

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



LSHTM Research Online

Thompson, JA; (2018) Improving the Design and Analysis of Stepped-Wedge Trials. PhD thesis, London School of Hygiene & Tropical Medicine. DOI: <https://doi.org/10.17037/PUBS.04647855>

Downloaded from: <https://researchonline.lshtm.ac.uk/id/eprint/4647855/>

DOI: <https://doi.org/10.17037/PUBS.04647855>

Usage Guidelines:

Please refer to usage guidelines at <https://researchonline.lshtm.ac.uk/policies.html> or alternatively contact researchonline@lshtm.ac.uk.

Available under license. To note, 3rd party material is not necessarily covered under this license: <http://creativecommons.org/licenses/by-nc-nd/3.0/>

<https://researchonline.lshtm.ac.uk>

LONDON
SCHOOL *of*
HYGIENE
& TROPICAL
MEDICINE



Improving the Design and Analysis of Stepped-Wedge Trials

Jennifer Anne Thompson

Thesis submitted in accordance with the requirements
for the degree of Doctor of Philosophy of the University
of London

March 2018

Department of Infectious Disease Epidemiology
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine
University of London

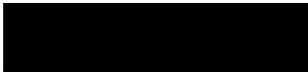
Funded by the Medical Research Council London Hub for Trials
Methodology Research, Grant code MR/L004933/1-P27

This thesis is dedicated to my parents

for their love and support

Declaration

I, Jennifer Anne Thompson, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: 

Date: 19/09/2017

Abstract

Despite their growing use, there is limited literature on the design and analysis of stepped-wedge trials (SWTs). The design is characterised by some or all clusters experiencing the control condition follow by the intervention condition. This enables within-cluster comparisons, but confounds the intervention effect with secular trends.

In this thesis, I aim to use statistical methodology to improve the design efficiency and identify robust methods of analysis.

I provide a new formulation of a design effect for an SWT. From this, I identify that it is more efficient for a trial to begin after the first clusters switch to the intervention, and end before the final clusters switch to the intervention.

SWTs are commonly analysed using a mixed-effect model with a random effect for cluster and fixed effects for periods and the intervention. Through a wide-ranging simulation study, I found that this “standard” model is sensitive to deviations from model assumptions and suggest adding a random effect for period. However, this alternative model still suffered from confidence interval under coverage in some scenarios.

I introduce a novel method of analysis that excludes the within-cluster comparisons. Each period of the trial is analysed separately to give within-period intervention effect estimates. An inverse-variance weighted average provides an overall effect and permutation tests provide a p-value and confidence intervals. In a simulation study, I find that this novel method provides unbiased inference in a range of scenarios, but had lower power than the standard model when the standard model was correctly specified. I introduce a new Stata command I have developed to conduct this novel method in order to encourage wider use of this new methodology.

These finding will lead to trialists using more efficient trial designs and more robust analysis methods.

Acknowledgements

I would firstly like to thank my supervisors, Katherine Fielding and James Hargreaves, for our long, hand-waving conversations about horizontal comparisons and period effects. Your encouragement and guidance have been invaluable. I would also like to thank Richard Hayes, Andrew Copas, and Calum Davey for their helpful input into this work.

The last three years have been made so much more enjoyable because of friends both within LSHTM and outside. To the friends with whom I have shared an office, thank you for the countless tea breaks. To my friends outside academia, thank you so much for trying to learn the meaning of the term stepped-wedge trial.

This thesis would not have been possible without my parents, to whom this thesis is dedicated. Their unconditional love throughout my childhood has given me the confidence to reach for the stars. My dad passed away nine years ago, but continues to influence my life every day. My mum is a constant source of support. I will forever be grateful for everything that they have done for me.

Lastly, I am indebted to my partner John Hales, who proof-read this thesis and has encouraged me be the best version of myself. Thank you for your unwavering support and interest in my work.

Contents

Declaration	3
Abstract	4
Acknowledgments	5
List of Figures	9
List of Tables	12
Glossary	14
1 Introduction	21
1.1 Cluster Randomised Trials	22
1.2 Stepped-Wedge Trials	23
1.3 Related Cluster Randomised Designs	25
1.4 Reasons for Choosing a Stepped-Wedge Trial Design	27
1.5 Thesis Outline	28
1.6 Role of the Candidate	30
2 Stepped-Wedge Trial Case Studies	37
2.1 Deworming School Children Trial	37
2.2 Tuberculosis Diagnostic Test Trial	41
2.3 Hypothetical NHS Health-Check Trial	45
2.4 Summary	46
3 Current Recommendations on Design and Analysis	50
3.1 The Breadth of Stepped-Wedge Trial Designs	50
3.2 Analysis Methods	57
3.3 Power and Sample Size	73
3.4 Summary	80
4 Aims and objectives	91

5	Methods: Background to Simulation Studies	92
5.1	Data Generation	93
5.2	Analysing Simulated Data	95
5.3	Number of Simulations	95
5.4	Analysing the Simulation Study	96
5.5	Summary	99
6	Paper A: The Optimal Design of the Stepped-Wedge Trials with Equal Allocation to Sequences, and a Comparison to Other Trial Designs	101
	Paper	103
	Appendix	112
7	Paper B: Bias and Inference from Misspecified Mixed-Effect Models in Stepped-Wedge Trial Analysis	131
	Paper	134
	Supporting Information	147
8	Paper C: Robust Analysis of Stepped-Wedge Trials using Cluster-Level Summaries Within Periods	170
8.1	Introduction	174
8.2	Application to Tuberculosis Diagnostic Trial	178
8.3	Simulation Study	180
8.4	Discussion	187
8.5	Supporting Information	191
9	Paper D: A Stata Command for Conducting Permutation Tests for Stepped-Wedge Cluster Randomised Trials	203
9.1	Introduction	206
9.2	Technical Details	207
9.3	The swpermute Command	212
9.4	The Dialog Box	217
9.5	Example	217
9.6	Concluding Remarks	225
9.7	Supporting Information	227
10	Discussion and Conclusions	232
10.1	Synthesis of Findings	233
10.2	Dissemination and Increasing Impact	240

10.3 Strengths and Limitations	241
10.4 Future Research Directions	244
10.5 Concluding Remarks	246
Appendix A: License Agreement for Paper A	253
Appendix B: Ethics approval for Paper B	255
Appendix C: License Agreement for Paper B	257
Appendix D: Ethics approval for Paper C	264
Appendix E: Ethics approval for Paper D	266

List of Figures

Declaration	3
1 Introduction	21
1.1 Schematic of a standard stepped-wedge trial	24
1.2 Schematics of other cluster-randomised trial designs	26
2 Stepped-Wedge Trial Case Studies	37
2.1 Schematic of deworming trial design	38
2.2 Deworming trial school attendance in each year, by sequence . .	40
2.3 Schematic of TB diagnosis trial design	42
2.4 Graphs of TB diagnostic trial outcomes in each study month by diagnostic test	44
2.5 Graph of NHS health-check acceptance over time.	46
2.6 Schematic of hypothetical NHS health-check SWTs	46
3 Current Recommendations on Design and Analysis	50
3.1 Schematics of modified stepped-wedge trials	52
3.2 Diagram of trial participant exposure and observation	54
3.3 Diagram of a design-pattern matrix	80
4 Aims and objectives	91
5 Methods: Background to Simulation Studies	92
6 Paper A: The Optimal Design of the Stepped-Wedge Trials with Equal Allocation to Sequences, and a Comparison to Other Trial Designs	101
6.1 Diagrammatic illustrations of trial designs	104
6.2 Optimal number of sequences by the cluster-mean correlation . .	107
6.3 Sample size of the SWT with no observations outside rollout relative to the optimised hybrid against the cluster-mean cor- relation	107

7	Paper B: Bias and Inference from Misspecified Mixed-Effect Models in Stepped-Wedge Trial Analysis	131
7.1	Schematic of motivating example	136
7.2	Simulation study period effect scenario	137
7.3	Mean and spread of intervention effect log odds ratio in each scenario with the same in effect in groups one and two by analysis model	141
7.4	Mean standard error and confidence interval coverage of estimated intervention effect in each scenario with the same in effect in groups one and two by analysis model	142
7.5	Mean and spread of intervention effect log odds ratio in each scenario with the different effects in groups one and two by analysis model	142
S5a:	Mean and spread of estimated intercept log odds by simulation parameters	153
S6a:	Mean and spread of period effect log odds ratio estimates by simulation parameters	156
S9a:	Mean and spread of intercept between-cluster variance estimates by simulation parameters	163
S10a:	Type-one error by simulation parameters	166
8	Paper C: Robust Analysis of Stepped-Wedge Trials using Cluster-Level Summaries Within Periods	170
8.1	Simulation study scenarios secular trends and ICC	181
8.2	Trial schematics used in simulation study	182
8.3	Mean and spread of intervention effect estimates in each scenario and trial design, by analysis method	184
8.4	Coverage of 95% confidence intervals in each scenario and trial design, by analysis method	185
9	Paper D: A Stata Command for Conducting Permutation Tests for Stepped-Wedge Cluster Randomised Trials	203
9.1	Schematics of cluster-randomised trial designs	207
9.2	Proportion of patients in the TB diagnostic trial with a confirmed diagnosis in each study month by type of test used	220
9.3	Distribution of the permutation test log odds ratio under the null hypothesis of no effect	222
S1:	Screen shots of dialog boxes to run example 5.3	227

10 Discussion and Conclusions	232
--------------------------------------	------------

List of Tables

Declaration	3
1 Introduction	21
2 Stepped-Wedge Trial Case Studies	37
2.1 Deworming trial school and pupil characteristics by randomised sequence	39
2.2 TB diagnostic trial patient characteristics by diagnostic test . .	43
3 Current Recommendations on Design and Analysis	50
4 Aims and objectives	91
5 Methods: Background to Simulation Studies	92
6 Paper A: The Optimal Design of the Stepped-Wedge Trials with Equal Allocation to Sequences, and a Comparison to Other Trial Designs	101
6.1 Illustrative example of the number of clusters required by dif- ferent trial designs	109
A1: Sample size of an SWT with the optimal proportion of observa- tions outside rollout and with an increased number of sequences relative to an original number of sequences	128
7 Paper B: Bias and Inference from Misspecified Mixed-Effect Models in Stepped-Wedge Trial Analysis	131
7.1 Summary of simulation study data scenarios	137
7.2 Intervention effect estimates from motivating example from dif- ferent analysis models	143
S3: Table of convergence of analysis models by simulation parameters	149
S4: Mean of intervention effect log odds ratio estimates by simulation parameters	151

S5b: Mean of intercept log odds estimates by simulation parameters .	154
S6b: Mean period effect log odds ratio estimates by simulation parameters	157
S7: Mean of standard error estimates by simulation parameters . . .	159
S8: Coverage of 95% confidence intervals by simulation parameters .	161
S9b: Mean of intercept between-cluster variance estimates by simulation parameters	164
S10b: Type-one error rate by simulation parameters	167
8 Paper C: Robust Analysis of Stepped-Wedge Trials using Cluster-Level Summaries Within Periods	170
8.1 Stages in estimating the risk difference using the cluster-summary analysis	179
8.2 Statistical power in each scenario and trial design, by analysis method	186
S3: Number of times the heuristic adjustment was used in each scenario and trial design	196
S4a: Mean of intervention effect log odds ratio estimates	197
S4b: Standard deviation of intervention effect log odds ratio estimates	198
S5: Intervention effect 95% confidence interval coverage	199
9 Paper D: A Stata Command for Conducting Permutation Tests for Stepped-Wedge Cluster Randomised Trials	203
10 Discussion and Conclusions	232

Glossary

Acronyms

CI	Confidence interval
CONSORT	Consolidated standards of reporting trials
CRT	Parallel cluster-randomised trial
CRXO	Crossover cluster-randomised trial
GEE	Generalised estimating equations
GLMM	Generalised linear mixed model
GP	General practice
HIV	Human Immunodeficiency Virus
ICC	Intraclass correlation coefficient
IQR	Inter-quartile range
LMM	Linear mixed model
NHS	National health service
OR	Odds Ratio
RCT	Randomised controlled trial
SWT	Stepped-wedge cluster-randomised trial
TB	Tuberculosis
WHO	World Health Organisation

Notation

Correlation parameters

R	The cluster-mean correlation, assuming equal correlation within clusters
R^*	The cluster mean correlation, allowing for repeated observations of individuals and for cluster-specific period effects
ρ	ICC
π	Cluster-level autocorrelation
σ^2	The total variance of the outcome. Subscripts are added to denote components of the total variance

Trial design parameters

b	Number of periods before baseline
g	Number of sequences. This is denoted k in chapter 6
c_{0j}, c_{1j}	Number of clusters in the control and intervention conditions respectively in period j
m	Total cluster size
M	Total number of observations in the trial
n	Number of observations in a cluster-period

Analysis model parameters

i	Index for cluster $i = 1, \dots, I$
j	Index for period $j = 1, \dots, J$
k	Index for observation $k = 1, \dots, K$. Also the number of sequences in chapter 6
y_{ijk}	Outcome of observation k in period j in cluster i
μ	Mean outcome in the first period in the control condition
β_j	Period effect comparing period j to the first period
t_j	The time between period 0 and period j
X_{ij}	Indicator variable equal to 1 if cluster i receives the intervention in period j , 0 otherwise

S_i	Variable indicating the period in which cluster i switches to the intervention
θ	Intervention effect. The subscript A denotes a specific value of the intervention effect. The subscript j a within-period intervention effect. $\hat{\theta}$ denotes an estimate of the intervention effect
θ'	Change in the intervention effect with longer in the intervention condition
d_{ik}	Random effect for an observation k in cluster i $N(0, \sigma_d^2)$
e_{ijk}	Random effect for an observation in cluster i in period j for observation k $N(0, \sigma_e^2)$
u_i	Random effect for cluster i $N(0, \sigma_u^2)$
v_{ij}	Random effect cluster-specific period effects $MVN(0, \Lambda_{v_j})$
q_{ij}	Random effect for cluster i in period j $N(0, \sigma_q^2)$
z_i	Random effect for cluster-specific intervention effects $N(0, \sigma_z^2)$
Λ	A covariance matrix
γ_i	Power given to the second term of the period effect (chapter 8 only)

Within period analysis parameters

p_{ij}	Probability of an outcome in cluster i in period j . p_{ij}^* denoted the probability after applying a heuristic adjustment for all cases or all controls
s_{0j}, s_{1j}	The variance of the cluster summaries in the control and intervention conditions respectively in period j
\hat{w}	Vector of estimated weights given to each period-specific intervention effect
\hat{w}_j	Weight given to the intervention effect in period j

Permutation test parameters

f	Number of permutations out of all possible permutations that give a parameter estimate the same or more extreme than that observed. f^* denotes the number of Monte-Carlo permutations.
F	Total number of permutations of clusters to sequences. F^* denotes the number of Monte-Carlo permutations

General

α	Type-one error rate. In chapter 6 this is the proportion after rollout
β	Statistical power. In chapter 6 this is the proportion before rollout. In chapter 7 this is the period effect since there are only two periods, so no subscript is used
$E(x)$	The expectation of x
$Var(x)$	Variance of x
$Cov(x, y)$	Covariance of x and y
$\Phi(x)$	Cumulative standard normal distribution function at a value x
$\Phi^{-1}(x)$	Inverse cumulative standard normal distribution at a value x
N	Number of simulation runs

Terminology

Terms used in this thesis	Description
Stepped-wedge trial (SWT)	A trial where all or some clusters switch from the control condition to the intervention condition during the trial. Clusters only ever switch from control to intervention, never intervention to control. Also referred to elsewhere as a multiple baseline trial, or unidirectional crossover trial. See figures 1.1 and 3.1.
Parallel cluster randomised trial (CRT)	A trial where half of the clusters are randomised to receive a control condition for the duration of the trial and the remaining clusters receive the intervention for the duration of the trial. See figure 1.2a.
Parallel cluster randomised trial with baseline observations	A CRT with a period of data collection before any clusters receive the intervention. See figure 1.2b.
Cluster crossover trial (CRXO)	A trial where half of the cluster are randomised to receive the control condition followed by the intervention condition, and the remaining cluster are randomised to receive the intervention condition followed by the control condition. See figure 1.2c.
Hybrid trial	A type of SWT with an unequal allocation of clusters to sequences. Some proportion of the clusters are randomised to receive the control or intervention condition for the duration of the trial, the remaining clusters are randomised to an SWT with half a period before and half a period after rollout. See figure 3.1d.
Condition	Refers to whether clusters are receiving the control or the intervention. Known elsewhere as arms, or treatments.

