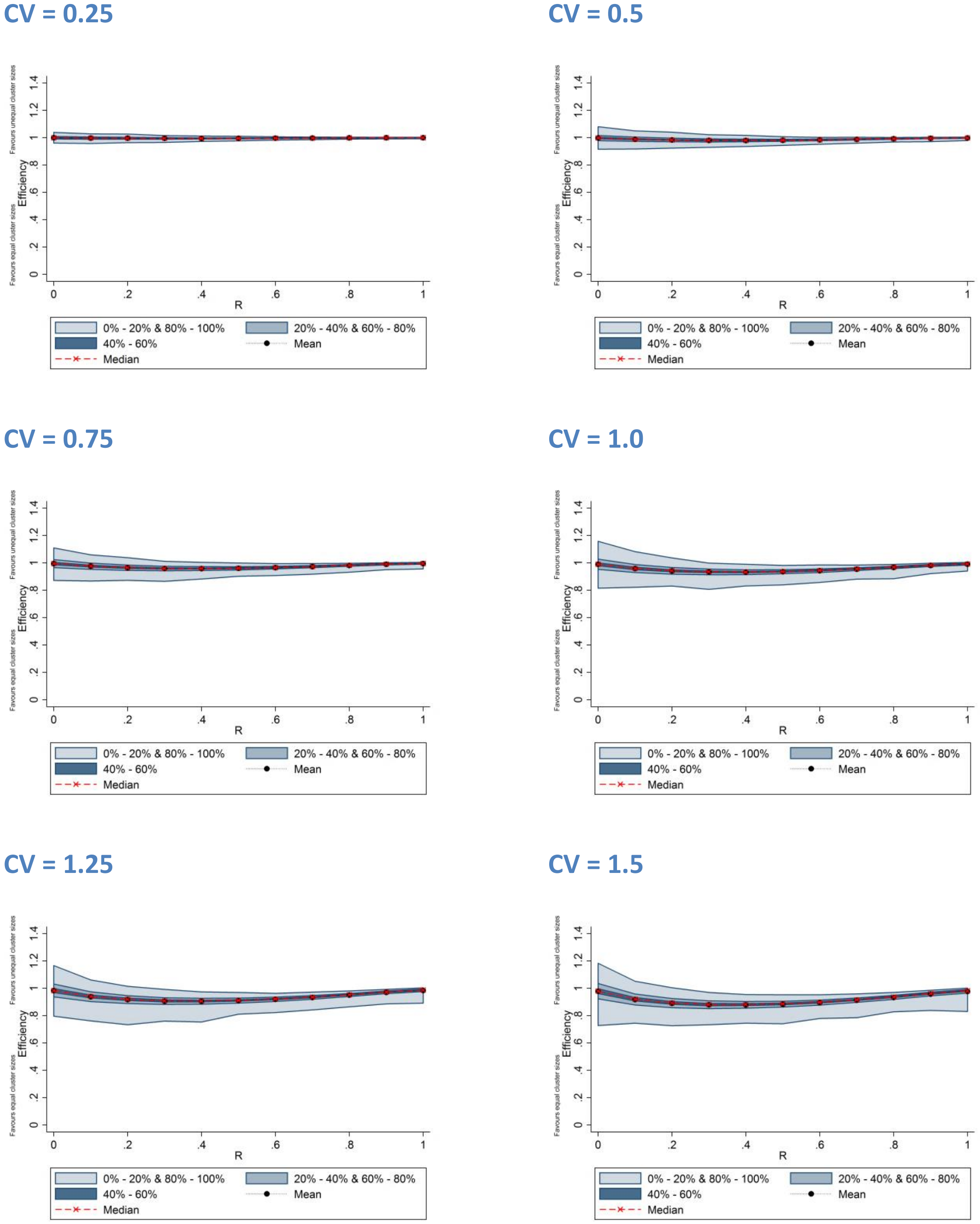


CV = 1.5



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 96 clusters and 12 randomisation steps. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

CV = 1.25

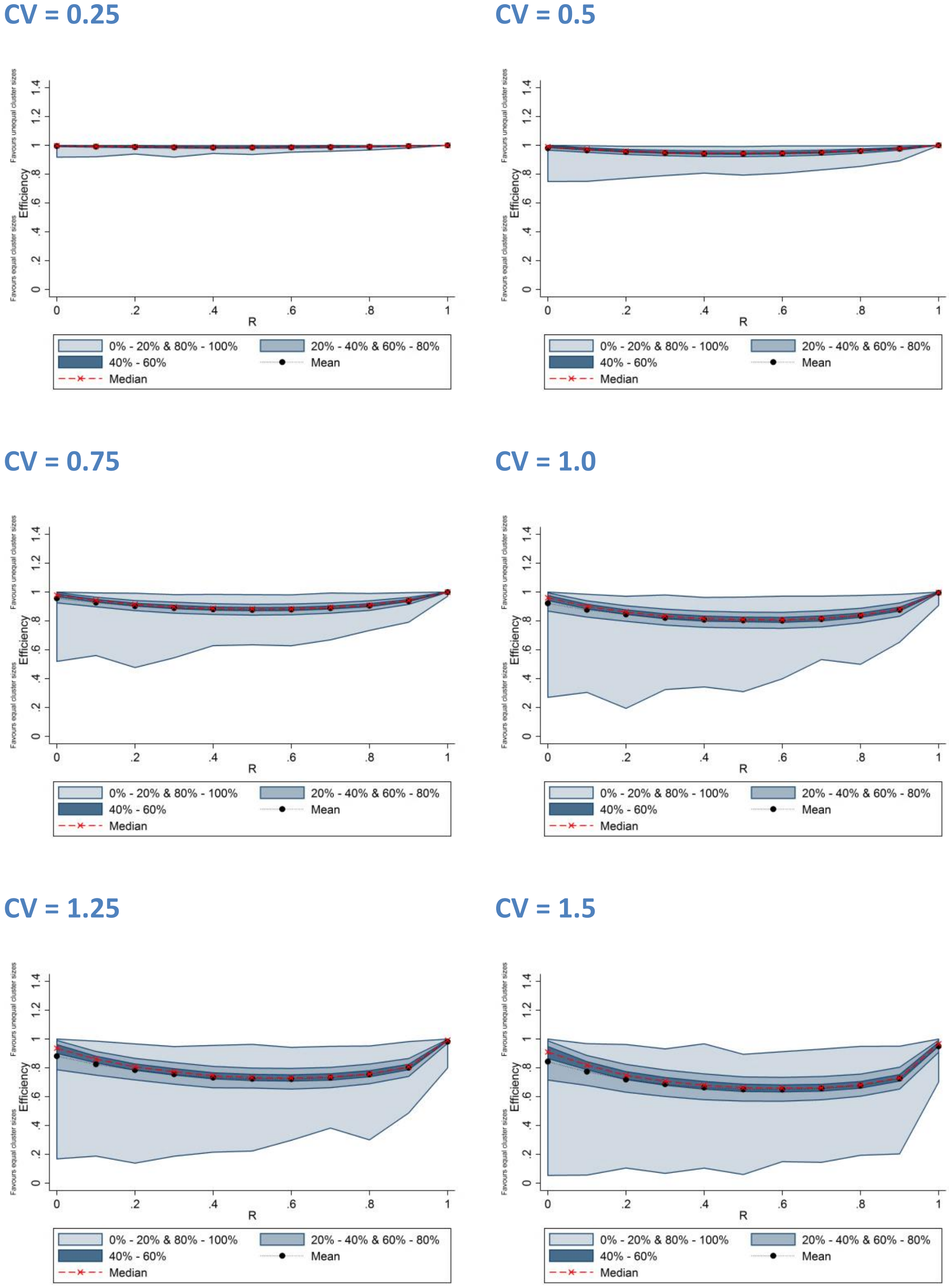
CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Appendix L

Below, we present the efficiency curves for the simulation study to assess the impact of between-cluster variation in a P-CRT. Results are presented grouped by the number of clusters in the P-CRT.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a P-CRT with 12 clusters. Cluster size = 72.