Terms used in this thesis	Description
Clusters	The collections of individuals and unit of randomisation. Indicated by the subscript $i = 1, \dots, I$. Known elsewhere as groups.
Switch-points	Describes the times at which clusters switch from the control to the intervention condition. Known elsewhere as steps, crossover-points, and uptake-times.
Sequences (chapter 1-6, 8, & 9) / Groups (chapter 7),	Defines the randomisation of clusters. Clusters are randomised to one of g sequences/ groups that switch from control to intervention at different times (switch-points) during the trial. Elsewhere, these are also known as steps or arms.
Rollout	The time between the first sequence switching to the intervention and the final sequence switching to the intervention. Before rollout, all clusters are in the control condition. After rollout, all clusters are in the intervention condition.
Periods	The time between subsequent sequences switching from control to the intervention. Indicated by the subscript $j = 1, \dots, J$. There may also be a period before the first sequence switches and/or after the final sequence switches. Elsewhere, known as epochs or steps.
Standard SWT design	An SWT with one period before rollout and one period after rollout. Each period contains the same number of observations and an equal number of clusters are randomised to each sequence. See figure 1.1.
Incomplete SWT	An SWT with no observations collected in some clusters in some periods. In a complete SWT, observations are collected in all clusters for all periods.
Total cluster size	The number of observations collected in each cluster across the whole trial.

Terms used in this thesis	Description
Secular trends/period effect	Changes in the outcome over time.
Vertical comparison	Comparison of outcomes from cluster in the control condition to outcomes from clusters in the intervention condition within a period.
Horizontal comparison	Comparison of outcomes from periods in the control condition to outcomes from periods in the intervention condition within a cluster.
Hussey and Hughes model, (chapter 1-6, & 9) / Standard model (chapter 7 & 8)	A mixed-effect model with a random effect for cluster and fixed effects for the periods and intervention, as shown in model 3.1. This is the most common analysis for SWTs.
Cluster-period interaction model	A mixed-effect model with a random effects for cluster and the interaction between clusters and periods, with a fixed effect for intervention, as shown in model 3.2.
Random-period model	A mixed -effect model with random effects for clusters and periods, and a fixed effect for intervention as shown in model 3.4. The random effects have an unstructured covariance matrix.
Random-intervention model	A mixed-effect model with random effects for cluster and intervention, and a fixed effects for periods, as shown in model 3.6. The random effects have an unstructured covariance matrix.

1 Introduction

Since the early beginnings of medical research, there have been concerns about statistical methods being used inappropriately [1]. Methodological literature aims to improve our understanding of statistical techniques, and to disseminate this knowledge to researchers wishing to use these methods. This has been an issue in many types of research, including randomised controlled trials (RCTs).

In an RCT, individuals are randomised to either receive a control condition, usually the standard of care, or a new intervention condition. By randomising the subjects, all known and unknown factors that could confound the intervention effect should be balanced between the two groups; this allows any difference between the two groups to be attributed to the intervention [2]. In the past, RCTs have been subject to poor implementation of statistical techniques [1], but there have been improvements [3]. These improvements have been reinforced by the development of the Consolidated Standards of Reporting Trials (CONSORT) statement in 1996 that gave guidelines for the essential information that should be reported for a clinical trial [4, 5].

With the advent and growing use of more complex trial designs, the need for better understanding of the statistical methods being used to analyse trials is even greater.

In this thesis, I will provide some important contributions to the methodological literature for stepped-wedge cluster-randomised trials (SWTs); this is a study design that is being used more frequently but where the methodology literature is in its infancy.

This chapter will give an overview of the SWT and related trial designs and discuss some of the reasons for using the SWT design. First, I will introduce cluster-randomised trials and how clustering is described mathematically. In section 1.2, I will describe the SWT design and the terminology used in this field. In section 1.3, I will describe related trial designs that will be referred to in the thesis. Section 1.4 will describe why the SWT design is growing

in popularity and some of drawbacks to using this design. The chapter will conclude with an outline of this thesis and my contributions to the thesis.

1.1 Cluster Randomised Trials

In a cluster randomised trial, sets of individuals are randomised rather than randomising the individuals themselves. These are usually naturally occurring sets of individuals, such as residents in villages, patients in hospital wards, or pupils in schools. This is sometimes done for practical reasons because the intervention can only be introduced to an entire cluster of individuals, for example, providing water wells to villages [6]. Alternatively, it might be that the trialist wants to explore the effect of the intervention within a whole community, for example, introducing vaccines to the majority of children in an area will reduce the chance of the unvaccinated children getting infected (known as herd immunity) [6].

In this type of setting, individuals within the same clusters are likely to be more similar to one another than to individuals from a different clusters; by randomising clusters, we are not getting the same amount of information about the effect of the intervention as if we randomised individuals. The independence assumptions of most simple analysis methods are inappropriate. The degree of clustering is characterised by the intraclass correlation coefficient (ICC, ρ), defined as

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

where σ_b^2 is the variability between clusters, and σ_w^2 is the variability within a cluster [6]. The ICC is the proportion of total variability in the data that is due to between-cluster variability. A high ICC means that observations in the same cluster are relatively more similar to one another than to observations in a different cluster. In health research, the ICC rarely exceeds 0.05 [7, 8].

Girling and Hemming introduced a slightly different concept to describe this correlation that incorporates the total cluster size (the total number of observations in each cluster across the whole of the trial) [9]. Assuming equal correlation within clusters, the cluster-mean correlation, R , is defined as:

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

where m is the total cluster size. Rather than partitioning the total variability, the cluster-mean correlation partitions the variability of the cluster-means into between-cluster variability and within-cluster variability. The cluster-mean correlation varies between 0 and 1, increasing as the ICC or cluster size increase. A cluster-mean correlation of 0 implies all the cluster-mean variability is within-cluster variability, whereas, a cluster-mean correlation of 1 implies that all of the cluster-mean variability is between-cluster variability. Unlike the ICC, which is usually small, the cluster-mean correlation can span from 0 to 1 in health research [9].

Whilst there are other measures of clustering, these are not used in this thesis. There are many types of cluster-randomised trial study designs, but in this thesis, I will focus on the SWT design.

1.2 Stepped-Wedge Trials

In an SWT, sometimes referred to as a multiple baseline design [10] or unidirectional crossover design [11], clusters are randomised into sequences. The trial consists of a number of periods, and clusters in each sequence are exposed to the control condition for a different number of periods then spend the remaining periods exposed to the intervention condition. For example, in a trial with three periods a sequences could consist of two periods in the control condition followed by one period in the intervention condition, or one period in the control condition followed by two periods in the intervention condition. The times at which clusters switch from the control to the intervention condition are known as switch-points.

This can lead to a range of different designs, as will be described in section 3.1. An example of the most commonly used design [12], herein referred to as the standard design, is shown in figure 1.1. In this design, the switch-points are equally spread throughout the study and observations are collected before the first and after the final switch-point.

A key characteristic of these trials is that the intervention effect has been confounded with changes in the outcome over time (secular trends) because the intervention observations are, on average, later in time than the control observations.

The use of the SWT design is growing; the same number of protocol or trial result articles were published in 2013 and 2014 as were published at any time

before 2013 [12]. SWTs have been seen in many areas of research including economics, education, and medical research, but they are most common in health research [12, 13]. They have frequently been used to evaluate health-education interventions [14]. They are now more commonly used in high-income countries [12], but this was not always the case [15].

Figure 1.1: Schematic of a standard stepped-wedge trial

Sequence	Cluster	Period				
		1	2	3	4	5
1	1	CONTROL				
	2					
2	3					
	4					
3	5					
	6					
4	7					
	8					INTERVENTION
		Before rollout	Rollout		After rollout	

The broad use of this design across a range of disciplines, along with a lack of reporting guidelines, has led to inconsistent terminology being used to describe the characteristics of the trial [13]. In this section, the different terminology will be outlined, highlighting the terminology used in this thesis. This terminology is also given in the glossary in the preliminaries of this thesis.

The defining feature of an SWT is that clusters switch from control to the intervention at more than one time-point. In this thesis, I will use the term switch-points throughout to describe the times at which clusters switch, but elsewhere they are sometimes referred to as steps [16], uptake times [9], or crossover points [17].

The time between the first and final switch-point is known as rollout because there are some clusters still in the control condition whilst others have switched to the intervention condition [18]. There are no other common names for this elsewhere in the literature.

The times between switch-points are referred to as periods throughout this thesis, and this terminology is common in the literature [13, 18]. They have also been referred to as epochs [19], or again, steps [20, 21].

This thesis will primarily use the term sequence to describe the randomisation of clusters, although Chapter 7 uses the term groups because it was published earlier. Elsewhere, sequences have been called arms [22], or steps [5].

SWTs are a relatively new type of cluster randomised trial and are more complex than alternative designs.

1.3 Related Cluster Randomised Designs

Often, an SWT is not the only feasible trial design. In this section, I will describe other cluster randomised trial designs that are referred to later in this thesis.

The simplest form of cluster randomised trial is a parallel cluster randomised trial (CRT), where the clusters are randomised to two groups, one of which receives a control condition for the duration of the trial, and the second receives an intervention condition for the duration of the trial (figure 1.2a). CRTs can be designed with observations of follow-up alone or with baseline observations (figures 1.2a and 1.2b respectively). In a study with follow-up alone, clusters are only observed after the intervention clusters have received the intervention. In a CRT with baseline observations, there is an additional baseline period before the intervention is introduced to any clusters, as well as the follow-up period after introduction of the intervention, and the outcome is observed in both periods of the study. The baseline period could contain the same number of observations as the follow-up period, or it could be larger or smaller.

Like SWTs, CRTs are used in many fields of research including medical, education, and economics [23]. They became popular in the 1980s but several reviews noted poor methodological quality of CRTs in the following 20 years [24, 25]. Although there are still methodological issues, there have been improvements in the reporting and analysis of CRTs [26], largely attributed to an increase in methodological papers and books, and the development of CONSORT guidelines [26–28]. However, suboptimal reporting and use of methodology remains common, highlighting that publication of methodology is not sufficient to improve trial methodology, efforts to increase methodology uptake are also required [26].

A lesser used design is a cluster crossover trial (CRXO, figure 1.2c) [29]. This design consists of two periods, with the same number of observations in each. Across these two periods, cluster are randomised to receive either the control

Figure 1.2: Schematics of other cluster-randomised trial designs

Cluster	
1	
2	CONTROL
3	
4	
5	
6	INTERVENTION
7	
8	

(a) Parallel cluster randomised trial (CRT)

Cluster	Baseline period	Follow-up period
1		
2		CONTROL
3		
4		
5		
6		INTERVENTION
7		
8		

(b) Parallel cluster randomised trial with baseline observations

Cluster	Period 1	Period 2
1		
2	CONTROL	INTERVENTION
3		
4		
5		
6	INTERVENTION	CONTROL
7		
8		

(c) Cluster randomised crossover trial (CRXO)

condition in the first period followed by the intervention condition in the second period, or the intervention condition in the first period followed by the control condition in the second period. Like an SWT, clusters are exposed to both the control and intervention conditions. Unlike an SWT, in the CRXO design the intervention effect is not confounded with time effects. A CRXO is only suitable where the intervention can be removed and the intervention effect disappears. As an example, Johnson *et al* [30] conducted a CRXO to assess the impact of parental training and support on developmental outcomes of preterm babies at two years old. The clusters were neonatal units in the UK. Half of the units were randomised to receive the control condition followed by intervention condition and the remaining units received the intervention condition followed by the control condition. This study overcame problems of caused by removing the intervention by firstly recruiting different parents in each period, and secondly, using different staff to implement the intervention on top of the usual care that all parents received.

1.4 Reasons for Choosing a Stepped-Wedge Trial Design

Although still relatively uncommon compared to CRTs, the SWT design is being increasingly used [12]. SWTs have primarily become popular for logistical reasons, but ethical reasons have also been given [31]. However, there is some debate in the literature around these justifications [31–33]

Arguments that restricted resources limit the rollout of the intervention have been countered by the suggestion of a staggered CRT that maintains the balance of calendar time between the two arms [31, 32]. Such a design also has the logistical benefit that only half of the clusters ever receive the intervention, so less effort is required for the intervention rollout [31]. A downside to the staggered rollout in SWT is that they can lead to trials that take longer to complete [18].

Using an SWT is thought to improve recruitment of clusters since all clusters know that they will receive the intervention; this is again countered with a modified CRT design where control clusters are given the intervention after data collection ends if the intervention is shown to be effective [32].

Some think that an SWT is more ethical than a CRT when the intervention is expected to have a positive effect as all clusters eventually receive the intervention [34]. However, even with an SWT there will often still be many individuals within clusters that do not receive the intervention; where it is unethical to withhold the intervention from some clusters, it would also be unethical to delay rollout of the intervention [31, 33].

One justification that is less criticised is conducting an SWT when the intervention is going to be rolled out anyway. By randomising the process of rollout and collecting data while the intervention is being rolled out, the intervention can be assessed at the same time as rollout occurs [31, 35]. Randomisation will mean this assessment is more robust than an assessment of a non-randomised rollout [36].

Lastly, in many settings, SWTs are thought to estimate the intervention effect with a smaller variance than a CRT, meaning trials can require a smaller sample size and maintain the same level of power [34]. This is only the case when there is a sufficiently large correlation between individuals in the same cluster and clusters are sufficiently large [16, 37, 38]. The reason for this

potential increase in statistical power is that the intervention and control conditions can be compared within-clusters as well as between clusters, and most analysis methods utilise all of these comparisons [39–41]. In a CRT, the difference between the intervention and control conditions is assessed by comparing clusters in each condition (between-cluster comparisons). In an SWT, these comparisons are also possible. In each period of the trial, clusters in each condition can be compared to provide an intervention effect; these comparisons are known as vertical comparisons. Only allowing for vertical comparisons, the CRT would have the most power because there are the same number of clusters in each condition at all times [42]. For SWTs, the intervention can also be compared within clusters; these comparisons are known as horizontal comparisons [17, 43]. The horizontal comparisons can be estimated with greater certainty because the between-cluster variability has been removed. However, they are confounded with time because, in an SWT, the intervention is on average later in time than the control condition [17]. This means that they require assumptions about the secular trends of clusters [17, 43]. The confounding with time is a primary feature of SWTs, and is also the design’s main drawback. Failing to appropriately adjust for this known confounder could lead to a biased intervention effect [17].

Despite their growing popularity, I will show in chapter 3 that this is a poorly understood design, with many gaps in the methodological literature. Just as with CRTs in 1980s, the methodological literature is lagging behind the uptake in applying this trial design. Methodological literature is required to ensure SWTs are being designed and analysed appropriately. As also seen with CRTs, further efforts are also required once methodology is available to encourage the uptake of these methods.

1.5 Thesis Outline

This thesis adds to the current literature on designing and analysing SWTs, with a focus on health research. I begin with a broad overview of the current methodological literature for these designs, followed by my research improving the design efficiency and the robustness of analysis.

I use three case studies of SWTs in this thesis to explore different analysis options and to design simulation studies. These are described in chapter 2.

Chapter 3 describes the current literature on SWTs. I begin with a discussion

of what is classed as an SWT, as the term covers a wide range of designs. I discuss whether the most commonly used analysis methods are sufficient to provide unbiased estimates of the intervention effect. I follow this with an exploration of the methods for sample size calculations and how these differ to methods for CRTs.

This review of current literature revealed many knowledge gaps. In chapter 4, I describe the gaps this thesis aims to address; efficient design and robust analysis.

Chapter 5 provides details of the methods used in the simulation studies presented in chapters 7 and 8. All other details of methods are presented within the chapters of novel work.

The next four chapters each give a paper of novel work under taken during this PhD. These are written in the “paper style” format and are either published or ready for publication.

In Chapter 6, I identify changes to the SWT design that reduce the sample size required for a given power. In particular, the first and last period with all clusters in the control and intervention conditions respectively are inefficient and designs which exclude these periods require a smaller sample size.

Chapters 7, 8, and 9 describe work exploring methods of analysis that ensure robust adjustment for secular trends and correctly account for correlation. I have focused on binary outcomes because these are common in practice [44].

Chapter 7 demonstrates that simply adjusting for period effects as fixed categorical variables is not always sufficient and doing so can lead to under-covered confidence intervals. Allowing the secular trends to vary between the clusters corrects this problem. I also identify a similar result that assuming that the intervention effect is common to all clusters can lead to biased intervention effect estimates and under-covered confidence intervals.

In Chapter 8, a novel analysis strategy is introduced that uses cluster-summary analyses within each period, conditioning on time, and so removing all confounding with secular trends. I give an example of this analysis method in trial data and conduct a simulation study to assess its performance against the most commonly used analysis method.

At present, the method described in chapter 8 requires some coding knowledge to run the analysis. In order to simplify use of this method, and improve its uptake, I wrote a Stata command to run this analysis. In Chapter 9, I describe this Stata command and provide a tutorial of its use.