CV = 0.25

CV = 0.5

CV = 0.75

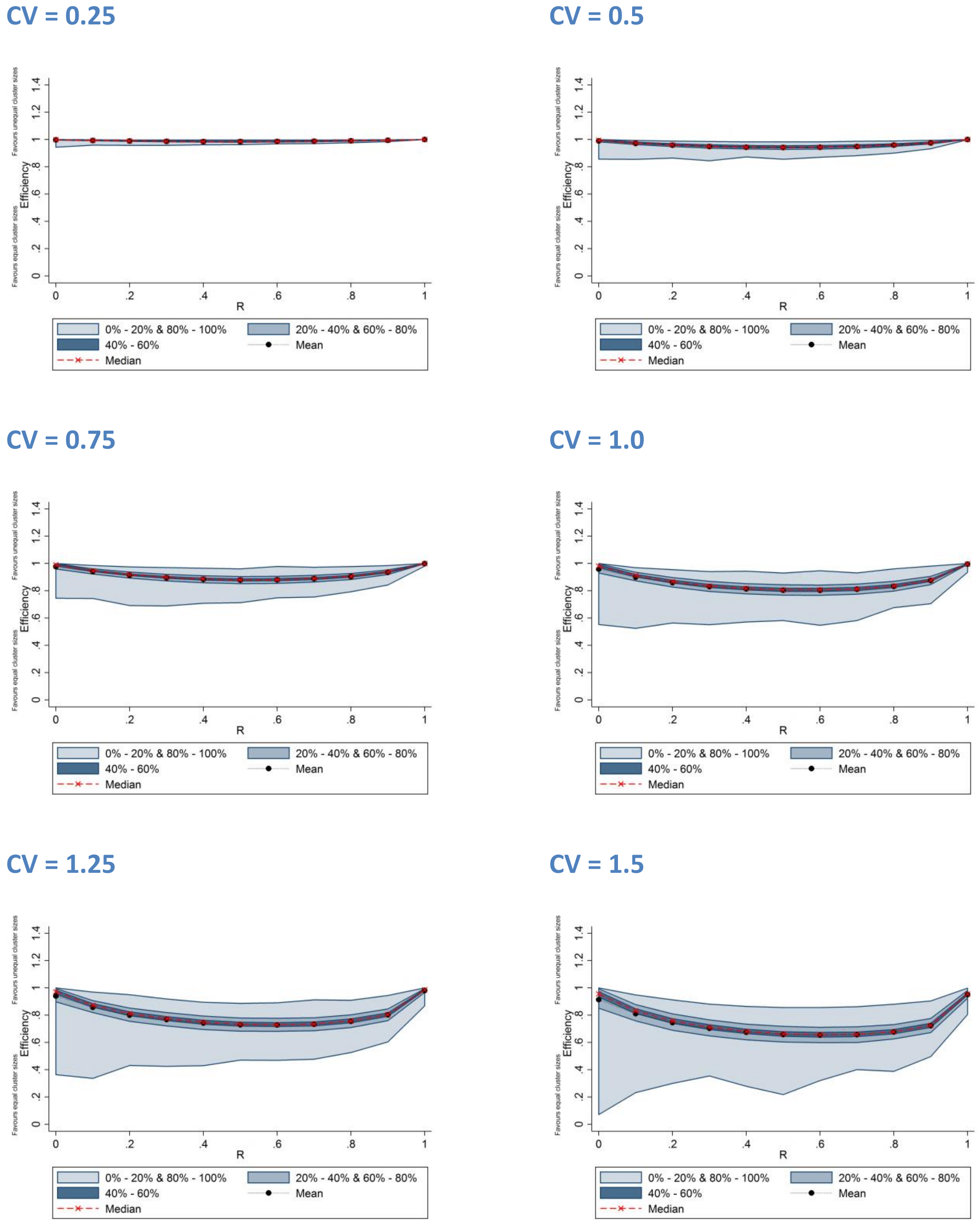
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a P-CRT with 24 clusters. Cluster size = 72.


CV = 0.25

CV = 0.5

CV = 0.75

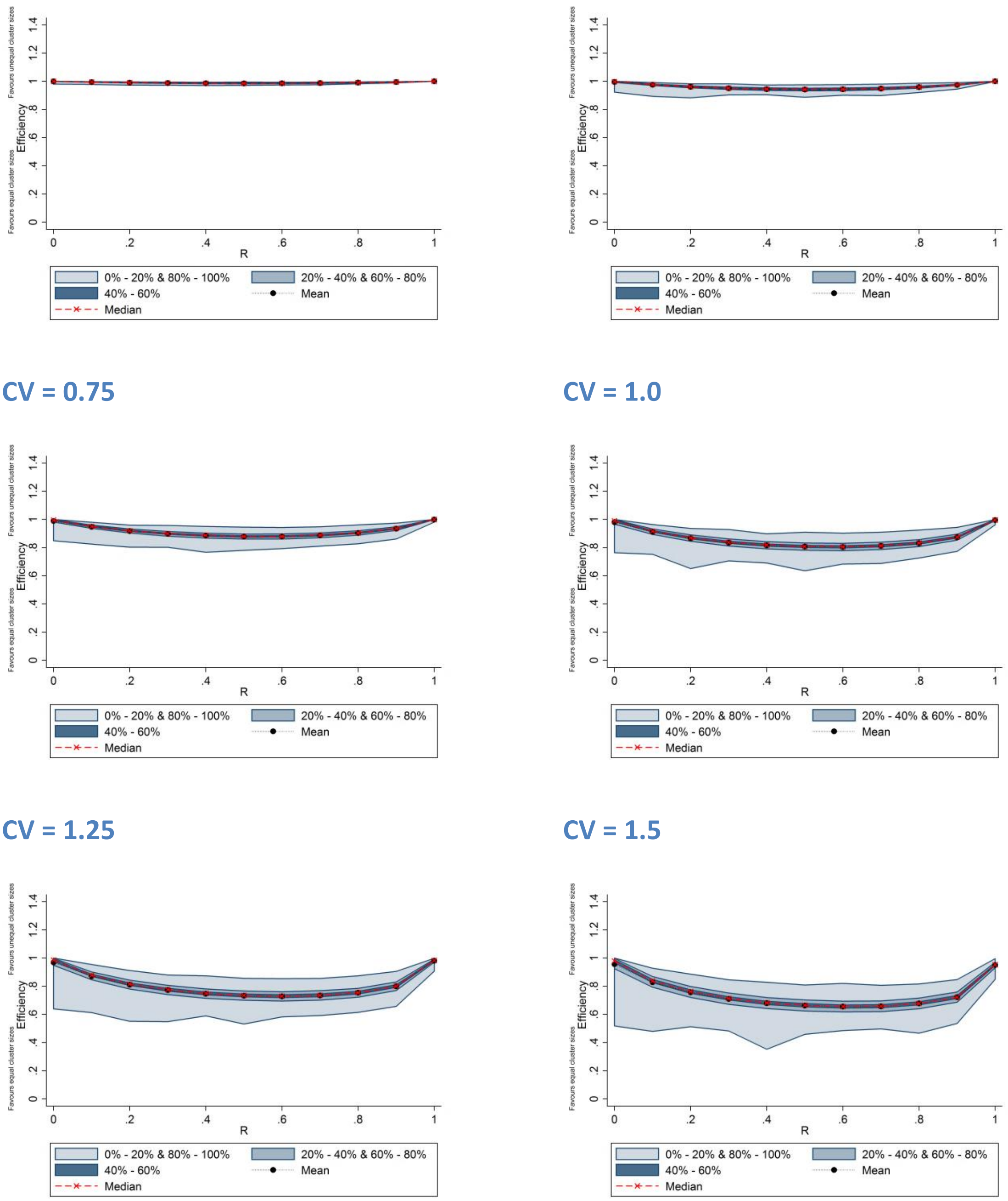
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a P-CRT with 48 clusters. Cluster size = 72.