Lastly, Chapter 10 gives a summary of the main conclusions of this work in the context of the broader literature. I describe the implications for SWTs going forward and ideas for future research leading on from my findings.

1.6 Role of the Candidate

I conceived the aims and objectives in this thesis, with support from my supervisors Katherine Fielding and James Hargreaves. In this section, I will describe my role addressing these aims in the description of the current literature (chapter 3), and the four chapters of novel work (chapters 6 to 9). All other chapters in this thesis were written by myself incorporating feedback from Katherine Fielding and James Hargreaves. I had no role in conducting the case studies used in this thesis.

1.6.1 Review of Current Literature (Chapter 3)

I was part of the data extraction team for a series of five articles in the *Trials Journal*. The series focused on a systematic review of SWTs [14], however I did not design this systematic review. I was involved in discussions and paper editing for the article on SWT designs [18]. I played a large role in interpreting the results for the article on SWT analysis [17], and contributed to editing the text of this paper. This series formed a starting point for chapter 3.

I led the remaining contributions of chapter 3, describing the current state of methodology literature for SWTs with guidance from Katherine Fielding and James Hargreaves.

1.6.2 Novel Work on Efficient Design of SWTs (Chapter 6)

Chapter 6 gives a published paper providing recommendations for improving the design efficiency of SWTs. The idea for this paper was initially identified by Andrew Copas along with the idea of reparameterising an existing design effect, but in a simpler form than that given in the published paper. I developed this idea, increasing the flexibility of the design effect beyond the original suggestions of Andrew Copas, and derived all results given in the paper. I also wrote the first draft of the paper. Andrew Copas checked the derivations and suggested substantial edits to the text. Katherine Fielding and James Hargreaves suggested edits to the text.

1.6.3 Novel Work Identifying Problems with Current Analysis Method (Chapter 7)

Chapter 7 gives a published paper detailing the results of a simulation study comparing bias, confidence interval coverage, and power of several analysis models. The motivation for this work stemmed from the paper by Davey *et al* [45], but I lead development and design of the simulation study, conducted the analysis, and wrote the first draft of the paper. Richard Hayes provided guidance in interpretation of results. All co-authors provided comments and edits to the manuscript text.

1.6.4 Novel Work Developing a Cluster-Summary, Within-Period Analysis (Chapter 8)

The idea for the analysis method described in chapter 8 came from the group discussion during the development of the Trials Journal series described above. Calum Davey and I were both keen to explore the idea of an analysis that only includes the vertical comparisons.

The work and the ideas for this chapter were divided between myself and Calum Davey. I led identifying the methods behind conducting a simulation study and we both decided on the design of the simulation study. We both looked for datasets to use as the basis for the simulation study, and I determined representation of the chosen data in terms of statistical distributions. Coding the simulation studies was shared between us; I created a first draft of the simulation study, whilst Calum Davey created a first draft of the code to conduct the cluster-summary, within-period analysis, which we then combined and finalised. The analysis and interpretation of results was divided between us. I conducted the example analysis using the within-period analysis method on a real SWT.

Calum Davey drafted a plan for the paper before I wrote the first draft. Katherine Fielding, James Hargreaves, and Richard Hayes provided guidance throughout this process and suggested edits to the text.

1.6.5 Novel Work Creating a Stata Command (Chapter 9)

I developed the idea to create a Stata command to facilitate the use of permutation tests for SWTs. Katherine Fielding, James Hargreaves, Richard Hayes

and Calum Davey contributed ideas for functionality for the command. I then coded the command and conducted the majority of testing. The command was also tested by Calum Davey, James Hargreaves, and Katherine Fielding. I wrote the first draft of the manuscript and all co-authors provided comments on the text.

Bibliography

- [1] Altman DG and Simera I. A history of the evolution of guidelines for reporting medical research: the long road to the EQUATOR Network. *Journal of the Royal Society of Medicine* 2016. 109. (2):67–77.
- [2] Kirkwood B and Sterne J. *Essential Medical Statistics*. Wiley, 2010.
- [3] Falagas ME, Grigori T and Ioannidou E. A systematic review of trends in the methodological quality of randomized controlled trials in various research fields. *Journal of Clinical Epidemiology* 2009. 62. (3):227–231.e229.
- [4] Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C and Gaboury I. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Medical Journal of Australia* 2006. 185. (5):263–267.
- [5] Schultz TJ, Kitson AL, Soenen S, Long L, Shanks A, Wiechula R, Chapman I and Lange K. Does a multidisciplinary nutritional intervention prevent nutritional decline in hospital patients? A stepped wedge randomised cluster trial. *e-SPEN Journal* 2014. 9. (2):e84–e90.
- [6] Hayes RJ and Moulton LH. *Cluster Randomised Trials*. 1st ed. USA: Chapman and Hall/CRC, 2009.
- [7] Pagel C, Prost A, Lewycka S, Das S, Colbourn T, Mahapatra R, Azad K, Costello A and Osrin D. 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.
- [8] Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S and Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology* 2004. 57. (8):785–794.

- [9] Girling AJ and Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Statistics in Medicine* 2016. 35. (13):2149–2166.
- [10] Rhoda DA, Murray DM, Andridge RR, Pennell ML and Hade EM. Studies with staggered starts: multiple baseline designs and group-randomized trials. *American Journal of Public Health* 2011. 101. (11):2164–2169.
- [11] Zhan Z, Bock GH de and Heuvel ER van den. Statistical methods for uni-directional switch designs: Past, present, and future. *Statistical Methods Medical Research* 2017:962280216689280.
- [12] Martin J, Taljaard M, Girling A and 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.
- [13] Grayling MJ, Wason JM and Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. *Trials* 2017. 18. (1):33.
- [14] Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, Thompson JA, Fielding KL, Omar RZ, Ononge S, Hargreaves J and Prost A. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 2015. 16. (1):353.
- [15] Brown CA and Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology* 2006. 6. (1):54.
- [16] Woertman W, Hoop E de, Moerbeek M, Zuidema SU, Gerritsen DL and Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology* 2013. 66. (7):752–758.
- [17] Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ and Fielding KL. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials* 2015. 16. (1):358.
- [18] Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G and Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials* 2015. 16. (1):352.
- [19] Hemming K, Haines T, Chilton P, Girling A and Lilford R. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015. 350:h391.

- [20] Hemming K and Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata Journal* 2014. 14. (2):363–380.
- [21] Hooper R and Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ* 2015. 350:h2925.
- [22] Twisk JW, Hoogendijk EO, Zwijsen SA and Boer MR de. Different methods to analyze stepped wedge trial designs revealed different aspects of intervention effects. *Journal of Clinical Epidemiology* 2016. 72:75–83.
- [23] Eldridge SM, Ashby D, Feder GS, Rudnicka AR and 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.
- [24] Simpson JM, Klar N and Donnor A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *American Journal of Public Health* 1995. 85. (10):1378–1383.
- [25] Donner A, Brown KS and Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *International Journal of Epidemiology* 1990. 19. (4):795–800.
- [26] Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, Skea Z, Brehaut JC, Boruch RF, Eccles MP, Grimshaw JM, Weijer C, Zwarenstein M and Donner A. 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.
- [27] Campbell MK, Piaggio G, Elbourne DR, Altman DG and Group C. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012. 345:e5661.
- [28] Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Medical Research Methodology* 2004. 4. (1):21.
- [29] Arnup SJ, Forbes AB, Kahan BC, Morgan KE and McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials. *Journal of Clinical Epidemiology* 2016. 74:40–50.
- [30] Johnson S, Whitelaw A, Glazebrook C, Israel C, Turner R, White IR, Croudace T, Davenport F and Marlow N. Randomized trial of a parenting intervention for very preterm infants: outcome at 2 years. *Journal of Pediatrics* 2009. 155. (4):488–494.

- [31] Prost A, Binik A, Abubakar I, Roy A, De Allegri M, Mouchoux C, Dreischulte T, Ayles H, Lewis JJ and Osrin D. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. *Trials* 2015. 16. (1):351.
- [32] Kotz D, Spigt M, Arts IC, Crutzen R and 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):1249–1252.
- [33] Doussau A and Grady C. Deciphering assumptions about stepped wedge designs: the case of Ebola vaccine research. *Journal of Medical Ethics* 2016.
- [34] Zhan Z, Heuvel ER van den, Doornbos PM, Burger H, Verberne CJ, Wiggers T and Bock GH de. 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):454–461.
- [35] Mdege ND, Man MS, Taylor Nee Brown CA and 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):936–948.
- [36] Hargreaves JR, Prost A, Fielding KL and Copas AJ. How important is randomisation in a stepped wedge trial? *Trials* 2015. 16. (1):359.
- [37] Hemming K and 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):1427–1428.
- [38] Hemming K, Girling A, Martin J and 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):1058–1059.
- [39] Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials* 2007. 28. (2):182–191.
- [40] Ji X, Fink G, Robyn PJ and Small SS. Randomization inference for stepped-wedge cluster-randomised trials: An application to community-based health insurance. *Annals of Applied Statistics* 2017. 11. (1):1–20.

- [41] Scott JM, deCamp A, Juraska M, Fay MP and Gilbert PB. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. *Statistical Methods in Medical Research* 2017. 26. (2):583–597.
- [42] Moulton LH, Golub JE, Durovni B, Cavalcante SC, Pacheco AG, Saraceni V, King B and Chaisson RE. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clinical Trials* 2007. 4. (2):190–199.
- [43] Matthews JNS and Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. *Statistics in Medicine* 2017:1–18.
- [44] Barker D, McElduff P, D’Este C and Campbell MJ. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. *BMC Medical Research Methodology* 2016. 16. (1):69.
- [45] Davey C, Aiken AM, Hayes RJ and Hargreaves JR. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. *International Journal of Epidemiology* 2015. 44. (5):1581–1592.

2 Stepped-Wedge Trial Case Studies

Throughout this thesis I will use three examples of stepped-wedge trials as case studies. Two are completed SWTs, and one is a hypothetical SWT in a real setting. In the four papers that make up the original contributions of my thesis, these case studies are used as examples and as the basis of simulation studies.

2.1 Deworming School Children Trial

In 1998 a group of researchers used an SWT to explore the impact of deworming school children on their school attendance.

Helminths are a type of parasitic worm that includes roundworm, hookworm, and whipworm. Helminth infections are very common, and it is estimated that, in 2017, 1.5 billion people (24% of the worlds population) are infected [1]. Prevalence of infection are particularly high in deprived areas with poor sanitation [1]. In the Busia district of Kenya, where this trial was conducted, over 90% of school children were infected with a Helminth infection in 1998 [2].

With such high prevalence of infection, an intervention to treat infections, and so prevent onward transmission, is appealing. To this end, the World Health Organisation (WHO) recommend periodic deworming treatment of all at-risk people living in areas with a prevalence above 20% [1], and school based deworming programs are now common place [3].

Whilst it is generally accepted that the use of deworming drugs and improving hygiene practices can induce a large reduction in the number of helminth infections [4], the secondary impacts of such programs are debated. In particular, some researchers claim that these programs increase school attendance

in children [5], but a recent Cochrane review has concluded that the evidence is lacking [3].

In Chapter 6 of this thesis, I use data from one of the most highly cited deworming trials investigating the impact on school attendance [6].

In 1998 and 1999, trialists ran a quasi-randomised SWT to explore the effect of a deworming intervention on school attendance [5]. The intervention they tested was administration of deworming drugs annually to all eligible children in participating schools along with education on how to prevent infection, such as hand-washing advice and advising children to wear shoes. 75 schools in the Busia district of Kenya were quasi-randomised into 3 sequences. Schools in the first sequence received the deworming intervention in both years, schools in the second sequence received the intervention only in the second year, and schools in the third sequence did not receive the deworming intervention in either year (Figure 2.1). This is described as quasi-randomised because the allocation of schools to sequences was determined by a list sorted alphabetically by area, then by school size. The first school in the list was allocated to sequence 1, the second to sequence 2, and so on down the list. While this approach has potential limitations [7], for the purpose of this thesis I will treat the trial as though it was randomised.

This is an example of an SWT that does not have periods before or after rollout.

Figure 2.1: Schematic of deworming trial design

Sequence Cluster	Period 1: 1998	Period 2: 1999
1		
.		
.		
25		
26		
.		
.		
50		
51		
.		
.		
75		

Many outcomes were collected for children enrolled at these schools, but I will focus on school attendance. This was measured by field workers making unannounced visits to schools: school were observed for an average of 3.8 visits in each year. Some characteristics of the clusters, pupils, and observations are given in Table 2.1. A median of 393 (interquartile range (IQR) 259-528) pupils

were observed in each school across the 2 years with a median of 8 (IQR 6-9) observations per pupil. 14328/27596 (52%) of pupils were male with a median age of 12 (IQR 10-14) at their first observation.

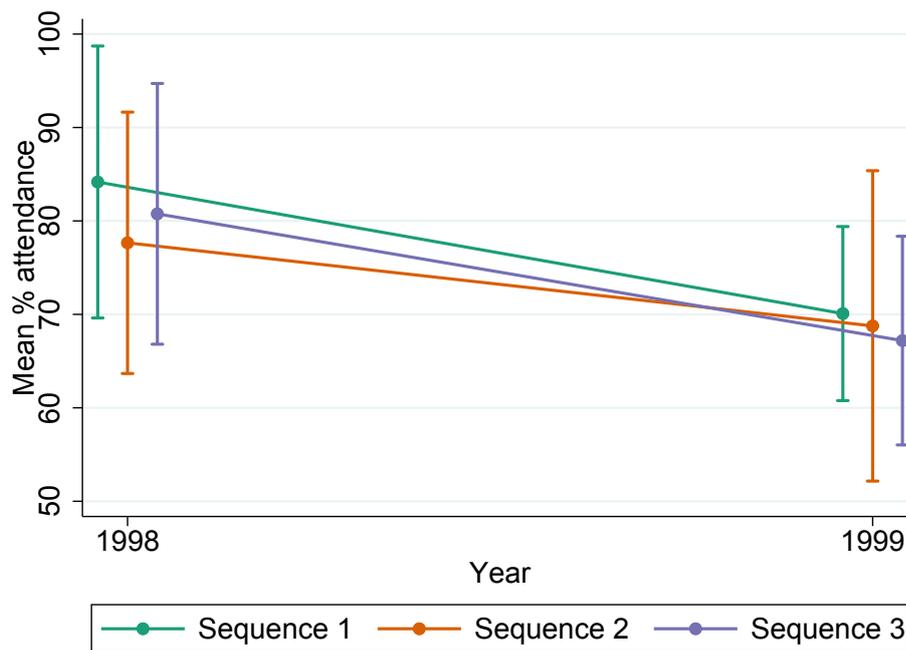
Table 2.1: Deworming trial school and pupil characteristics by randomised sequence

	Sequence 1	Sequence 2	Sequence 3	Total
School level characteristics				
Number of schools	25	25	25	75
Median pupils observed per school (IQR)	393 (281-501)	370 (380-525)	396 (247-528)	393 (259-528)
Median observations per school (IQR)	2866 (1952-3680)	2640 (1812-3866)	2743 (1770-3375)	2743 (1774-3727)
Pupil level characteristics				
Number of pupils	10,319	10,397	9,786	30,502
Median observations per pupil (IQR)	8 (5-9)	8 (6-9)	8 (5-9)	8 (6-9)
Sex				
Female	4648 (45%)	4708 (45%)	3912 (40%)	13268 (44%)
Male	5256 (51%)	4838 (47%)	4238 (43%)	14328 (47%)
Missing	415 (4%)	851 (8%)	1640 (17%)	2906 (10%)
Median age (IQR), years	12 (10-14)	12 (9-14)	13 (10-14)	12 (10-14)
Initial Standard N (%)				
0-1	2521 (24%)	2863 (28%)	2448 (25%)	7832 (26%)
2-3	2636 (26%)	2703 (26%)	2475 (25%)	7814 (26%)
4-5	2333 (23%)	2049 (20%)	2132 (22%)	6514 (21%)
6-8	2400 (23%)	2177 (21%)	2162 (22%)	6739 (22%)
Not known	429 (4%)	605 (6%)	569 (6%)	1603 (5%)

The mean and standard deviation of attendance for schools by year and randomisation sequence is shown in Figure 2.2. The capped bars span one standard deviation either side of the mean school attendance. Sequence 1, which received the intervention in both years, had slightly higher attendance than

sequence 3, which did not receive the intervention in either year. All schools experienced a decrease in attendance between the two years, but there does appear to be a smaller decrease for sequence two, the sequence that switched from the the control condition to the intervention condition. The variability between the attendance at different schools is large relative to the difference between sequences, and it is larger in the first year than the second year.

Figure 2.2: Graph of mean \pm standard deviation of school attendance in each year, by randomised sequence



Overall, the original trialist concluded that the deworming interventions increased school attendance by 5.1% (95% CI 0.8%, 9.4%; $p=0.02$) [5]. However, a recent replication of this trial analysis identified several errors that led to the result above changing to an increase in school attendance of 5.7% (95% CI 3.0%, 8.4%; $p<0.001$) [8].