CV = 0.25

CV = 0.5

CV = 0.75

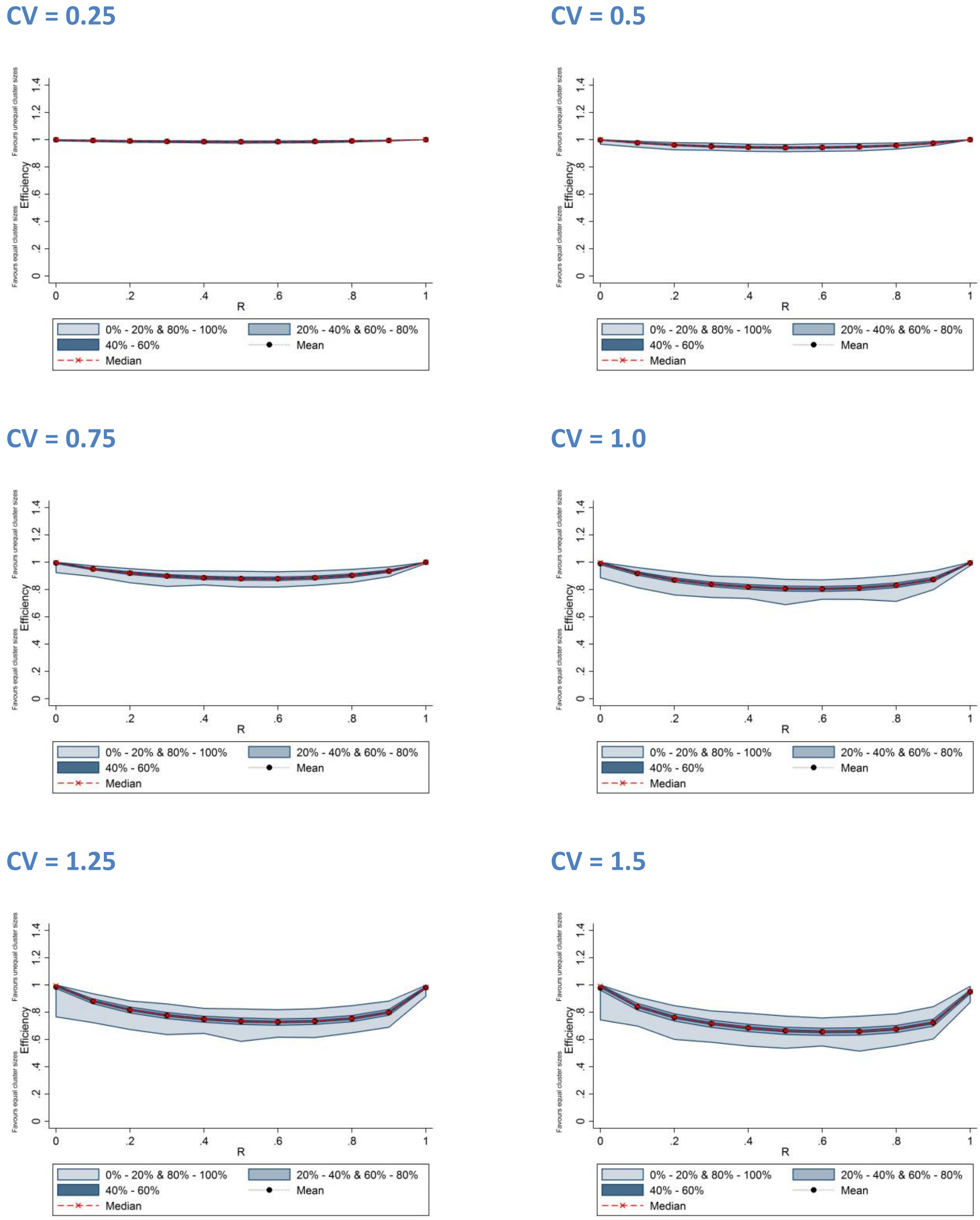
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a P-CRT with 96 clusters. Cluster size = 72.


CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

CV = 1.25

CV = 1.5

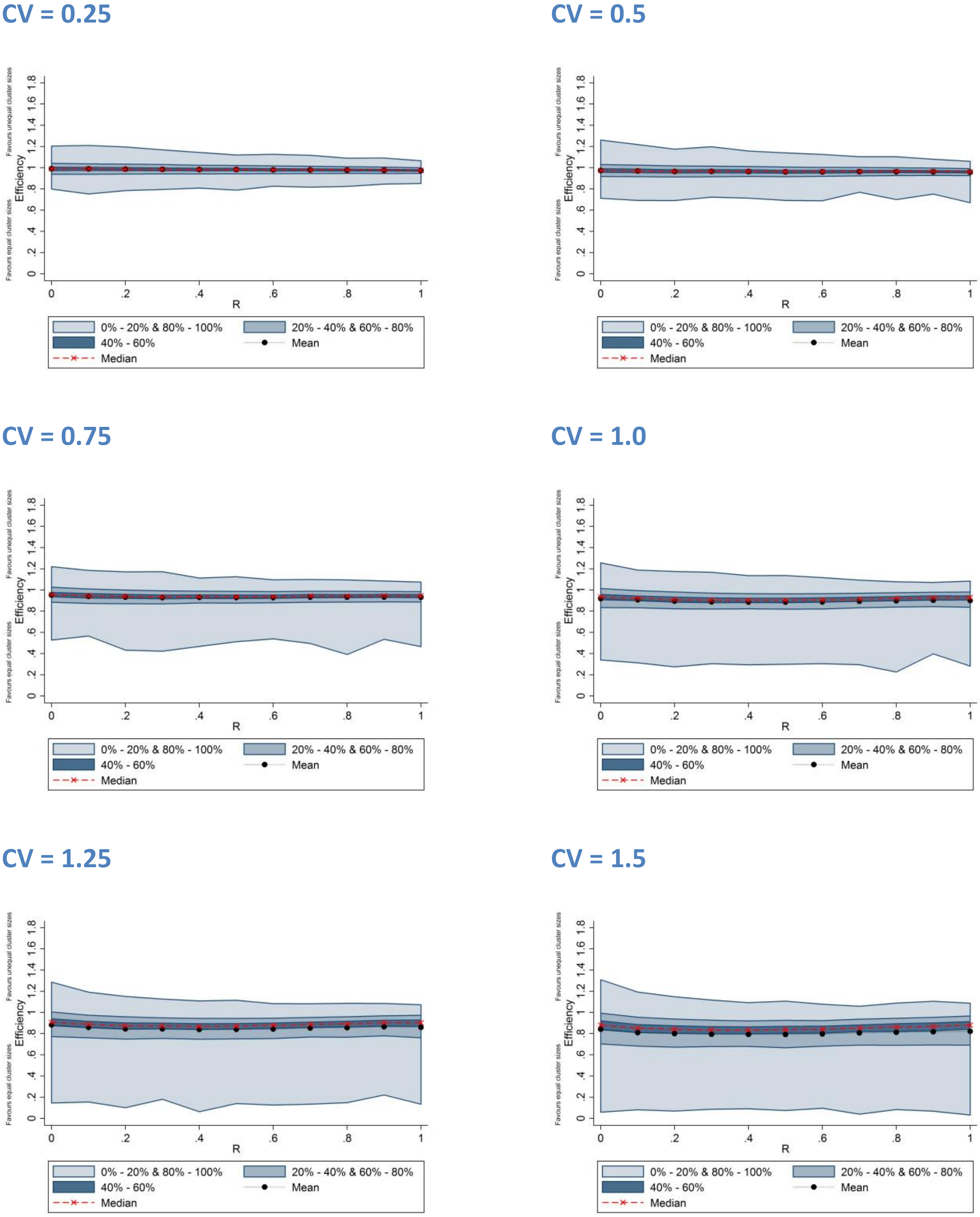
Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Appendix M

Below, we present the efficiency curves for the simulation study to assess the impact of within-cluster variation in a SW-CRT. Results are presented grouped by the within-cluster coefficient of variation.

Within-cluster coefficient of variation = 0.25

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 12 clusters and 2 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = 0.25.