The partner paper to this replication, a reanalysis of the data, identified an intriguing result [9]. This reanalysis showed that, when analysed with both years' data combined into one model and adjusting for period effects, there was strong evidence that the deworming intervention improved school attendance: in an adjusted analysis, odds ratio (OR)=1.82 (95% CI 1.74, 1.91; $p<0.001$). One would expect the effect estimate from this overall analysis to be between the estimates from the two individual years. But, analysing each year separately with the same analysis model (excluding the time effect because we are

stratifying by time) gave estimates in both years that were smaller than this overall estimate: in 1998 OR=1.48 (0.88, 2.52; p=0.15) and in 1999 OR=1.23 (1.01, 1.51; p=0.04).

This counter-intuitive result motivated the research in Chapter 7. I conduct a simulation study mimicking this trial and reanalyse the trial using different analysis models. In line with the reanalysis of this trial [9], I will ignore clustering within pupils and analyse the trial as though different pupils were observed at each visit.

2.2 Tuberculosis Diagnostic Test Trial

The second case-study I will use in this thesis is an SWT conducted in Brazil that investigated the effect of different diagnostic tests for tuberculosis (TB) on the patient's outcome and whether their TB diagnosis was bacterially confirmed [10].

WHO estimated a TB incidence in Brazil of between 72 and 97 TB notifications per 1,000 person-years in 2015, putting Brazil in the top 20 countries with the highest burden of TB [11]. Treatment success rate is below the global average, despite high rates of TB treatment, and it is hypothesised that early diagnosis of TB and detection of drug resistance could contribute to ending the country's TB epidemic [11].

The standard diagnostic test for TB, sputum smear-microscopy, is quick and cheap, but it has low sensitivity, particularly in patients with human immunodeficiency virus (HIV) [12]. Because of this, many patients are diagnosed with TB based on clinical symptoms alone, even if their test comes back negative [12]. A new TB diagnostic test (Xpert MTB/RIF) is known to be more sensitive than the standard smear-microscopy method of diagnosis, and it also provides a result for rifampicin drug resistance at the time of diagnosis [13]. Whilst the accuracy of the new diagnostic test is well established [13], the impact of this on patient outcomes is not so clear. Several studies have found increases in bacterial confirmation of TB, but this has had little impact on patient outcomes [14, 15].

My second case study is an SWT in Brazil that ran in 2012 looking at this issue.

In 2012 as part of a pilot rollout of the Xpert test in Brazil, this trial team used an SWT to explore the impact of switching from smear microscopy to

the Xpert test [16]. The trial enrolled 14 laboratories covering most diagnoses in the cities of Rio de Janeiro and Manaus in Brazil. At the initiation of the study, all laboratories were using sputum smear microscopy to diagnose TB. Following a month of baseline data collection, the Xpert diagnostic test was rolled out to two laboratories at random each month, so that 7 months later all laboratories were using the Xpert diagnostic test (figure 2.3). The primary outcomes were notification of cases to the national notification system and time to treatment initiation [16]. A second paper (Trajman *et al* [10]) reported on the secondary outcomes of (i) TB treatment outcomes and (ii) prevalence of bacteriological confirmation of individuals diagnosed with TB.

Figure 2.3: Schematic of TB diagnosis trial design

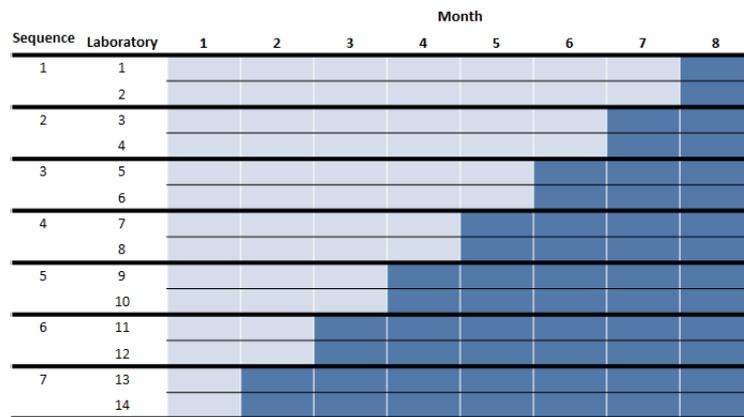


Table 2.2 summarises the patient characteristics from individuals contributing to the secondary outcomes. All patients who were diagnosed with TB during the study in the participating laboratories were included in the study; a median of 34 (IQR 21-45) patients per lab per study month. This gave a total of 3926 patients. 3148 (80%) patients were from Rio de Janeiro and more laboratories in Rio de Janeiro switched to the Xpert test later in the trial, resulting in an imbalance between the diagnostic tests. 2546 (65%) patients were male, most were aged between 15 and 29 (2082, 53%) and were HIV negative (2126/2518, 84% of those with known status).

In this thesis, I focus on two secondary outcomes of the trial; patient outcomes and bacterial confirmation of TB.

In Chapter 8, I use the impact of the diagnostic test on TB patient outcomes as a case-study. For each patient diagnosed with TB during the study, their TB-treatment outcome was collected 15-23 months after diagnosis. A patient's outcome was defined as favourable, meaning that they successfully completed

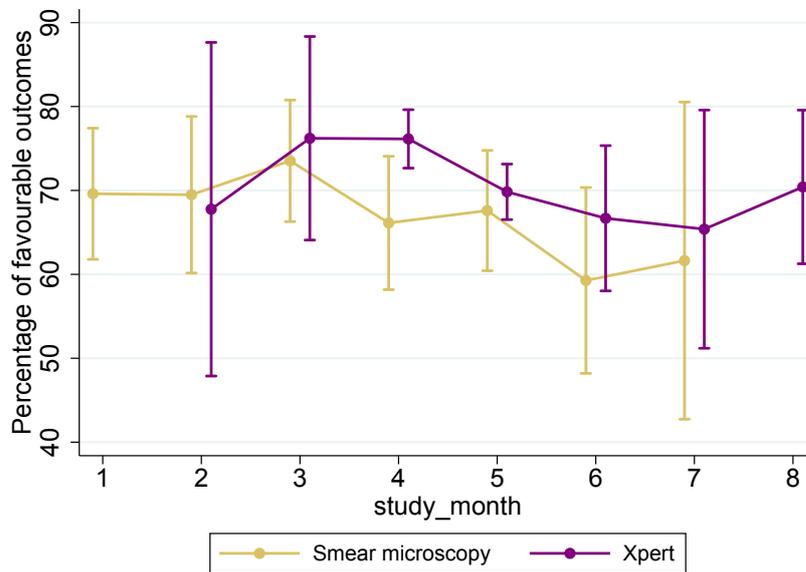
Table 2.2: TB diagnostic trial patient characteristics by diagnostic test. Taken from Trajman *et al* [10]

	Sputum smear microscopy N(%)	Xpert MTB/RIF N(%)	Total N(%)
Total	1777	2147	3924
City			
Rio de Janeiro	1618 (91%)	1528 (71%)	3146 (80%)
Manaus	159 (9%)	619 (29%)	778 (20%)
Male Sex	1140 (64%)	1405 (65%)	2545 (65%)
Age			
<15	51 (3%)	40 (2%)	91 (2%)
15-29	929 (52%)	1153 (54%)	2082 (53%)
30-59	592 (33%)	697 (32%)	1289 (33%)
≥60	205 (12%)	257 (12%)	462 (12%)
HIV status			
Negative	975 (55%)	1151 (54%)	2126 (54%)
Positive	196 (11%)	195 (9%)	391 (10%)
Unknown	606 (34%)	801 (37%)	1407 (36%)

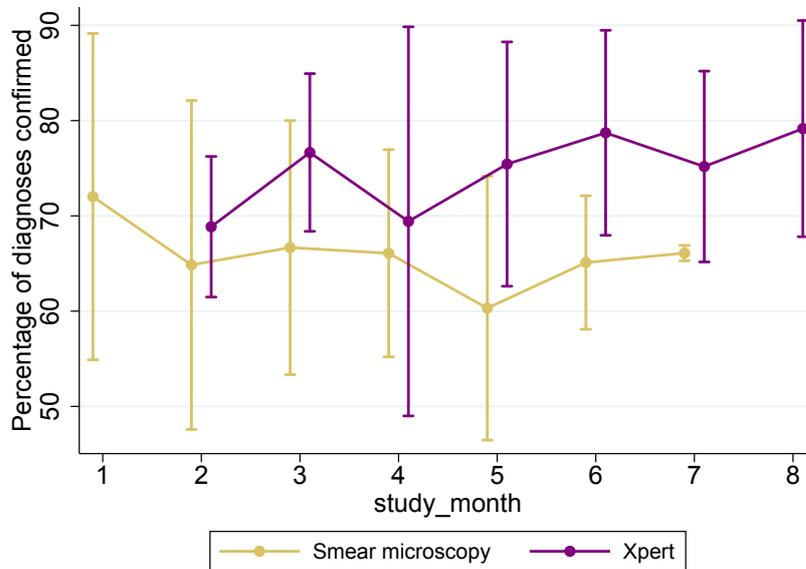
treatment without evidence of treatment failure, or unfavourable. An unfavourable outcome included death from any cause and transfer to another clinic (this included specialist clinics treating drug-resistant TB).

Trajman *et al* [10] reported that there was no evidence in this trial that the Xpert diagnostic test reduced unfavourable patient outcome (OR=0.93 95% CI 0.79-1.08) . However, their analysis did not adjust for secular trends. Figure 2.4a shows the mean and standard deviation of the percentage of patients in each laboratory with a favourable outcome in each study month. In most study months, the percentage of patients with favourable outcomes is greater for those diagnosed after their laboratory switched to the Xpert test , but the difference appears small. There appears to be a small decrease over time in the percentage with favourable outcomes.

Figure 2.4: Graph of mean \pm standard deviation of laboratory-level outcomes in each study month by diagnostic test



(a) Favourable patient outcomes



(b) Bacterial confirmation of diagnosis

In Chapter 9, I use the impact of the diagnostic test on whether patients had their TB diagnosis bacterially confirmed [10]. Each patient's diagnosis was recorded as clinical, with either a negative test or no test done, or as bacterially confirmed by the diagnostic test. Trajman *et al* [10] reported that, overall, more patients diagnosed with the Xpert test had a bacterially confirmed dia-

gnosis than those diagnosed using a sputum smear microscopy test (76.2% and 68.0% respectively). Figure 2.4b shows the mean and standard deviation of the percentage of patients in each laboratory with a bacterially confirmed diagnosis in each study month. The percentage of confirmed diagnoses is higher in those diagnosed with an Xpert test in all study-months, and there are no obvious secular trends in this outcome.

2.3 Hypothetical NHS Health-Check Trial

The final case study I will consider in this thesis is an observational dataset from the National Health Service (NHS) in England [17]. This is an example of a high-income setting where SWTs are commonly performed [18–20]. I will create hypothetical trials based on this observational data for a simulation study.

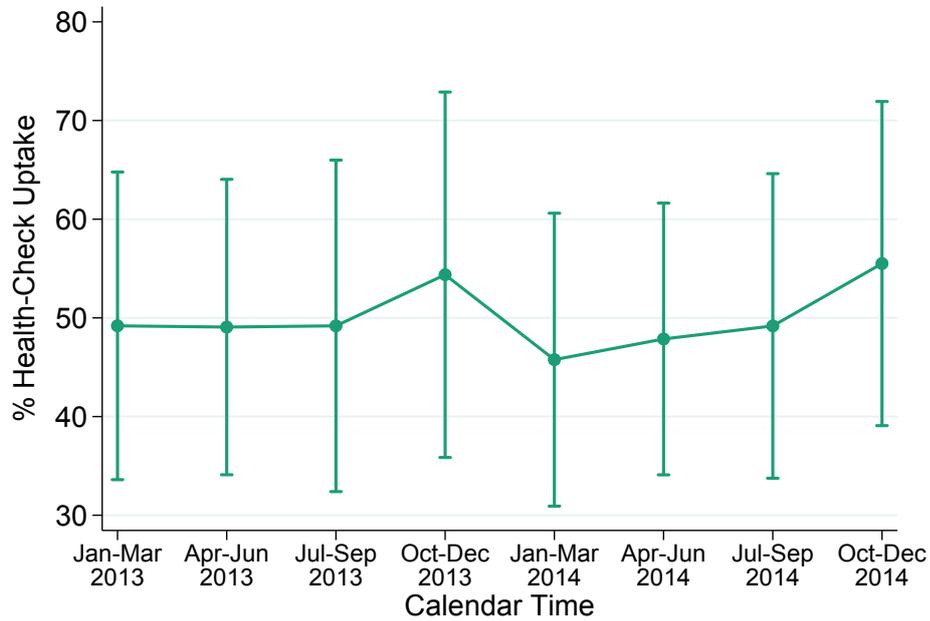
In England, the NHS offers health-checks to all adults aged 40-74 every five years. The checks are carried out by general practices and third parties, and they assess the patient’s risk of diabetes, heart disease, kidney disease, stroke, and dementia [21]. The percentage of those eligible attending a health-check increase from 5.8% between April 2009 and March 2010, to 30% between April 2012 and March 2013 [21].

Data on the uptake of health-checks are publicly available for each local authority in each quarter [17]. For each local authority in England, in each quarter, the number of patients eligible for the health-checks, offered a health-check, and that accepted a health-check are given. Unlike the paper by Robson *et al* [21], which reported the proportion of those eligible that receive a health-check, I focused on the proportion of those offered a health-check that accepted the offer. For this case study, I use data from January 2013 to December 2014.

Figure 2.5 shows the mean and standard deviation of the percentage of acceptance in each quarter. On average, 49% of patients in the first quarter of 2013 accepted the offer of a health-check; this increased to 54% in the last quarter. At the start of 2014, this dropped to 46% before increasing to 56% in the last quarter.

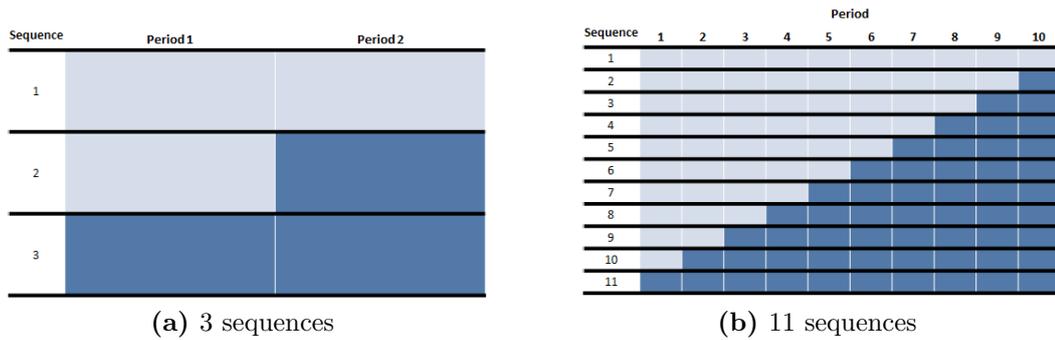
In Chapter 8, I simulate hypothetical trials that test an intervention aimed to increase the acceptance of health-checks. Two SWT designs are considered; in the first, local authorities are randomised to 3 sequences as shown in figure

Figure 2.5: Graph of mean \pm standard deviation of local authority level proportion of NHS health-check acceptance over time.



2.6a, and in the second, local authorities are randomised to 11 sequences as shown in figure 2.6b.

Figure 2.6: Schematic of hypothetical NHS health-check SWTs



2.4 Summary

In this chapter I have introduced the case studies that will be used throughout this thesis. These case studies motivated the aims of this thesis, in particular the counter intuitive results of the deworming trial. The designs used in these case studies varied substantially; the deworming study and NHS hypothetical

trials did not include periods outside rollout, while the TB diagnostic trial used the standard design. The number of sequences varied from 3 to 11. However, there are commonalities to these case studies; all three used binary outcomes. This is common in SWTs more generally [19], and so will be the focus of this PhD. I have not provided details of the analysis methods used in these case studies, or whether these were appropriate in the setting, as this will be further explored throughout this thesis.

In the next chapter, I will provide an overview of the literature on designing and analysing SWTs to give a broader context for SWTs than these case studies have been able to provide.

Bibliography

- [1] World Health Organisation. Fact sheet: Soil-transmitted helminth infections. URL: <http://www.who.int/mediacentre/factsheets/fs366/en/> (visited on 20/05/2017).
- [2] Brooker S, Miguel EA, Moulin S, Luoba AI, Bundy DA and Kremer M. Epidemiology of single and multiple species of helminth infections among school children in Busia District, Kenya. *East African Medical Journal* 2000. 77. (3):157–161.
- [3] Taylor-Robinson DC, Maayan N, Soares-Weiser K, Donegan S and Garner P. Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance. *Cochrane Database of Systematic Reviews* 2015. (7):CD000371.
- [4] Horton J. Global anthelmintic chemotherapy programs: learning from history. *Trends in Parasitology* 2003. 19. (9):405–409.
- [5] Miguel E and Kremer M. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 2004. 72. (1):159–217.
- [6] Replication data for: Worms: Identifying impacts on education and health in the presence of treatment externalities. Havard Dataverse Network V1. URL: <http://dx.doi.org/10.7910/DVN/28038> (visited on 19/11/2015).
- [7] Shadish WR, Cook TD and Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. 2nd ed. Houghton Mifflin, 2002.

- [8] Aiken AM, Davey C, Hargreaves JR and Hayes RJ. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. *International Journal of Epidemiology* 2015. 44. (5):1572–1580.
- [9] Davey C, Aiken AM, Hayes RJ and Hargreaves JR. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. *International Journal of Epidemiology* 2015. 44. (5):1581–1592.
- [10] Trajman A, Durovni B, Saraceni V, Menezes A, Cordeiro-Santos M, Cobelens F and Van den Hof S. Impact on Patients' Treatment Outcomes of XpertMTB/RIF Implementation for the Diagnosis of Tuberculosis: Follow-Up of a Stepped-Wedge Randomized Clinical Trial. *PloS One* 2015. 10. (4):e0123252.
- [11] World Health Organisation. Global Tuberculosis Report. Report. 2016.
- [12] Siddiqi K, Lambert ML and Walley J. Clinical diagnosis of smear-negative pulmonary tuberculosis in low-income countries: the current evidence. *Lancet Infectious Diseases* 2003. 3. (5):288–296.
- [13] Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC and Dendukuri N. Xpert(R) MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database of Systematic Reviews* 2014. (1):CD009593.
- [14] Theron G, Zijenah L, Chanda D, Clowes P, Rachow A, Lesosky M, Bara W, Mungofa S, Pai M, Hoelscher M, Dowdy D, Pym A, Mwaba P, Mason P, Peter J, Dheda K and team TN. Feasibility, accuracy, and clinical effect of point-of-care Xpert MTB/RIF testing for tuberculosis in primary-care settings in Africa: a multicentre, randomised, controlled trial. *Lancet* 2014. 383. (9915):424–435.
- [15] Churchyard GJ, Stevens WS, Mametja LD, McCarthy KM, Chihota V, Nicol MP, Erasmus LK, Ndjeka NO, Mvusi L, Vassall A, Sinanovic E, Cox HS, Dye C, Grant AD and Fielding KL. Xpert MTB/RIF versus sputum microscopy as the initial diagnostic test for tuberculosis: a cluster-randomised trial embedded in South African roll-out of Xpert MTB/RIF. *Lancet Global Health* 2015. 3. (8):e450–457.

- [16] Durovni B, Saraceni V, Hof S van den, Trajman A, Cordeiro-Santos M, Cavalcante S, Menezes A and Cobelens F. Impact of replacing smear microscopy with Xpert MTB/RIF for diagnosing tuberculosis in Brazil: a stepped-wedge cluster-randomized trial. *PLoS Medicine* 2014. 11. (12):e1001766.
- [17] Public Health England. Explore NHS Health Check Data. URL: http://www.healthcheck.nhs.uk/commissioners_and_providers/data/ (visited on 20/08/2015).
- [18] Barker D, McElduff P, D'Este C and Campbell MJ. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. *BMC Medical Research Methodology* 2016. 16. (1):69.
- [19] Martin J, Taljaard M, Girling A and 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.
- [20] Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, Thompson JA, Fielding KL, Omar RZ, Ononge S, Hargreaves J and Prost A. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 2015. 16. (1):353.
- [21] Robson J, Dostal I, Sheikh A, Eldridge S, Madurasinghe V, Griffiths C, Coupland C and Hippisley-Cox J. The NHS Health Check in England: an evaluation of the first 4 years. *BMJ Open* 2016. 6. (1):e008840.

3 Current Recommendations on Design and Analysis

The literature on the methodology of SWTs is in its infancy, meaning that many topics have yet to be explored. This is in part because SWTs cover a large range of designs.

The first section of this chapter will introduce the breadth of designs covered by the term “stepped-wedge trial”: the designs can vary in how clusters are randomised and the rollout of the intervention, as well as how individuals within clusters are observed, and in the effect of the intervention. Much of this section is based on a series of articles that I was involved in at the start of this PhD [1–6]. The series focused on a systematic review of SWTs published between 1987 (when the first SWT was conducted) and 2010 [1]. My main involvement was in exploring the designs used by the SWTs identified in the review including issues with these designs [2], and the analyses they used highlighting key areas for consideration in analysis [3]. This first section primarily focuses on the review of common designs.

The remaining sections describe the current state of the literature on SWTs. Section 3.2 describes methodology literature on analysing SWTs, including which methods are used in practice [3, 7–9]. Finally, section 3.3 details the methodology literature on sample size calculations for SWTs and the implication this has for efficient trial design. These final two sections are focused on literature identified through writing the papers in chapters 6 to 9.

3.1 The Breadth of Stepped-Wedge Trial Designs

An issue with SWT methodology is that the term covers a wide range of designs (figure 3.1). The systematic review, which I co-authored at the start of this PhD explored this variation [2]. In this section I will highlight the considerations relevant to this thesis.

There are many ways in which clusters can be randomised, for example, the number of sequences can vary, and the time between switch-points can vary between trials. Even once the randomisation and allocation of the intervention has been determined, there is variability in how observations are made of the individuals. Many of these variations exist in other cluster randomised trial designs, but the impact is more important in SWTs because of the confounding of the intervention effect with time. This section will describe the range of designs possible, the impact of such design choices, and the focus of this PhD.

3.1.1 Cluster-Level Designs

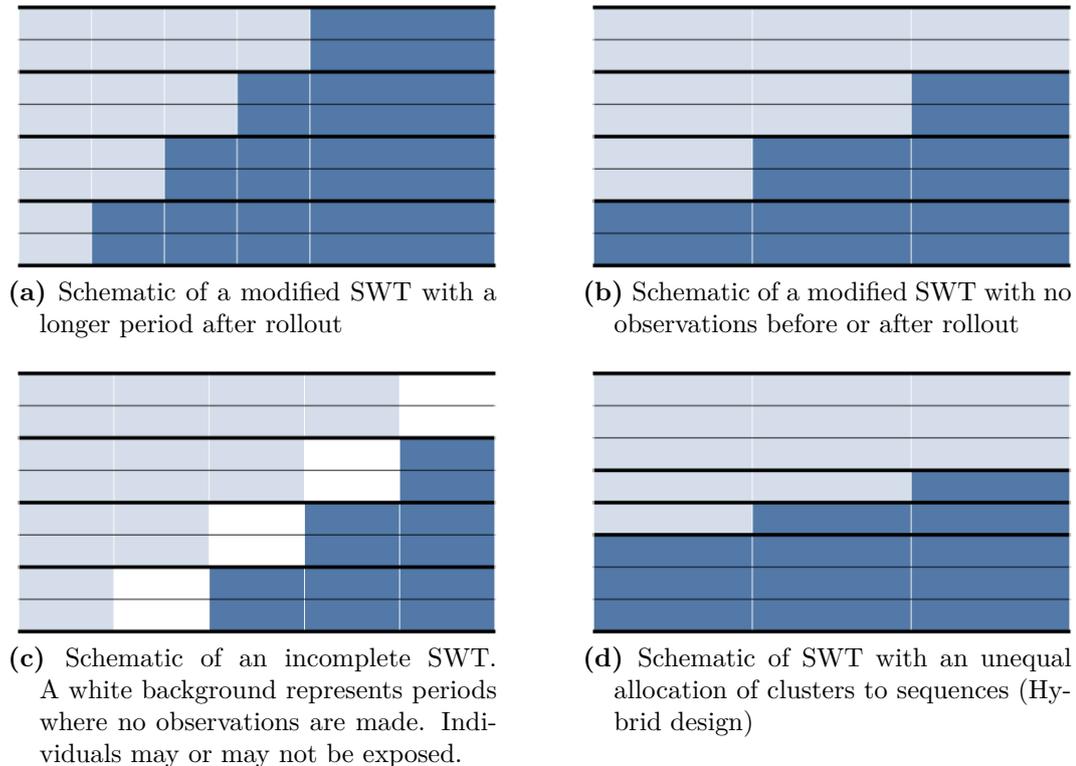
The most commonly used design, which I will refer to as the standard design, includes a period before rollout and a period after rollout (figure 1.1). The periods of the trial are all of equal size with the same number of observations in each, and there are the same number of clusters in each sequence. Systematic reviews have shown that most SWTs have a median of 17 clusters [1, 7], randomised to a median of 4 sequences (IQR 2 - 6) [1, 7, 8].

Other SWT designs include longer periods before or after rollout (figure 3.1a) [10], and others have excluded the periods before and after rollout altogether (figure 3.1b) [11]. Martin *et al* [7] found approximately 34% of SWTs included extended periods before or after rollout.

In non-standard SWT designs, the number of observations in each period or in each cluster may vary. Usually, this is because the number is not controlled by the trialist, such as in the TB diagnostic case-study described in chapter 2 where the number of observations in each period was determined by the number of patients diagnosed with TB in each laboratory and not by the trialists. This meant that the number varied, both between periods and between laboratories. Few studies report that their cluster size varied, but this may be due to poor reporting rather than low prevalence of such designs [7].

In incomplete designs, there are some periods where no observations are collected. Reasons to use incomplete designs include not collecting observations in the period immediately following a cluster switching to the intervention to allow time for the cluster to transition (figure 3.1c) [12], or improving the efficiency of the design by optimising when observations are collected [13]. These designs are unusual in practice [7].

Another way SWTs can vary is in how they allocate clusters to sequences.

Figure 3.1: Schematics of modified stepped-wedge trials

Rather than uniform allocation, SWTs can be designed to have more clusters randomised to some sequences than to others. An example of such a design is called a Hybrid design (figure 3.1d) [14]. This is an SWT with no observations before or after rollout and with the first and final periods half the size of the remaining periods. The number of clusters randomised to the sequences that remain in the control or intervention for the duration of the trial is allowed to differ to the number of clusters randomised to the other sequences of the design. This is often described as a combination of an SWT and a CRT, where some proportion of the clusters are randomised to an SWT with the periods before and after rollout half the size of the remaining periods, and the remaining clusters randomised to a CRT. To my knowledge, this design has not yet been implemented in practice.

Because the range of SWT designs is so large, I limited the scope of designs considered in this thesis. This thesis will only consider complete SWTs with uniform allocation of clusters to sequences. However, I will consider designs with no observations outside rollout as well as designs with periods before and after rollout. SWTs comparing more than one intervention condition have also been suggested [15, 16], but I will only consider designs with a control and

one intervention condition. In Chapter 6, I have assumed an equal number of observations in each cluster in each period, but in Chapters 7-9, the number of observations varies between periods and clusters.

3.1.2 Participants and Observations

In a cluster randomised trial, each cluster consists of a population of participants. How those participants are exposed, sampled and observed will vary from trial to trial. In this section, I will discuss how participant sampling and observation can vary and why this is particularly important in SWTs. Lastly, I will describe the types of sampling and observation that this PhD will focus on.

The data type of observations collected has implications for the analysis. The most common are continuous outcomes such as a test score, binary outcomes such as school attendance, and rate or time-to-event outcomes such as time to infection [7]. The most common type for SWTs is binary [7].

SWTs require that observations of the clusters are collected during each period of the trial (or most periods for an incomplete design). These could either be collected from the same individuals observed multiple times during the study (a cohort design), or from different individuals each observed only once during the study (a repeated cross-sectional design) [17].

These two broad categories can be further refined.

Cohort designs can be divided into open and closed cohorts [2]. This distinction is usually determined by the study setting and not by the trialist.

In a closed cohort, individuals are sampled at the start of the trial and observed multiple time through to the end of the trial: once the sample has been selected, no one can join it (figure 3.2a). An example is a year-long trial of the effect of free school breakfasts on school attendance [18]. In this study, children that joined the participating schools during the study did not have their attendance monitored so did not join the study cohort (but were still able to receive free school breakfasts). Only a few children were lost to follow up, so the same children were observed in each period.

In an open cohort, participants are able to join the sample during the trial (figure 3.2b). For example, a trial of preventive TB therapy given to people living with HIV and attending an HIV clinic [19]. In this study, clinic patients

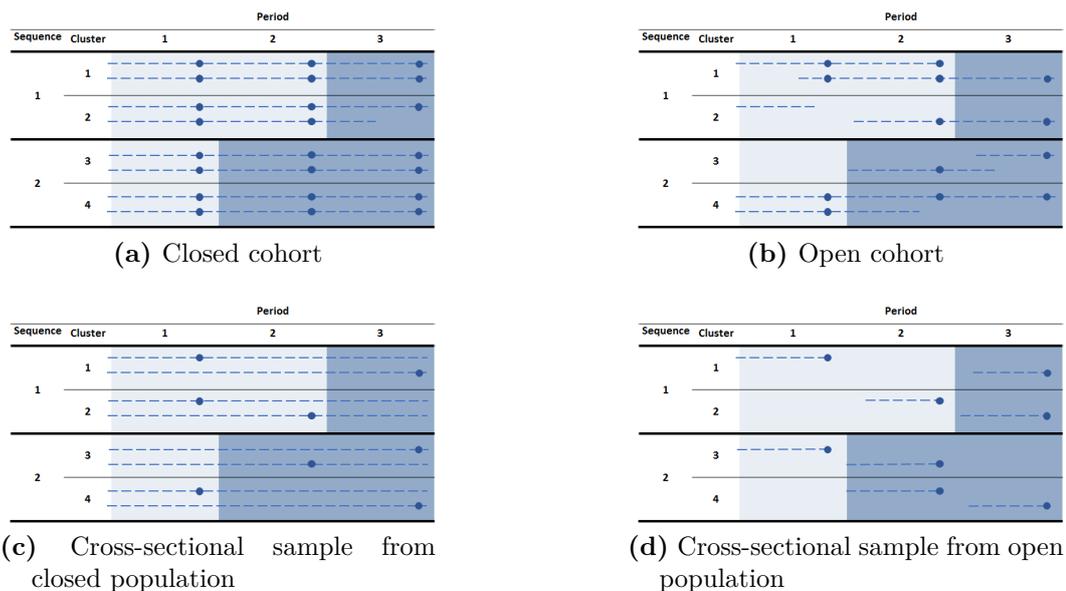
joined the study when they visited the clinic and left at their final visit during the study: this meant that the study sample changed throughout the trial. In both types of cohort design, some individuals may leave the sample during the study, but this would be more common in an open cohort.

Similarly, repeated cross-sectional designs can be classified based on the populations from which the cross-sectional samples are taken: again, this will usually be determined by the setting and not by the trialist.

In some trials, the population is largely closed to migration into and out of the cluster and so cross-sections are from the same population of individuals in each period (figure 3.2c). An example of repeated cross-sectional samples of a closed population is a trial which sent postal surveys to a sample of members of the clusters in each period of the trial, so that different residents were observed in each period [20].

Alternatively, some populations have larger levels of migration into and out of the cluster and so the cross-sections are from a different population of individuals in each period (figure 3.2d). An example is sampling women in a maternity ward over a 10 month trial [21].

Figure 3.2: Diagram of trial participant exposure and observation. Background: light is control condition, darker is intervention condition. Dashed lines represent the time a participant is exposed. Dots represent observations of individuals



There has been disagreement about how common each type of design is; most

reviews have found that between 35% and 55% use cross-sectional designs [1, 8, 9], but one review only identified 16% as cross-sectional designs [7].

Whilst these forms of sampling and observation are present in all cluster randomised trials, the distinction is particularly important in an SWT because of the confounding of the intervention effect with time and because clusters are exposed to both conditions.

Confounding with time means that changes to the sample over time could bias the intervention effect estimate if confounding factors become imbalanced between the control and intervention conditions [2]. In a cohort design, this could be caused by non-random attrition of the sample. Conversely, samples taken from a largely closed population, with few individuals joining or leaving the clusters during the trial, are likely to be similar at the start and end of the trial.

In trials conducted in closed populations where individuals are exposed to both the control and intervention condition, the intervention effect could be confounded with changes in the individuals' health [2]. An example is a trial aiming to improve diabetes control of a cohort of patients [22]; patients who receive the intervention later in the trial may have been in a worse state of diabetes control at the time they switch than those that received the intervention earlier, which could alter the effect of the intervention.