CV = 0.25

CV = 0.5

CV = 0.75

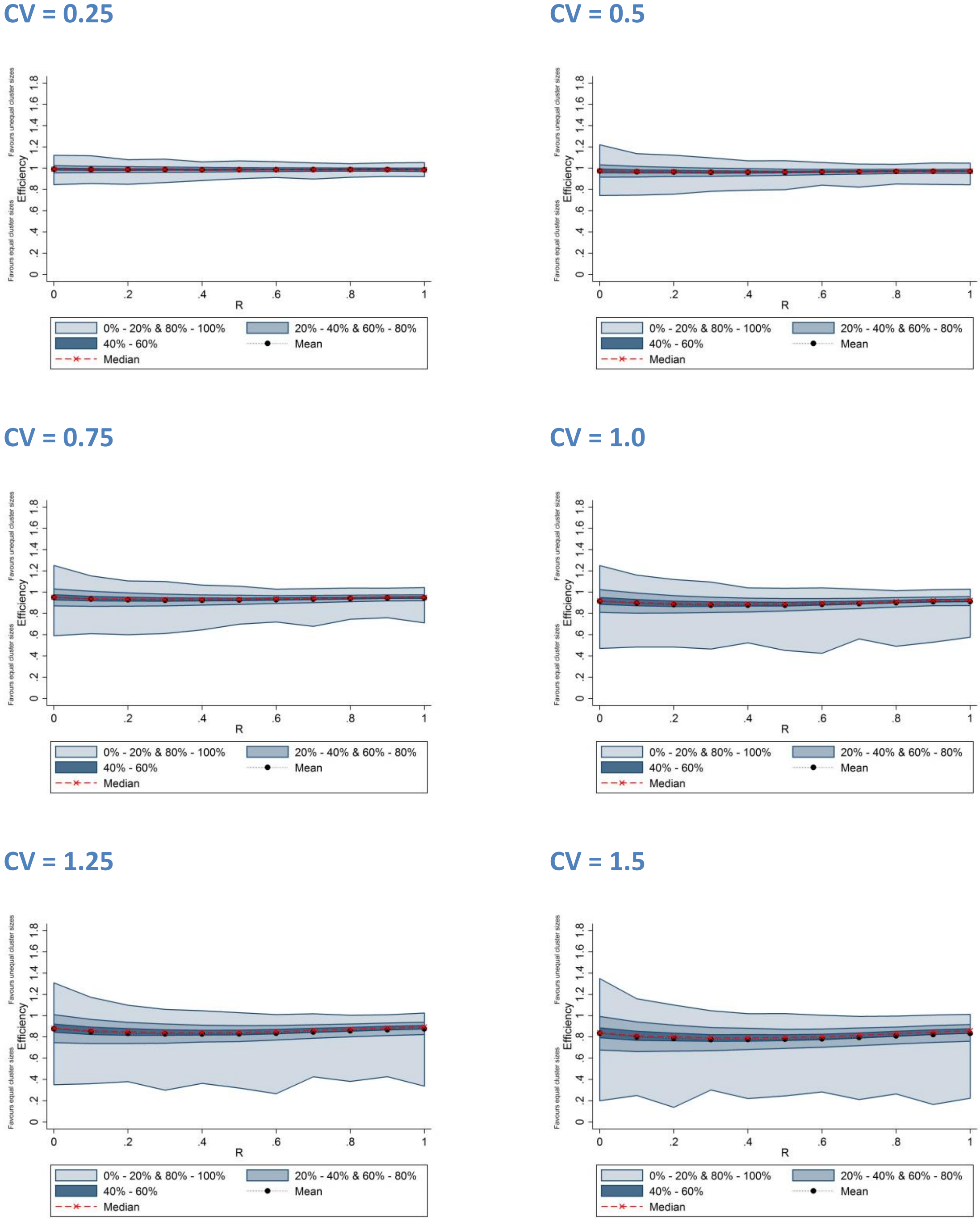
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 12 clusters and 12 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = 0.25.



CV = 0.25

CV = 0.5

CV = 0.75

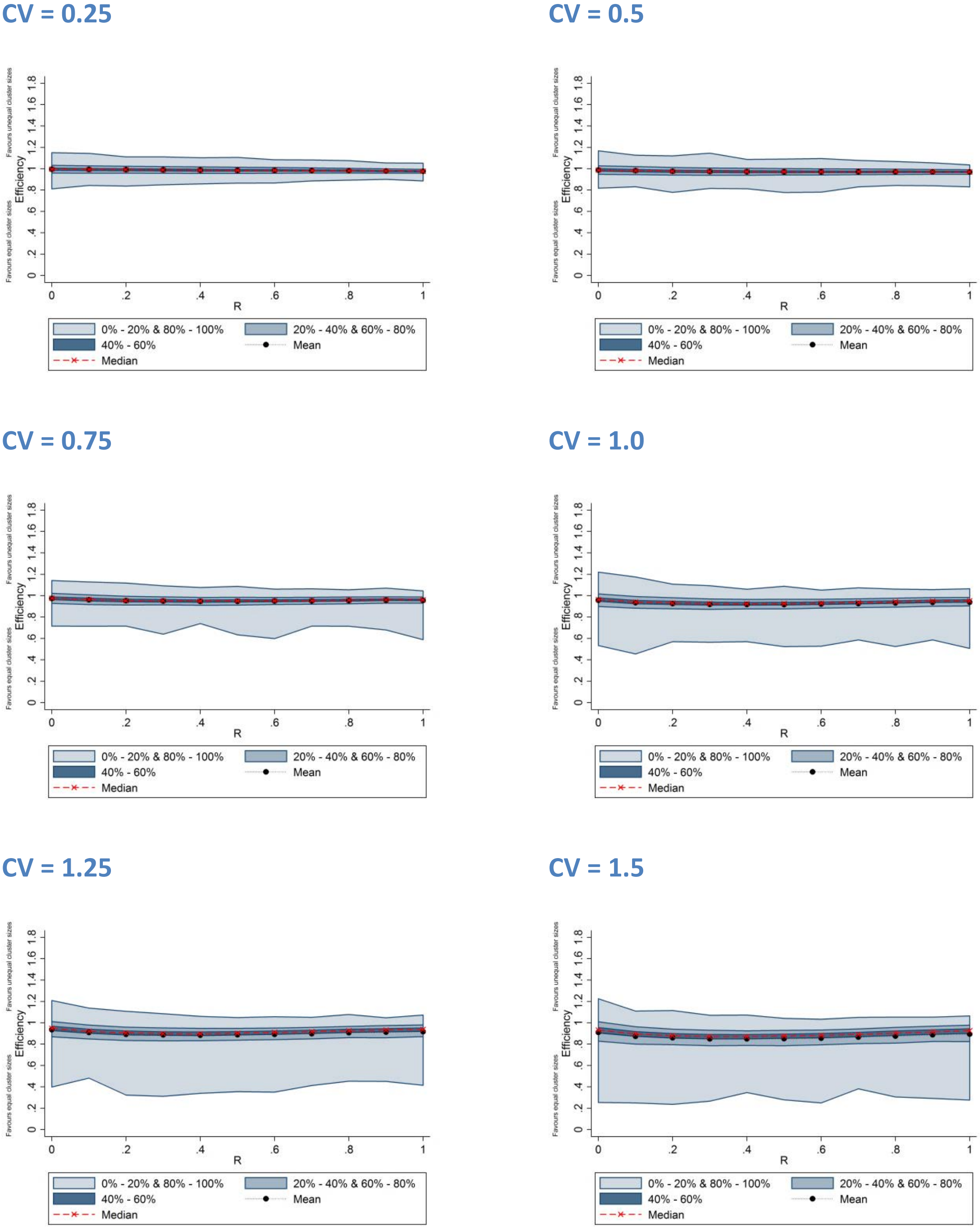
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 2 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = 0.25.



CV = 0.25

CV = 0.5

CV = 0.75

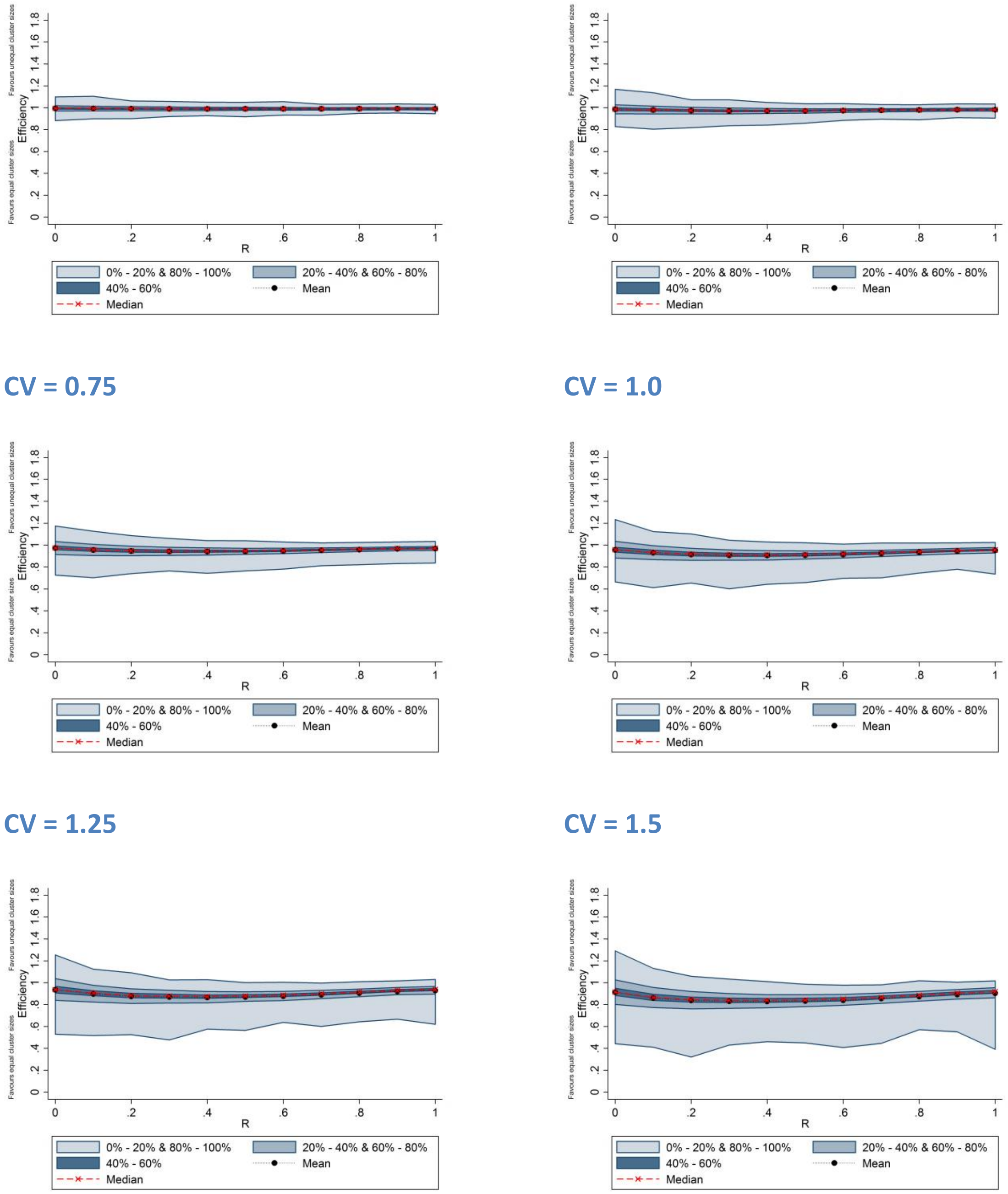
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 12 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = 0.25.


CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

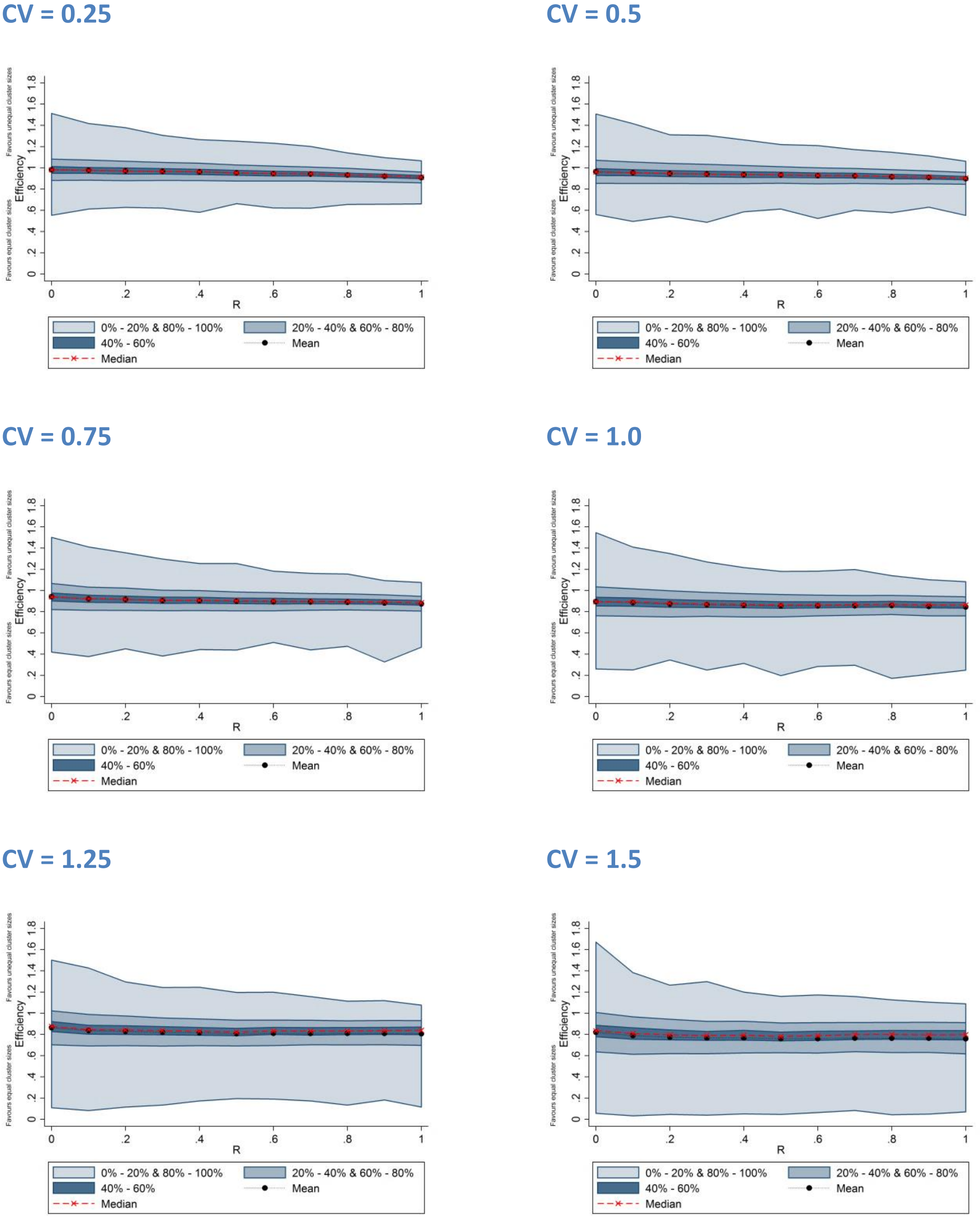
CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Within-cluster coefficient of variation = 0.5

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 12 clusters and 2 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = 0.5.



CV = 0.25

CV = 0.5

CV = 0.75

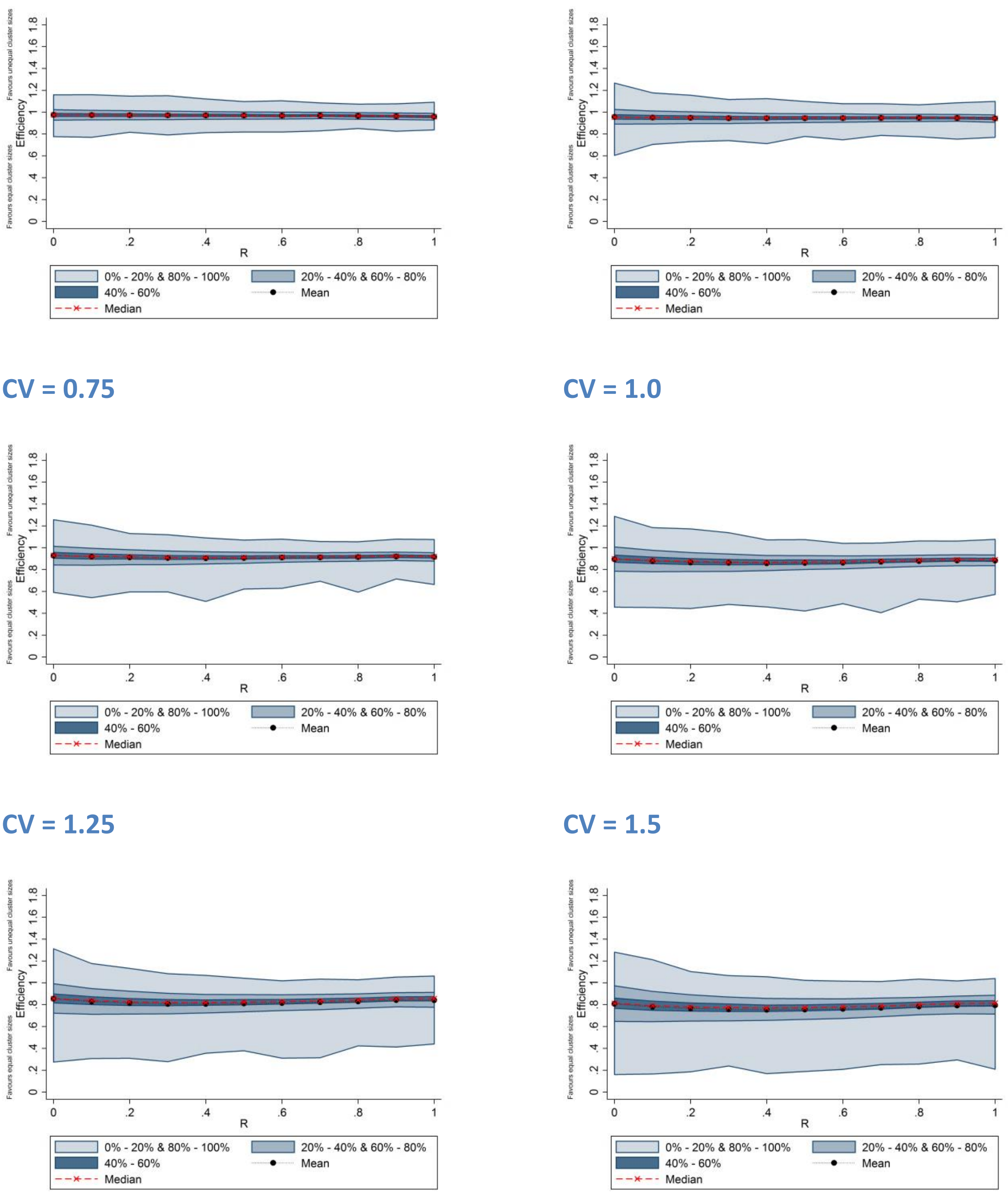
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 12 clusters and 12 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = 0.5.



CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

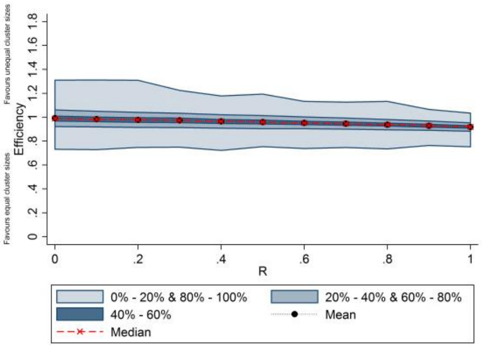
CV = 1.25

CV = 1.5

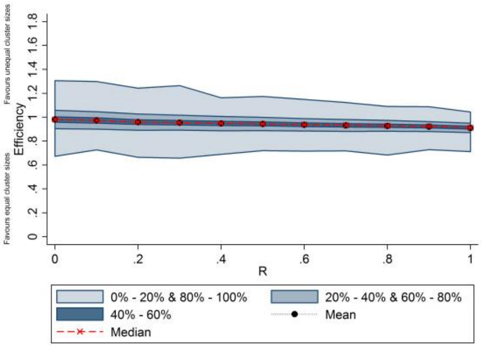
Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 2 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = 0.5.

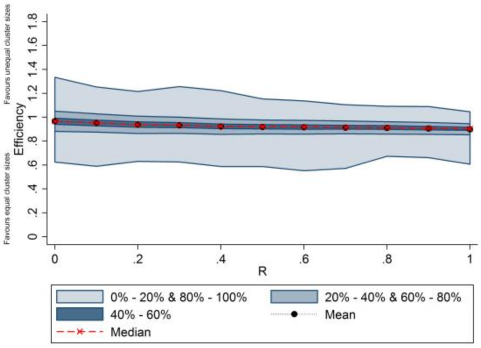
CV = 0.25



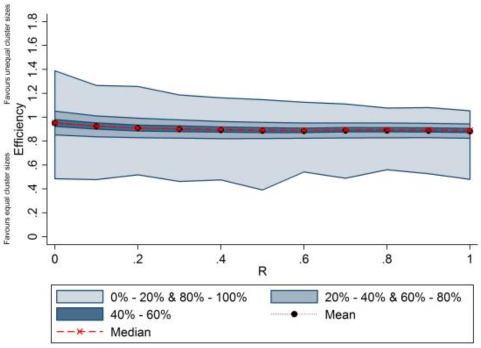
CV = 0.5



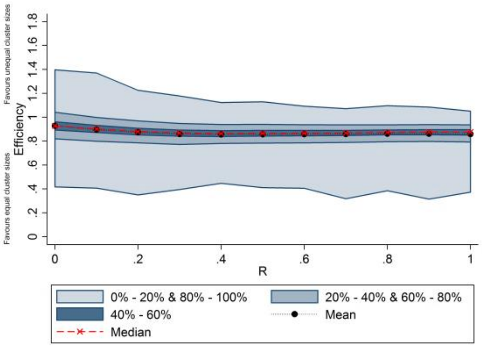
CV = 0.75



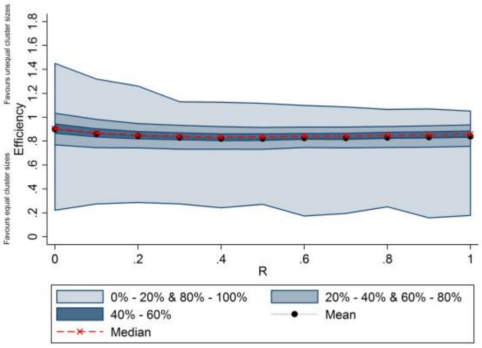
CV = 1.0



CV = 1.25

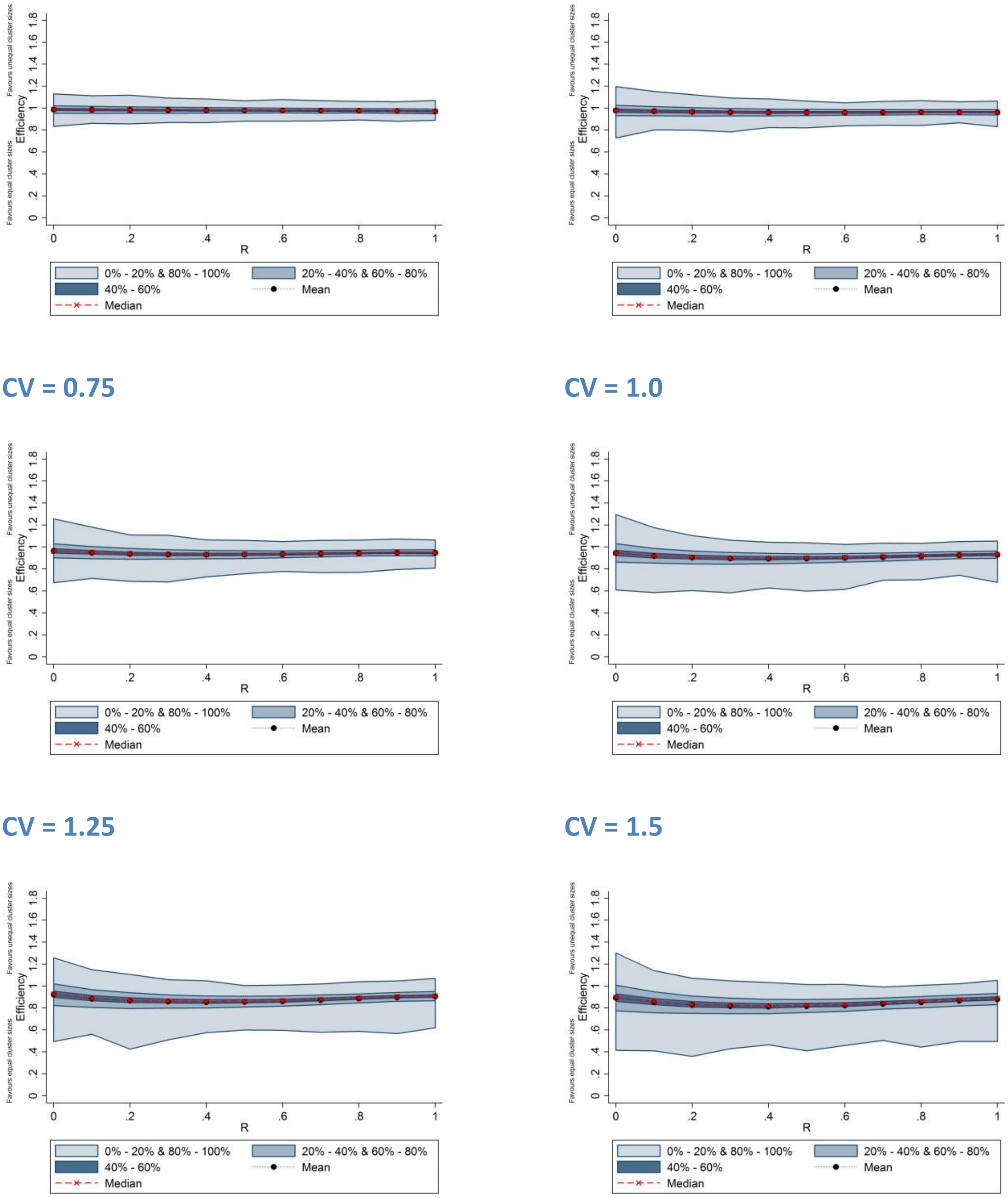


CV = 1.5



Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 12 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = 0.5.



CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

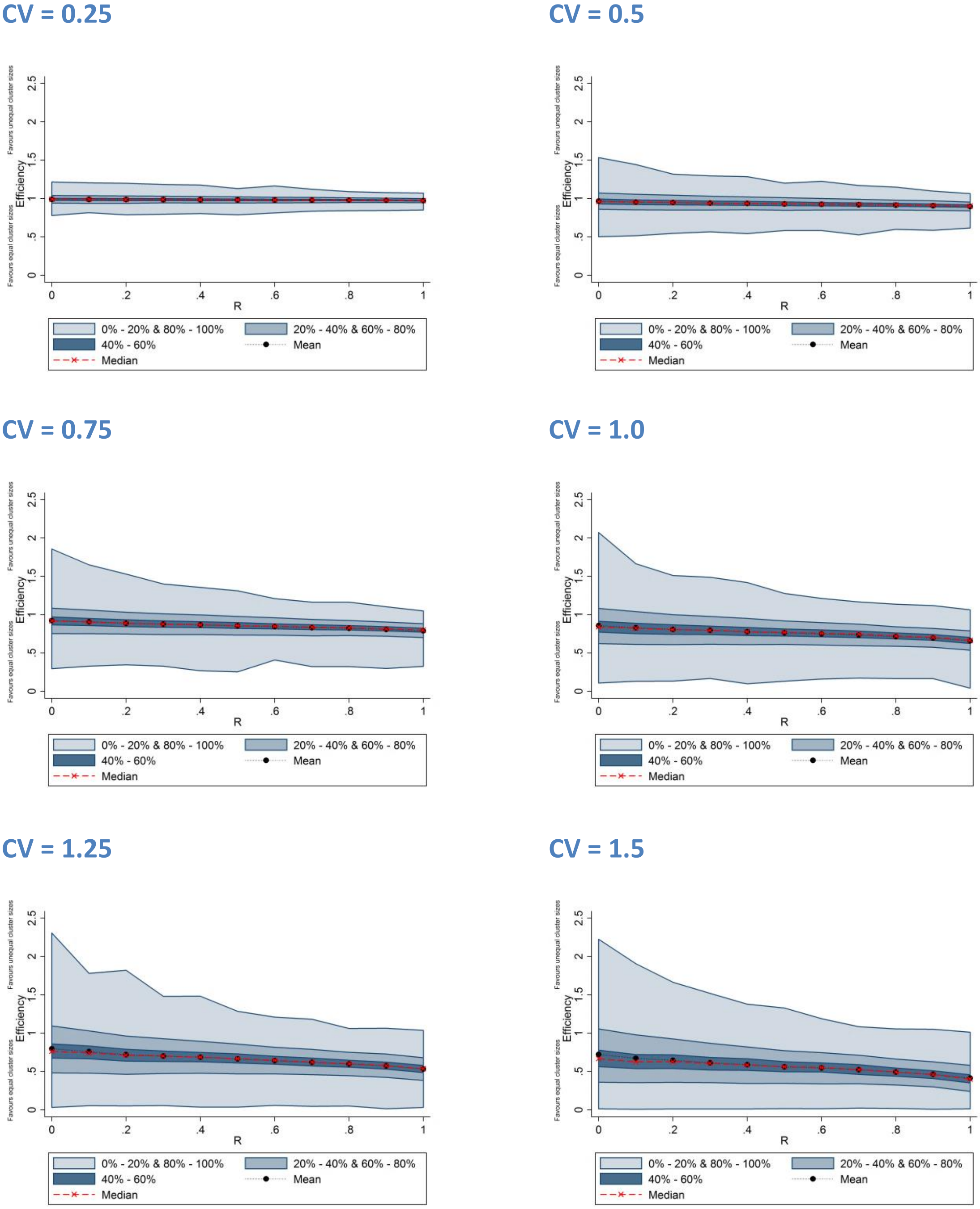
CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Within-cluster coefficient of variation = between-cluster coefficient of variation

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 12 clusters and 2 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = between-cluster coefficient of variation.



CV = 0.25

CV = 0.5

CV = 0.75

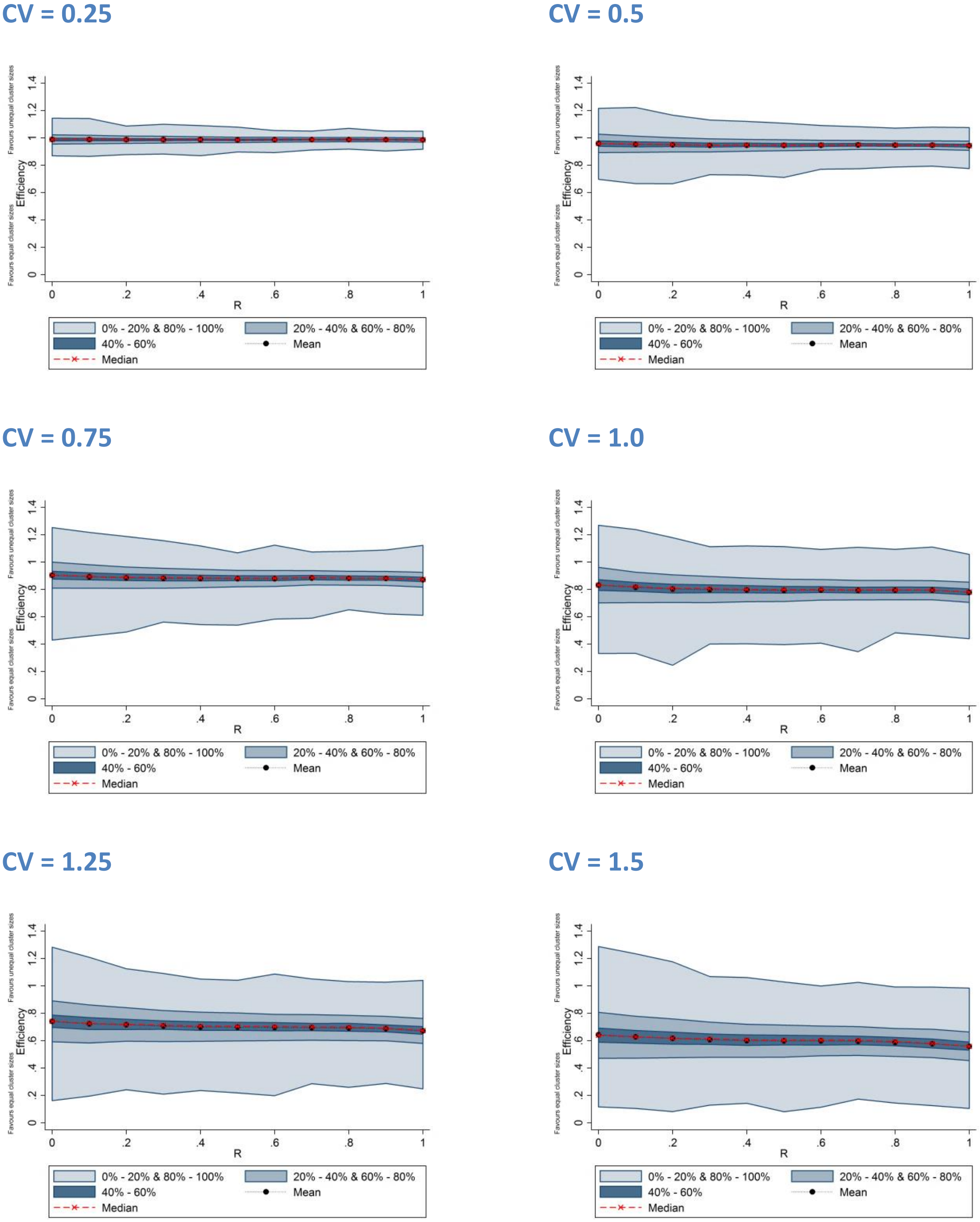
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 12 clusters and 12 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = between-cluster coefficient of variation.