In this thesis, I focus on repeated cross-sectional sampling, in both open and closed populations, because these designs are more common and are simpler to analyse as there is only one level of clustering to account for in the analysis. Throughout this thesis, I allow for secular trends in the outcome, but assume that there are no other confounding factors and that changes in the clusters' populations do not affect the intervention effect.

3.1.3 The Intervention Effect

The final aspect of trials design that is particularly important in SWTs is the intervention effect and whether it changes over time or between clusters.

The aim of RCTs is to estimate the effect of an intervention. In most RCTs, the outcomes under each condition are compared after a fixed amount of time and exposure to the intervention. For example, a trial in Scotland studied the impact of an education programme on sexual activity [23]. The trial randomised 25 secondary schools to either receive the programme or not and measured the

outcome in a cohort of pupils that were surveyed 2 years after the trial began. In most SWTs, this is not the case; instead, observations are collected with different lengths of exposure to the intervention. Because of this, assumptions have to be made about how the intervention effect is expected to change over time.

One way in which the intervention effect may change over time is if there is a lag to its effectiveness after the clusters switch, similar to the concept of carry over in a crossover trial [2]. This could either be a lag in implementing the intervention, or it could be that some time is needed after implementation for it to have an effect. Trials where individuals are exposed to both control and intervention conditions are more likely to be affected [2].

The second way in which the intervention effect may change is if the effect wanes after some period of time. Some interventions, such as a vaccine, have long lasting effects after an individual is exposed. Other interventions, such as educational interventions, require ongoing exposure to have an effect and so the effect is more likely to wane with time since exposure.

Settings where waning of the intervention effect is not an issue allow for designs where individuals are only exposed for a short period of time [2]. Observation of an individual can occur some time after exposure as long as the individuals are no longer subject to the clusters' exposure. An examples of such a design is the Gambia trial [24] that is measuring the effect of a hepatitis vaccine (the intervention) on incidence of liver disease in the 40 years following childhood vaccination. Such designs appear very different to designs where individuals are constantly exposed within the cluster and are observed within the time frame of the intervention rollout, but in terms of analysis they are similar. Alternatively, in setting where the effect of the intervention is only temporary, i.e. the effect will wane very quickly, SWTs can explore the removal of an intervention, with clusters moving from the intervention condition to the control condition [25].

Lastly, the intervention may have a differing effect in each cluster. This could be a result of differences in the clusters, leading to a greater or smaller susceptibility to the intervention, or in the implementation of the intervention. This phenomenon has been noted in CRTs [26, 27], and it is more likely in SWTs where the intervention is implemented at different calendar times in each sequence of clusters [4]. This is separate to confounding of the intervention effect with secular trends, which affect the control and intervention

observation equally.

Careful consideration of the trial design can safeguard against some of these issues. Longer periods give more time for the intervention to become effective, but may result in a lengthier trial and so a higher chance of a waning effect [2]. The use of incomplete designs, where no observations are collected in the period after a cluster switches, are a solution to lag. Differential intervention effects can be controlled to some extent through a carefully designed intervention defined by a strict protocol, although this will not be appropriate in every setting [4].

In the remainder of this thesis, I will not consider the complexities of lag or waning intervention effects and will instead focus on a simple case that assumes that the intervention is fully realised immediately after implementation, and remains the same throughout the trial. This is a common assumption in both methodology papers and trial analysis [8]. The implications of this simplification will be explored in the discussion. I will however explore the impact on analysis of the intervention effect varying between clusters.

3.2 Analysis Methods

In this section I will describe the current literature on the analysis of SWTs.

In the simple case of cross-sectional designs, the analysis must consider two main issues beyond those of other RCTs: secular trends and clustering.

In Chapter 1, I introduced the concepts of vertical and horizontal comparisons [3, 28]. To date, the majority of analysis methods discussed in the literature have incorporated both types of comparisons. This gives maximum power to detect an effect, but inclusion of the horizontal comparisons means that period effects and correlations over time must be appropriately considered.

Clustering can be accounted for in cluster randomised trials through mixed-effect modelling, generalised estimating equations (GEE), or by analysing at the level of the cluster [17].

Most literature for SWTs has focused on mixed-effect models, but some literature has described using GEE. Analysing at the cluster level has been less extensively explored for SWTs. In subsection 3.2.2, I will detail the most commonly used analysis model for SWTs and problems with this model. In

subsections 3.2.3 and 3.2.4 I will describe modifications that have been suggested for this analysis model. Subsections 3.2.5 and 3.2.6 describe the literature on GEE and cluster level analysis respectively. Subsection 3.2.7 described the current literature on analyses that only utilise vertical comparisons to avoid confounding the intervention with time. Subsection 3.2.8 will describe the ease of implementation of all the discussed methods in commonly used software. I have focused on literature relating to robust estimation of a singular intervention effect. There are several other areas of methodological research that I have not detailed, such as group-sequential designs [29], estimating complier average causal effects [30], and analysis methods for trials with more than one intervention [16].

3.2.1 Properties of Analysis Methods

Before describing the analysis methods used for SWT, I will first describe the characteristics of an analysis which are desirable.

First, the analysis method should provide an unbiased estimate of the parameter of interest. That is

$$E(\hat{\theta}) = \theta_A$$

Second, the analysis method should provide standard error estimates that appropriately reflect the uncertainty around the estimate. This can be seen through the confidence interval coverage and the type-one error rate.

A 95% confidence interval is defined as an interval which we would expect to contain the true effect in 95% of repetitions of a study [31]. Therefore, we would expect 95% of the simulated confidence intervals to contain the true parameter effect. When this is not the case, either the parameter estimates are biased, the parameter standard-errors are biased, or the test statistic is being compared to an inappropriate distribution.

The type-one error rate is related to this: it is defined as the probability of a $p - value < \alpha$ when the parameter effect is truly null and is 1 - the coverage of $(1 - \alpha) \%$ confidence intervals when the parameter is simulated to have no effect [31]. It is common to use $\alpha = 0.05$.

Lastly, the analysis method should be efficient. From a choice of unbiased estimates, the one with the smallest standard error will be preferred. The power of the method is one way to assess the efficiency of an analysis method.

The power is the probability of detecting an effect with $p < \alpha$ when there is truly an effect present [31].

In the rest of this section I will describe the literature comparing these properties for different analysis methods for SWTs.

3.2.2 Hussey and Hughes Model

In their simplest form, mixed-effect models account for clustering by partitioning the error term into a cluster-level error and a within-cluster error by giving each cluster its own intercept term and assuming that these follow a normal distribution [17]. The error structure can be extended to account for more complex correlations within the data by, for example, introducing random slopes for covariates in the data. In general, mixed-effect models have been found to be sensitive to the correlation structure being correctly specified [32].

Hussey and Hughes published the first paper describing the analysis of SWTs [33]. This model is now widely used [3, 8] and is the basis of the first design effect for an SWT (see section 3.3). They suggested using the following linear mixed-effect model (LMM) for continuous outcomes, herein referred to as the Hussey and Hughes model, or the standard model:

$$y_{ijk} = \mu + \beta_j + \theta X_{ij} + u_i + e_{ijk} \quad (3.1)$$

where y_{ijk} is the outcome in cluster $i = 1, \dots, I$, in period $j = 1, \dots, J$, for individual $k = 1, \dots, K$, β_j is a period effect comparing period j to the first period with $\beta_1 = 0$, θ is the intervention effect, X_{ij} is 1 if cluster i received the intervention in period j and 0 otherwise, $u_i \sim N(0, \sigma_u^2)$ is a random effect for cluster, and $e_{ijk} \sim N(0, \sigma_e^2)$ is the within-cluster, individual-level error and $\sigma_u^2 + \sigma_e^2 = \sigma^2$.

This can be extended to binary and count outcomes through generalised linear mixed-effect models (GLMM). With a binary outcome and a logit link, this results in the model estimating cluster-specific estimates, not population-average estimates [34].

The model includes a random effect for cluster to account for the correlation of observations in the same cluster. It incorporates horizontal and vertical comparisons [28, 35].

The horizontal comparisons are adjusted for secular trends (period effects) by

the inclusion of a fixed, categorical variable for period. This leads to strong assumptions about the correlation structure of the data. Because time is treated as a fixed effect, the same change over time is assumed for all clusters and all observations within a cluster are assumed to be equally correlated whether they are from the same period or different periods (this is also known as exchangeability within clusters). The intervention effect is also assumed to be common to all clusters, which again means that observations within a cluster are equally correlated regardless of whether they are from the same or different conditions. Although Hussey and Hughes suggest using this model, their paper gives no limitations in the appropriate use of the model.

Matthews and Forbes [28] provide a greater insight into how this analysis model estimates an intervention effect. The analysis model takes a weighted average of the vertical and horizontal comparisons, with weights depending on the cluster-mean correlation. The model gives a different weight to each cluster-period cell of the trial, with some cluster-periods contributing very little to the analysis.

At the start of this PhD in 2014, there was very little literature assessing whether this widely used analysis model was appropriate, given the known sensitivity of mixed-effect models to misspecified random effects [36]. In the paper I co-authored with Davey *et al* [3], we highlighted the importance of ensuring that there was adequate adjustment of period effects, and questioned whether a fixed categorical variable would be sufficient in all situations.

In 2015, the Ebola outbreak gave rise to the need for pragmatic, efficient trials to assess the efficacy of new vaccines developed in response to the outbreak. Several groups gave strong consideration to using an SWT, but none selected the SWT design [37, 38], and some expressed concerns about the use of SWTs in this setting [39, 40]. Bellan *et al* [40] performed a simulation study to calculate the power and type-one error rate of several trial designs in a simulated Ebola vaccine trial in Sierra Leone. They found that analysis using a Cox model with frailty resulted in inflated type-one error. This was because of the spatial temporal trends in the outcome seen in this pandemic setting [40]: the incidence of Ebola changed differently over time in each district because infections spread from one place to another, violating the assumption of a common period effect. Conversely, they saw <5% bias in the intervention effect estimate. However, a pandemic setting is not representative of the settings that usually use SWTs, which may have less variability in the secular trends between clusters.

More recently, two papers have assessed this issue in other settings. Ji *et al* [41] and Wang and DeGruttola [42] both assessed the results of the model 3.1 in a more common setting when the period effect varied between the clusters through simulation studies. They found that model 3.1 gave inflated type-one error rates when the period effects varied between the clusters [41, 42].

Even when the assumptions of this model are met, mixed-effect models are known to have inflated type-one error rates when the number of clusters is small unless small sample corrections are used [43]. This has started to be explored for SWTs; Barker *et al* [44] found that model 3.1 has inflated type-one errors when used with six or fewer clusters, and Zhan *et al* [45] found biased cluster-level variance estimates with a total of 11 clusters using a Poisson GLMM.

Despite these warning, many trialists continue to pay little attention to the period effects in their analyses with at most fixed effects for periods included in their analysis model [8]. Alternative analysis methods are needed that maintain an appropriate type-one error rate in the presence of period effects that differ between clusters.

In light of these findings several adaptations of this model have been considered, but none are commonly used [8]. The next two subsections will outline suggested adaptations to model 3.1.

3.2.3 Non-Parametric Methods

Since research to date has found that the intervention effect estimate from model 3.1 is unbiased [40–42], and only the type-one error is affected, several authors have suggested the use of non-parametric inference with the intervention effect estimate from model 3.1 [40–42]. Two such non-parametric methods are bootstraps and permutation tests.

Bootstraps

Bootstrapping is where p-values and confidence intervals are calculated by repeatedly resampling with replacement from the observed data. The observed data are taken to be representative of the population from which they are sampled. By resampling from the observed data with replacement, one obtains a different sample that remains representative of the population. Repeating

this process many times will give a distribution of the intervention effect in the population [46].

To account for clustering in cluster randomised trials when using the bootstrap, the clusters are resampled. A second stage of resampling individuals within clusters has been shown to improve the confidence interval coverage with a small number of clusters [47].

Bellan *et al* [40] assessed the use of bootstraps with the Cox model in SWTs. Although bootstraps are widely thought to give very robust inference [46], Bellan *et al* [40] found that this method gave inflated type-one error; the authors conclude that this is because they only had 20 clusters in the simulation study. Bootstraps rely on asymptotic theory so require a sufficiently large sample to maintain favourable properties [46]. Most SWTs have fewer than 20 clusters [1, 7], so this prevents bootstraps being widely useful to correct the type-one error rate of model 3.1 unless refined methods are identified.

Permutation tests

A permutation test works on the premise that if the intervention has no effect on the outcome, the assignment of observations to control and intervention is arbitrary. For any assignment of observations to control and intervention, we can calculate an intervention effect. If we repeat this many times, each time collecting the intervention effect for that assignment, this gives a distribution of the intervention effect under a null hypothesis that the intervention has no effect. The proportion of assignments that gave an intervention effect the same or more extreme than the observed intervention effect is the p-value.

The permutation test only requires that two conditions hold: exchangeability of observations, and a strong null hypothesis [46].

Exchangeability means that any assignment of the observations to control or intervention is equally likely. For SWTs, this means that assignment of clusters to sequences should be permuted (reassigned), rather than simply permuting the individual observations.

A strong null hypothesis tests that the intervention has no effect on any observations, as opposed to a weak hypothesis that tests that there is no effect on average. This means that the intervention effect cannot vary within, or between the clusters, which also implies that observations must be equally

correlated regardless of whether they were from the control or intervention condition, as is assumed by model 3.1.

Whilst sensitivity of permutation tests to the strong null hypothesis assumption has not been assessed for SWTs, Gail *et al* [48] found that permutation tests tends to be relatively robust under a weak null hypothesis unless there is an imbalance in the number of clusters in each arm and the smaller arm has a larger variance.

Several authors [40–42] have suggested the use of permutation tests with model 3.1 and found that this gave a correct type-one error rate in the scenarios they considered.

3.2.4 Other Mixed-Effect Models

Whilst some researchers have investigated non-parametric inference, others have considered parametric modifications to model 3.1. Hemming *et al* [49] gives an overview of many of these modifications. Here, all models are written as LMM, but can be extended to GLMM for non-normally distributed outcomes.

Fixed period effects that vary within strata

Hemming *et al* [49] suggest fitting different period effects within some strata of clusters. This relaxed the assumption of a common period effect, but still requires that the period effects are common to all clusters within the defined strata. This model would also still assume exchangeability of observations within clusters. Moreover, identification of appropriate strata would be difficult in practice.

A Cluster-period interaction model

Several authors have suggested including a random effect for a cluster-period interaction will relax the assumptions of a common period-effect and exchangeability of observations within clusters [12, 14, 35, 49, 50]. The suggested model is as follows, herein referred to as the cluster-period interaction model:

$$y_{ijk} = \mu + \beta_j + \theta X_{ij} + u_i + q_{ij} + e_{ijk} \quad (3.2)$$

where $u_i \sim N(0, \sigma_u^2)$, $q_{ij} \sim N(0, \sigma_q^2)$ is a random-effect for each cluster-period, and $e_{ijk} \sim N(0, \sigma_e^2)$. The random effects are independent. In this model, observations in different periods are less correlated than observations in the same period and period effects are allowed to vary between clusters. The total variance is fixed to be the same in all periods of the trial. The ICC is now the correlation of two observations in the same cluster in the same period:

$$\rho^* = \frac{\sigma_u^2 + \sigma_q^2}{\sigma_u^2 + \sigma_q^2 + \sigma_e^2}$$

and the cluster-level autocorrelation, describing the correlation between observations in the same cluster but different periods, is:

$$\pi = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_q^2}$$

The cluster-level autocorrelation is the same for all pairs of periods, regardless of whether they are close in time to one another, or further apart.

The model has also been extended to allow for cohort designs [14, 50]:

$$y_{ijk} = \mu + \beta_j + \theta X_{ij} + u_i + q_{ij} + d_{ik} + e_{ijk} \quad (3.3)$$

where $d_{ik} \sim N(0, \sigma_d^2)$ is a random effect for individuals, and now e_{ijk} is a random effect for an individual in a specific period.

Despite several papers suggesting this model [14, 35, 49, 50], no literature has assessed whether it had better properties than model 3.1 for SWTs.

While there is a lack of advice available on whether a random effect for a cluster-period is necessary or sufficient for the analysis of SWTs, there is a more definitive answer for a CRXO, and a CRT with baseline observations. For these trial designs, mixed-effect models should include a random effect for cluster-period interactions, as well as for clusters [51, 52]. Turner *et al* [51] and Morgan *et al* [52] ran simulation studies comparing analysis methods for a CRXO. Morgan *et al* [52] directly compared model 3.1 and model 3.2, and found that model 3.2 had a type-one error rate closer to 5% than model 3.1. Despite advice from the the methodological literature, uptake of of the cluster-period interaction model is poor for CRXO designs [53]. In the field of CRTs with baseline observations, similar results have been found: Ukoumunne and Thompson [27] found that a model similar to model 3.1 gave invalid results. However, these designs only have 2 periods. SWTs can have many more periods

than this, and it may be unrealistic to assume the same correlation between all periods.

Random-period model

An alternative parameterisation fits a random effect for each period, herein referred to as the random-period model:

$$y_{ijk} = \mu + \beta_j + v_{ij} + \theta X_{ij} + u_i + e_{ijk} \quad (3.4)$$

where $v_{ij} \sim MVN(0, \Lambda_v)$ where Λ_v is a $(J - 1) \times (J - 1)$ covariance matrix with $\Lambda_{v11} = 0$, and v_{ij} and u_i may be correlated. This is modified from a version given by Heuvel *et al* [54], which treated the period effect as linear. This is a very flexible model that allows the total variability to change in each period, and allows the cluster-level autocorrelation to be different for each pair of periods. However, this model will grow in complexity as the number of periods increases, unless a linear period effect is assumed.

Heuvel *et al* [54] assessed this model in a simulation study, and found slightly inflated type-one error. But, the simulation study used a specific scenario with a small sample and assumed a linear time effect, so the results are not easily generalisable.

Intervention effect lag and wane

Heuvel *et al* [54] also suggested a model that allowed the intervention effect to increase with time in the intervention:

$$y_{ijk} = \mu + \beta t_j + \theta X_{ij} (t_j - S_i) + u_i + e_{ijk} \quad (3.5)$$

where t_j is the time from the first period to period j , S_i is the time at which cluster i switched to the intervention. The period effect is treated as continuous, and the intervention effect is assumed to increase with time in the intervention. This is one way of dealing with the lag in the intervention discussed in section 3.1.3. A more flexible approach to modelling change to the intervention over time, would be to have an interaction term between the intervention effect and time since switch [3]:

$$y_{ijk} = \mu + \beta t_j + \theta X_{ij} + \theta' X_{ij} (t_j - S_i) + u_i + e_{ijk}$$

This model now allows the intervention effect to increase or decrease with time in the intervention, allowing for a waning of effect over time, in which case θ' would be negative, or an increase in the intervention with time in the intervention, in which case θ' would be positive. Hemming *et al* [49] suggest a similar model with an interaction with period to allow the intervention effect to vary between the periods.

Another option is to treat the intervention effect as fractional, so that in the first period after the intervention is rolled out the intervention is assumed to be at half efficacy and for all later periods it is at full efficacy [33], or excluding observations from the period following a cluster switching to the intervention. However, the intervention effect estimate might be sensitive to the proportion of efficacy assigned to each period and this is difficult to know in advance. In addition, excluding observations is inefficient.

Other fixed-effect parameterisations can be used to explore different aspects of the intervention effect [49, 55].

Fixed intervention effects that vary within strata

Hemming *et al* [49] suggest modelling different intervention effects within specified strata of clusters. This relaxes the assumption of model 3.1 of a common intervention effect, but continues to assume equal correlation of observations with clusters. In addition, it may be difficult in its practice to identify strata where the intervention effect is likely to differ. This may be useful as a sensitivity analysis, but may have limited use as a primary, prespecified analysis.

Random-intervention model

All analysis methods mentioned so far have assumed that the intervention effect is common to all clusters and that observations are exchangeable between the control and intervention conditions. An alternative model that allows the intervention effect to vary between clusters, herein referred to as the random-intervention model is:

$$y_{ijk} = \mu + \beta_j + (\theta + z_i) X_{ij} + u_i + e_{ijk} \quad (3.6)$$

where $z_i \sim N(0, \sigma_z^2)$ and z_i and u_i may be correlated [56]. This is a reparameterisation of the model suggested by Hemming *et al* [49]:

$$y_{ijk} = \mu + \beta_j + \theta X_{ij} + u_{0i}^* X_{ij} + u_{1i}^* (1 - X_{ij}) + e_{ijk}$$

where $u_{0i} \sim N(0, \sigma_{u_0}^2)$ and $u_{1i} \sim N(0, \sigma_{u_1}^2)$ are random effects for clusters while in the control and intervention conditions respectively, which may be correlated.

In CRTs, ignoring variability in the intervention effect between clusters has been shown to bias the intervention effect estimate and underestimate the effect's standard error [57].

Robust variance and fixed cluster effects

Other modifications to model 3.1 include using a robust variance estimate, and treating cluster effects as fixed. Hussey and Hughes suggested using the jackknife variance estimator with their analysis [33]. Moulton *et al* [58] suggested using a robust variance estimator with the Cox model instead of a random effect for cluster: since the Cox model conditions on time, this model should only incorporate vertical comparisons.

It should also be possible to account for clustering using fixed effects rather than random effects because the intervention effect can be estimated within clusters (as horizontal comparisons). This has performed well in literature to date in terms of estimating an intervention effect and producing valid confidence intervals [44]. However, this method implies that the clusters included in the study are the only clusters relevant to the research question, and so estimates the intervention effect with greater precision than is appropriate for generalising the intervention effect to a larger population of clusters [59].

3.2.5 Generalised Estimating Equations

Models that utilise GEE calculate the mean effect of the covariates across all individuals, also known as the population-average effects, treating the correlation structure of the data as nuisance parameters. GEE have been shown to be robust to misspecification of the correlation structure [60], although this has not been assessed in the context of SWTs.

Only 2/10 studies reporting SWT results from 2010 to 2014 used GEE [1]. This is often because SWTs have few clusters, and the robust variance estimator commonly used in conjunction with GEE is anti-conservative with small samples [44, 61]. Scott *et al* [61] have suggested using small sample correction techniques or permutation tests to account for this problem; both techniques gave appropriate type-one error rates with as few as 10 clusters.

GEE incorporate horizontal and vertical comparisons so they must adjust for time. Scott *et al* [61] used the following analysis model:

$$y_{ijk} = \mu + \beta_j + \theta X_{ij} + \varepsilon_{ijk}$$

where ε_{ijk} are the total residuals that are assumed to have zero means given a set of covariates and follow a working correlation matrix.

There is no published advice on choosing a working correlation structure in the setting of SWTs. Scott uses an autoregressive one structure [61], but this may not be appropriate in other settings. The use of non-diagonal correlation matrices can lead to biased effect estimates with time varying covariates [62].

3.2.6 Cluster-Level Analysis

Cluster-level analyses are used in CRTs as a simple, straightforward way of accounting for clustering. They are used in trials with a small number of clusters where mixed-effect models and GEE are inappropriate [43, 60]. They have the additional benefit of allowing the simple calculation of the more epidemiologically relevant measure of the risk difference for binary outcomes [63, 64].

In the case of a CRT, the outcome is summarised at the cluster level: this could be risks, rates, or the mean of the outcome in each cluster. These cluster summaries can then be analysed using a standard two sample t-test to calculate the difference between the intervention and control outcomes [17].

Concerns have been raised about whether estimates of risk ratio and odds ratios are unbiased using this method. Ukoumunne *et al* [65] found that in scenarios with a small number of observations per cluster and a high ICC the cluster-level analysis produced biased estimates of the marginal risk ratio and odds ratio true effects. They hypothesised that there were 2 reasons for these biases. The bias with a small number of observations per cluster was due to the need for

a heuristic adjustment when there were no cases (or controls for odds ratios) in a cluster, a common adjustment is to add 0.5 to the number of cases and controls to such clusters [26, 65, 66]. The bias with a high ICC they hypothesised was due to the method calculating the difference in the geometric means of the log odds or log risk, rather than the arithmetic means. However, others have suggested that a cluster-level analysis estimates a cluster-specific effect [67]; this could explain the observed bias for the odds ratio, since Ukoumunne *et al* [65] compared the estimate to the marginal effect. The risk difference does not seem to be affected by any of these issues [65].

In SWTs, a cluster-level analysis is more difficult. The outcome cannot be summarised over all observations in the same cluster because some of the clusters' observations will be control and some will be intervention, and this would also ignore any period effects.

The only cluster-level analyses currently available, suggested by Hussey and Hughes [33], uses a model similar to model 3.1, but using summaries of the outcome in each cluster-period as the outcome. This model still requires a random effect for cluster because there are multiple cluster-period summaries per cluster. With equal cluster sizes and a continuous outcome, this would be equivalent to analysis with model 3.2. This method has been shown to produce biased estimates of the risk difference when the risk in each cluster-period is used as the outcome and the ICC is low [44]. It may also be unsuitable with a small number of clusters because it requires random effects.

3.2.7 Within-Period Analysis

A within-period, or vertical, analysis method is one which only utilises the vertical comparisons for estimating the intervention effect [3]. This is desirable because these are randomised comparisons that are not confounded with secular trends.

There are two published within-period analysis methods for SWTs, though methods for combining dependent effects have been explored for other longitudinal data contexts.

Granston explored the impact of modelling each period of the trial independently with model 3.1 (excluding the period effect) and combining the period-specific estimates as an inverse-variance weighted average [68]. However, Granston only considered scenarios where model 3.1 is correctly specified.

Similarly, Matthews and Forbes [28] derived algebraic formula for a vertical analysis. They looked at optimising the weights assigned to each period in such an analysis under different correlation structures. They found that inverse variance weights were optimal under the assumptions of model 3.1, and that only marginal improvements to efficiency were possible through introducing additional imbalance in the weights of the periods when assuming an autoregressive correlation structure between the periods, so that correlation between observations reduced over time. All correlation structures assumed that the total variance remained constant throughout the trial, so inverse variance weights were equivalent to weights by the imbalance in the number of clusters in each condition.

In the broader statistical literature, others have explored methods of combining dependent effects, mostly in the context of analysing longitudinal data with multiple observations of individuals.

Several papers have provided methods for testing a set of dependent estimates $\hat{\theta}_j$, $j = 1, \dots, J$ [69–71]. They were primarily concerned with the longitudinal data where individuals are observed multiple times. They wanted to analyse the effect of some parameter θ_j without specifying the correlation structure in the data and without assuming that the parameter effect was constant over time. In cases where θ_j can be assumed to be constant over time, they suggest combining the time-specific effects as a weighted average. Assuming complete data, they show that the weights for combining the estimates that maximise power are as follows:

$$\hat{w} = (\hat{w}_1, \dots, \hat{w}_J) = \left((1, \dots, 1) \hat{\Lambda}_w^{-1} (1, \dots, 1)' \right)^{-1} \hat{\Lambda}_w^{-1} (1, \dots, 1)'$$

where $\hat{\Lambda}_w$ is the covariance matrix for the time-specific parameters. Zeger [72] noted that the estimates $\hat{\theta}_j$ and covariance matrix $\hat{\Lambda}_w$ are identical to the estimates provided by GEE with an independent correlation structure and robust variance estimate.

Using these weights, the test statistic $\theta \left(\hat{w}' \hat{\Lambda}_w \hat{w} \right)^{-1/2}$ asymptotically follows a standard normal distribution. Moulton and Zeger [71] also suggest using bootstraps to calculate confidence intervals and a p-value for the overall effect.

This method has not been explored in the context of SWT. It provides a method of parametric inference without specifying a correlation structure for the data. However, the efficiency gains from estimating a covariance matrix are questionable given the results of Matthews and Forbes [28], and the as-

assumptions of a normal distribution may be unsuitable with the small number of clusters often seen in SWTs [1, 7].

3.2.8 Software Implementation of Analysis Methods

Despite some research warning about the suitability of model 3.1, the model remains the most popular choice in practice [8]. This could be because the literature on this topic is fairly recent, but a lack of available software has also been identified as a factor limiting the uptake of new statistical methods [73]. In this section, I will look at the availability of software to perform the analysis methods identified in the review to see where software might be limiting uptake.

LMM and GLMM can be implemented with relative ease in most common statistical software (Stata[®] [74], R [75], SAS[®]). Stata provides the commands *mixed* for LMM, and *meqglm* for GLMM. Several packages are available for R such as *lme4* [76]. In SAS, procedures such as *MIXED* and *GLIMMIX* can be used for LMM and GLMM. This wide availability of software may in part explain the popularity of model 3.1 and makes the development of more suitable model based alternative analyses appealing. These software can also fit the more complex correlation structures described in section 3.2.4.

The second most common analysis method of SWTs is GEE [8], which are also widely available in common statistical software. Since SWTs often have a small number of clusters, the small sample corrections suggested by Scott *et al* are recommended [77]. These are widely available in R [61], and SAS, but fewer corrections are available in Stata.

Bootstraps can also be implemented in these software. Stata's bootstrap command and the R package *multivcov* will perform bootstraps allowing for clustering in the data [78]; however, bootstraps must be manually coded in SAS. Bootstraps have not yet been shown to give appropriate inference for SWTs because of the small number of clusters the designs often use and this is likely the factor limiting their use.

Permutation tests have been shown to give robust inference when used with the model 3.1 [41, 42], but are not widely used in practice. This may be because they are less easily implemented. While there is an R package that performs permutation tests that are valid for SWTs [79], in Stata and SAS, the inbuilt commands are not suitable for use with SWTs. In both software, in-

dividual observations are permuted between intervention conditions and so do not maintain exchangeability for SWTs. Instead, the process must be manually coded in these software. Improving the ease of implementing the permutation test may increase the uptake of this more robust method, and so would be beneficial to SWT analysis.

3.2.9 Summary

The analysis model that prevails in practice is model 3.1. At the start of this PhD, there were few alternative methods to, and little criticism of, model 3.1 published. The intervening three years has seen developments in both of these areas. Model 3.1 has been shown to have inflated type-one error rates in the presence of cluster-specific period effects, consistent with findings from other trial designs, and many alternative analyses have been suggested. However, few of these analyses have been assessed to see if they have better performance than model 3.1 when observations in the same cluster but different periods are less correlated than observations in the same cluster and the same period.

Model 3.2 has performed well in other study designs with two periods (the CRXO and CRT with baseline observations), but this may not translate to an SWT, which can have many more than two periods, because the cluster-level autocorrelation is fixed to be the same for all periods. Model 3.4 has had little attention, despite its flexibility for allowing different correlation structures within clusters.

There has been little consideration in the literature to date of the impact of assuming a common intervention effect in mixed-effect models, despite these causing bias in CRTs [57].

The issues with type-one error have arisen because these models use the horizontal comparisons in their estimation of the intervention effect. The most robust analysis would be to exclude these comparisons as Granston considered [68], and accept that this comes at the cost of lower power. There is a need for methods that provide valid inference without requiring assumptions about the correlation structure between periods.

3.3 Power and Sample Size

As explained in section 3.1.1, a cluster randomised trial provides less information about the intervention effect than an individually randomised trial. Within cluster randomised trials, the amount of information available about the intervention effect varies by design.

In a CRT, the intervention effect can only be observed by comparisons between clusters in each condition (vertical comparisons). Power is maximised from these types of comparisons when there is an equal number of clusters in the control and intervention conditions [58].

In an SWT, the intervention effect can be observed by comparing clusters in each condition within each period (vertical comparisons), and by comparing periods in the control and intervention conditions within each cluster (horizontal comparisons). The vertical comparisons will provide less power than they do in a CRT because of the imbalance in the number of clusters in the control and intervention conditions. The horizontal comparisons are not available in a CRT so can increase the power of the SWT design relative to the CRT design. The amount of information taken from the horizontal comparisons will very much depend on the assumptions made about the period effects and the correlation within clusters between periods. The potential for this design to require a smaller sample size for the same power has resulted in some trials including a very small number of clusters. This raises problems for the use of mixed-effect models [80], and for generalisability of results [81].

Earlier sample size literature focused on model 3.1 assumptions, which include assuming equal correlation within clusters. More recently, sample size literature has focused on models with weaker assumptions, such as model 3.2, which assumes that observations in the same period are more correlated than observations in different periods. This model will weight the horizontal comparisons more or less depending on how strong the correlation is between periods.

All but two papers have focused on variance formulae and design effects for SWTs. Whilst these are simple to use, it has been noted that sometimes the results they provide are only approximate [82], particularly for outcomes that are not normally distributed [5]. One paper provides software to calculate the power of a given design with a given sample size [83], and one suggested using simulation studies [5].

As well as providing methods for calculating sample size requirements, this lit-

erature often also provides advice on how to minimise that sample size through design choices.

In this section, I will appraise this literature. I will begin with a description of how variance formulae and design effects are related and how they can inform trial design and sample size requirements. Subsections 3.3.2 - 3.3.4 detail literature that has provided design effects or variance formula, or used these methods to provide advice on trial design. Subsection 3.3.5 will describe results derived through simulation studies. Subsection 3.3.6 will describe implementation of sample size methods into statistical software.