CV = 0.25

CV = 0.5

CV = 0.75

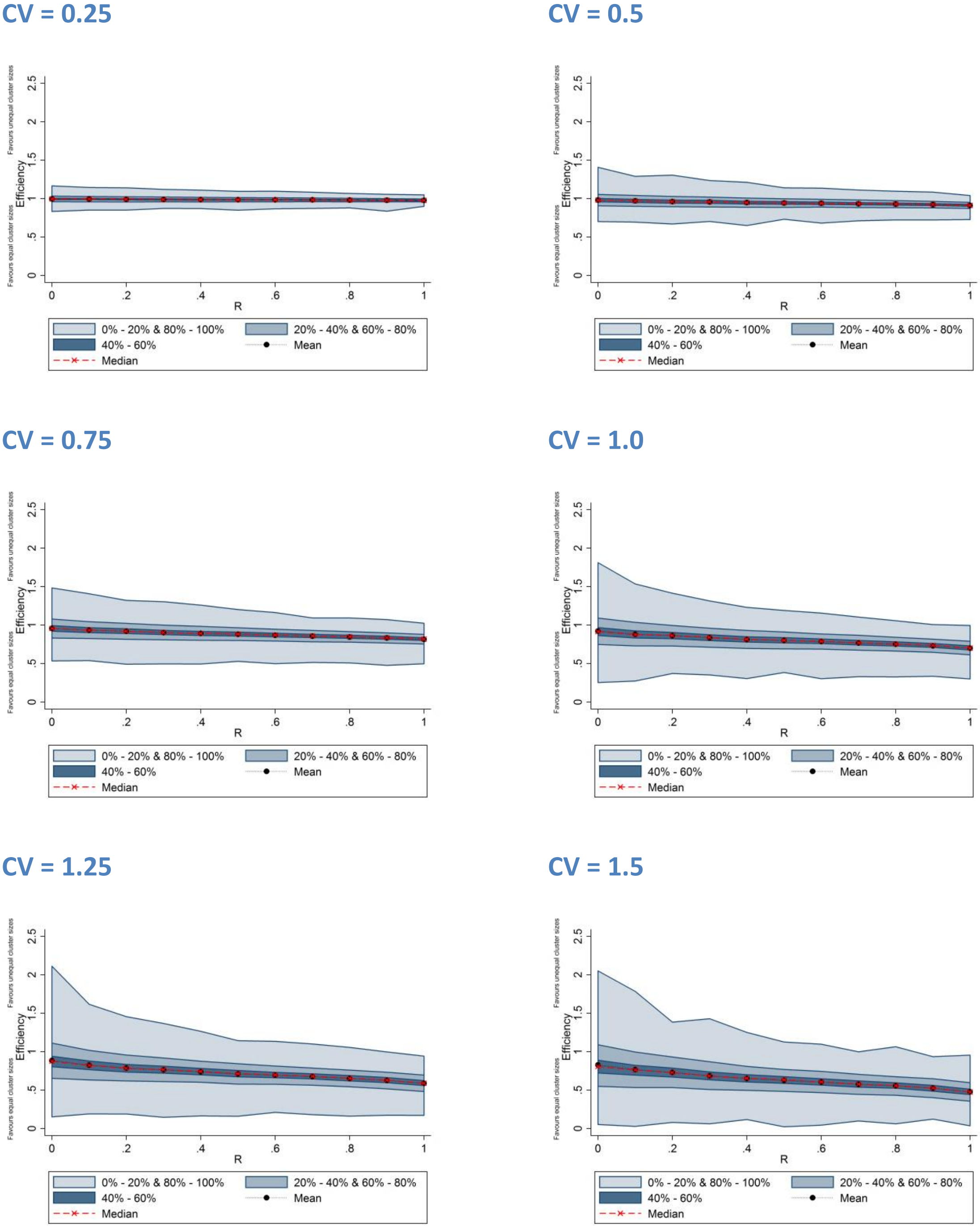
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 2 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = between-cluster coefficient of variation.



CV = 0.25

CV = 0.5

CV = 0.75

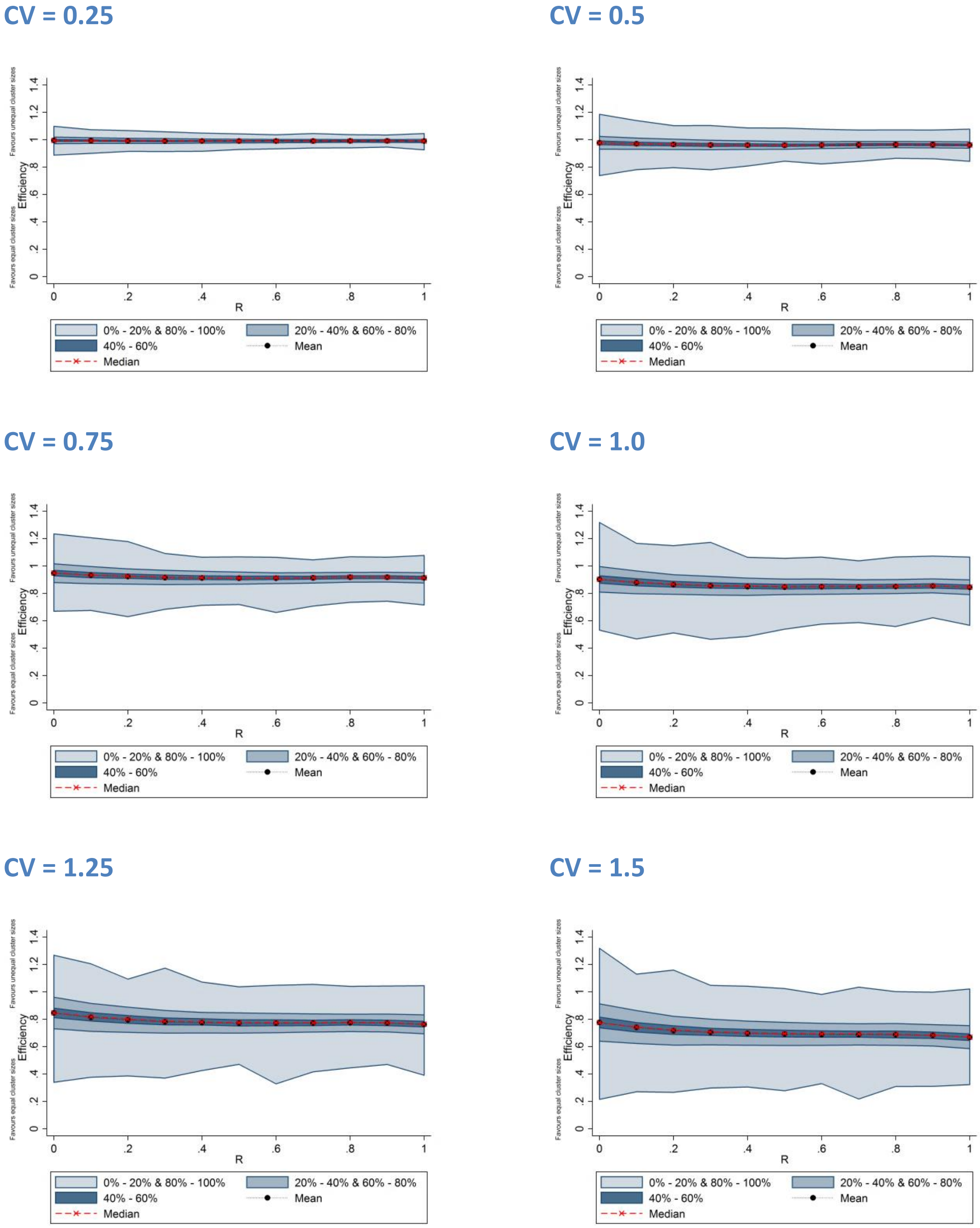
CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.

Efficiency vs cluster mean correlation (R) for a selection of between-cluster coefficient of variation (CV) values for a SW-CRT with 24 clusters and 12 randomisation steps. Cluster size = 72. The within-cluster coefficient of variation = between-cluster coefficient of variation.



CV = 0.25

CV = 0.5

CV = 0.75

CV = 1.0

CV = 1.25

CV = 1.5

Efficiency is calculated as the ratio of the precision in a SW-CRT with unequal cluster sizes compared to the precision in a SW-CRT with equal cluster sizes. R represents the cluster mean correlation.