3.3.1 Background to Design Effects

Using a design effect is a simple way to calculate the required sample size of a given trial design. A sample size is calculated assuming individual randomisation in a parallel trial, and this number is then multiplied by a design effect to give the appropriate sample size of a different design [17]. The justification of a design effect is as follows [84]:

Let α be the significance level, β be the statistical power, θ be an intervention effect, and $Var_{alt}(\hat{\theta})$ the variance of an estimate of the intervention effect using an alternative trial design such as an SWT. The power of a Wald test is as follows:

$$power = \Phi \left(\frac{\theta}{\sqrt{Var_{alt}(\hat{\theta})}} - \Phi^{-1}(1 - \alpha/2) \right)$$

where $\Phi(x)$ is the cumulative standard normal distribution function at a value x , and $\Phi^{-1}(x)$ is the inverse of this function so if $\alpha = 0.05$, $\Phi^{-1}(1 - \alpha/2) = 1.96$.

The power to detect a difference θ for a given sample size can be directly calculated from this formula, or the process of calculating a required sample size for a given power can be simplified by rearranging this formula:

$$M = \left[\Phi^{-1}(1 - \beta) + \Phi^{-1}(1 - \alpha/2) \right]^2 \frac{4\sigma^2 Var_{alt}(\hat{\theta})}{\theta^2 Var_{ind}(\hat{\theta})}$$

where M is the required total number of observations, σ^2 is the variance of the outcome, and $Var_{ind}(\hat{\theta})$ is the variance of the intervention effect estimate of an individually randomised, parallel trial design.

This is simply the formula for the sample size of an individually randomised trial multiplied by the ratio of the intervention effect estimate variance under the alternative trial design and the variance under the individually randomised trial. So, if one can calculate how much larger the variance of a given trial design will be compared to the variance of an individually randomised trial, called the design effect, one can use this ratio to calculate the sample size.

The design effect for a CRT depends on the ICC and the total cluster size [84]:

$$DE_{CRT} = 1 + (m - 1)\rho$$

where m is the total cluster size and ρ is the ICC.

The design effect for a CRT with 50% of observations at baseline is [85]:

$$DE_{CRTB} = \left(1 + \left(\frac{m}{2} - 1\right)\rho\right) \left(1 - \left(\frac{\frac{m}{2}\rho}{1 + \left(\frac{m}{2} - 1\right)\rho}\right)^2\right)$$

Design effects are very simple to use and the effect of some assumptions on the sample size, such as the ICC, can be explored through calculating the design effect for different values.

The formula for the design effect and $Var_{alt}(\hat{\theta})$ can also provide insights into how each parameter affects the power of a given design, for example the design effect for a CRT shows that the sample size of a CRT will increase as the ICC increases.

Some data assumptions are fixed in order to calculate a formula for $Var_{alt}(\hat{\theta})$. If these assumptions are not appropriate to a specific setting, the sample size calculated by the design effect will not necessarily provide the required power.

3.3.2 Hussey and Hughes Model Assumptions

Hussey and Hughes [33] provide a formula for the variance of an intervention effect estimate from their model 3.1. This formula provided the basis of all early work on sample size calculations for SWTs. Calculations of sample size based on this formula require the assumptions of model 3.1 to hold, namely common period effects and intervention effects, and so equal correlation of all observations within a cluster.

As described above, this formula can be directly applied to calculate the power to detect a given intervention effect for a given design and sample size.

Woertman *et al* [86] converted the variance formula into a design effect for an SWT. While it was not clear in the original paper that the formula gave the sample size required for each period of the trial (as opposed to the total sample size), this was later clarified [87]. The design effect derived by Woertman *et al* [86], corrected to give a total sample size requirement, is as follows:

$$DE_W = \frac{1 + \rho(gn + bn - 1)}{1 + \rho(0.5gn + bn - 1)} \frac{3(1 - \rho)}{2\left(g - \frac{1}{g}\right)} (g + b) \quad (3.7)$$

where ρ is the ICC, g is the number of sequences, n is the number of observations collected during each period, and b is the size of the baseline period relative to other periods. The baseline period is allowed to be bigger or smaller than all other periods of the design, but the final period, after the last sequence has switched to the intervention, must be the same size as the other periods of the design. There cannot be any cluster-periods excluded from the analysis and each sequence must contain the same number of clusters. This means that the total cluster size is given by $m = n(g + b)$ and is assumed to be the same for all clusters.

Assuming equal cluster size when in fact cluster size varies can cause under estimation of the required sample size in CRTs [67]. Initial investigations suggest that the SWT design is less sensitive to deviations from this assumption [88].

The availability of this design effect, and the variance formula derived by Hussey and Hughes have improved sample size calculations for SWTs: before these were published many trials ignored the SWT design altogether in their calculations and only accounted for the clustering [7]. Since this design effect has been published, it has become the most common method for sample size calculation [7].

The greater understanding of SWT sample size and power requirements led to comparisons with the power of other designs, in particular a CRT and CRT with baseline observations.

Some early comparisons of a CRT and SWT did not keep the total cluster size the same for the two designs, instead they compared an SWT with a CRT that had clusters the same size as one period of the SWT [86]. Later comparisons of these designs kept cluster size equal for each design and a general consensus emerged that the SWT is more powerful than a CRT when the ICC is high and the cluster sizes are large, but with low ICC or small cluster size, the

CRT is most powerful [87, 89, 90]. Hemming *et al* found that an SWT with 4 sequences was always more powerful than a CRT with baseline observations where 50% of observations were collected at baseline [85, 91].

There have also been many papers discussing improvements to the SWT design to increase its power. Numerical calculations have consistently shown that for the standard design SWT, the power increases with the number of sequences [5, 33, 86]. Here again, some comparisons increased the number of observations as they increased the number of sequences [5, 33], but even when the total number of observations was held constant, this pattern of increasing power persisted [86].

Lawrie *et al* [92] algebraically explored the impact of relaxing the assumption that each sequence contains an equal number of clusters. They found that it was optimal to have more clusters randomised to the first and final sequences to switch to the intervention than in the intermediary sequences.

3.3.3 Cluster-Period Interaction Model Assumptions

More recently, several researchers have developed design effects for an SWT assuming analysis with the cluster-period interaction model (model 3.2 or model 3.3) [14, 35, 50]. The design effect of Girling and Hemming for the standard SWT is:

$$DE_{girling} = \frac{\frac{m}{g+1} \frac{\sigma_q^2}{\sigma^2} + \frac{\sigma_e^2}{\sigma^2}}{\frac{2}{3} \left(1 - \frac{1}{g}\right) - \frac{1}{3} \left(1 - \frac{2}{g+1}\right)} R^*$$

where R^* is the cluster-mean correlation incorporating the additional random effects of model 3.3. R^* now has the interpretation of partitioning the cluster-means into time-dependent and time-independent variability:

$$R^* = \frac{\sigma_u^2 + \frac{\sigma_d^2}{n}}{\sigma_u^2 + \frac{\sigma_q^2}{g+1} + \frac{\sigma_d^2}{n} + \frac{\sigma_e^2}{m}}$$

For a repeated cross-sectional design, $\sigma_d^2 = 0$. Under the assumptions of model 3.1, Girling and Hemmings' design effect simplifies to the design effect 3.7 and $R^* = R$ [14].

These methods have only recently been published so their uptake in practice has not yet been assessed.

Girling and Hemming used their variance formula to identify the SWT design with minimal variance, and so a minimal sample size, allowing an unequal

allocation of clusters to sequences: the answer is the Hybrid design (figure 1.1d) [14]. The sample size of this design is minimised when the proportion of clusters randomised to the SWT was equal to the cluster-mean correlation. For example, if $R^* = 0.75$, 75% of the clusters should be randomised to the SWT and the remaining clusters randomised to the CRT [14].

Girling and Hemming also identified that an SWT with the periods before and after rollout half the size of the periods within rollout always had a smaller variance than the standard SWT design [14].

Hooper *et al* [50] derived a similar design effect and used it to explore the sensitivity of the sample size to changes in the design effect parameters. They found that sample size calculations were sensitive to changes in the cluster-level autocorrelation, with sample size increasing as the cluster-level autocorrelation reduced from one. This is concordant with the inflated type-one errors seen by others when this assumption is violated in analysis [40–42].

3.3.4 Within-Period Analysis

The additional power an SWT has compared to a CRT when the ICC or cluster size is large is because models 3.1 and 3.2 incorporate horizontal comparisons. In terms of vertical comparisons, the CRT should have the most power because this design has an equal number of clusters in each condition at all times.

Moulton *et al* [58] have quantified the drop in power due to this imbalance, though only for their specific design. They found that their SWT, assuming no clustering and independent periods with 14 sequences and a long period following rollout, would require 1.44 times the sample size of a CRT and found that this would increase to a limit of 1.5 as the number of sequences increased.

Others have explored the difference in efficiency comparing within-period analyses to model 3.1.

Both Granston [68] and Matthews and Forbes [28] compared the variance of their respective within-period methods with model 3.1 under the assumptions of model 3.1, and both found that the within-period methods resulted in a larger intervention effect variance than model 3.1. Matthews and Forbes show that the difference increased as the cluster-mean correlation R^* increased, driven by an increasing variance for the within-period method.

However, both of these studies have only compared the variance of the within-period method with model 3.1 under the assumptions of model 3.1 [28, 68]. As

shown in subsection 3.2.2, these assumptions are not always appropriate. The cluster-mean correlation will be lower where there is cluster-level autocorrelation, as defined for model 3.2, so the variance of a within-period method will be smaller. Conversely, Hooper *et al* [50] showed that the variance of model 3.2 is greater when there is cluster-level autocorrelation. This suggests that the difference between these methods will be smaller in practice than the current literature suggests.

3.3.5 Sample Size under Other Assumptions

Given the complexities of analysis that I described in section 3.2, design effects are not available for all potential chosen analysis models. For example, all previously described design effects assume that the intervention effect is common to all clusters, so would not be appropriate for analysis model 3.6. In such cases, simulation studies can be used to calculate the power of a particular study.

Baio *et al* [5] and Hughes [56] identified that ignoring period effects in the sample size calculation inflates the apparent power, so period effects must be considered from the start of planning a trial. This also means that all sample size methodology should incorporate period effects, however there are examples of methodology that have ignored this important confounder [93].

Baio *et al* [5] also found that when the intervention effect varied between the clusters this reduced the study's power to detect an effect, and showed that changes to the ICC affect cross-sectional designs more than cohort designs.

3.3.6 Software Implementation of Sample Size Calculations

As with analysis methods, the uptake of sample size calculation methods is dependent on their availability in commonly used software. There are two main software implementations available for SWT sample size calculations.

Calculation of power under the assumption of model 3.1 is available in several ways.

Hemming and Girling implemented the variance formula derived by Hussey and Hughes [33] as a command for the statistical software Stata [74, 83]. In their Stata command, they allow users to enter the design of their trial as a "design-pattern matrix" of 0,1, and "." for control periods, intervention periods,

or periods that are excluded respectively. An example of a design-pattern matrix for an incomplete SWT is given in figure 3.3. This allows trialists to explore a wide range of designs, for example excluding certain cluster-periods, or extending one particular period of the trial to twice as long. However, the assumptions of model 3.1 must apply for the resulting power to be valid.

Figure 3.3: Diagram of a design-pattern matrix

$$\begin{array}{ccccccc} 0 & 0 & 0 & 0 & 0 & . & 1 \\ 0 & 0 & 0 & 0 & . & 1 & 1 \\ 0 & 0 & 0 & . & 1 & 1 & 1 \\ 0 & 0 & . & 1 & 1 & 1 & 1 \\ 0 & . & 1 & 1 & 1 & 1 & 1 \end{array}$$

Hughes has also developed a spreadsheet for calculating the power using the Hussey and Hughes variance formula of a given design [33, 94].

The wide availability of software implementing methods making the assumptions of model 3.1 may explain their wide uptake [7].

Baio *et al* [95] has written an R package to simplify the process of using simulation studies to calculate the sample size of an SWT.

3.3.7 Summary

At the start of this PhD, the only literature available on the sample size of SWTs made the assumptions of model 3.1. These methods have since become popular [7]. The literature on designing SWTs focused on the impact of increasing the number of sequences on the trial power [86].

More recently, the sample size literature has relaxed the assumption of equal correlation within clusters and has been extended for cohort studies [14, 50]. These have yet to be implemented in any software to my knowledge.

Literature on making SWTs more efficient has looked at the effect of allowing an unequal allocation of clusters to sequences.

3.4 Summary

Much of the literature in this review has been published after I started this PhD; at the start of my PhD, literature on both design and analysis still

focused on model 3.1. Literature on analysis of SWTs to date, has largely focused on repeated cross-sectional designs. The design literature is starting to explore cohort designs, although many gaps remain in our understanding of the simpler repeated cross-sectional design. Therefore, this thesis will focus on furthering our understanding of the analysis of complete, cross-sectional designs.

Chapter 6 will explore improvements to the efficiency of the SWT design under the assumption of model 3.1. The finding of Girling and Hemming [14] and Lawrie *et al* [92] that it was optimal to have more clusters randomised to the sequences that remain in one condition for the longest time may provide some insight into how to improve SWT designs with an equal allocation of clusters to all sequences. A very simple modification that has not been explored is whether it is optimal to include the first period of the standard SWT design, before any clusters have switched to the intervention, and the last period, after all clusters have switched. This would lead to a design similar to the Hybrid design but with equal allocation. Equal allocation of clusters to sequences may be preferable where resources limit the rollout of the intervention, for example if a limited number of teams are available to introduce the intervention.

Chapters 7, 8 and 9 of this thesis will explore alternative analysis options.

In Chapter 7, I conduct a simulation study comparing model 3.1 to models with additional random effects for either period effects (model 3.4) or the intervention effect (model 3.6) in a range of different data scenarios. I extend the existing literature by exploring the deworming case-study in which the total variability of the outcome reduces in the second year. It is also novel to look at the impact of the intervention effect varying between clusters on the intervention effect estimate.

In Chapter 8, I describe an analysis method that only uses the vertical comparisons, similar to that of Granston [68] and Matthews and Forbes [28], but using cluster-period summaries and the permutation test. There are several advantages to my method, over the those proposed by Granston, and Matthews and Forbes: the use of cluster-summaries allows the simple calculation of risk differences as recommended by the CONSORT statement [96], and are easier to interpret than odds ratios. Secondly, because a cluster-level summary is used, there is no question over the number of clusters required for valid inference, as there is with mixed-effect models [17]. Lastly, using permutation tests should give valid inference without requiring assumptions about the correlation of

observations between periods, making this a much more robust analysis.

In Chapter 9, I address the need for software that computes permutation tests for SWTs and demonstrate a Stata command I have developed for this purpose. The command conducts the cluster-summary analysis described in Chapter 8, as well as enabling the methods described by Ji *et al* [41] and Wang and DeGruttola [42].

Bibliography

- [1] Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, Thompson JA, Fielding KL, Omar RZ, Ononge S, Hargreaves J and Prost A. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 2015. 16. (1):353.
- [2] Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G and Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials* 2015. 16. (1):352.
- [3] Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ and Fielding KL. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials* 2015. 16. (1):358.
- [4] Prost A, Binik A, Abubakar I, Roy A, De Allegri M, Mouchoux C, Dreischulte T, Ayles H, Lewis JJ and Osrin D. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. *Trials* 2015. 16. (1):351.
- [5] Baio G, Copas A, Ambler G, Hargreaves J, Beard E and Omar RZ. Sample size calculation for a stepped wedge trial. *Trials* 2015. 16. (1):354.
- [6] Hargreaves JR, Copas AJ, Beard E, Osrin D, Lewis JJ, Davey C, Thompson JA, Baio G, Fielding KL and Prost A. Five questions to consider before conducting a stepped wedge trial. *Trials* 2015. 16. (1):350.
- [7] Martin J, Taljaard M, Girling A and 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.
- [8] Barker D, McElduff P, D'Este C and Campbell MJ. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. *BMC Medical Research Methodology* 2016. 16. (1):69.

- [9] Grayling MJ, Wason JM and Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. *Trials* 2017. 18. (1):33.
- [10] Fuller C, Michie S, Savage J, McAteer J, Besser S, Charlett A, Hayward A, Cookson BD, Cooper BS, Duckworth G, Jeanes A, Roberts J, Teare L and Stone S. The Feedback Intervention Trial (FIT)–improving hand-hygiene compliance in UK healthcare workers: a stepped wedge cluster randomised controlled trial. *PloS One* 2012. 7. (10):e41617.
- [11] Miguel E and Kremer M. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 2004. 72. (1):159–217.
- [12] Hemming K, Lilford R and Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Statistics in Medicine* 2014. 34. (2):181–196.
- [13] Hooper R and Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ* 2015. 350:h2925.
- [14] Girling AJ and Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Statistics in Medicine* 2016. 35. (13):2149–2166.
- [15] Pol MC, Ter Riet G, Hartingsveldt M van, Krose B, Rooij SE de and Buurman BM. Effectiveness of sensor monitoring in an occupational therapy rehabilitation program for older individuals after hip fracture, the SO-HIP trial: study protocol of a three-arm stepped wedge cluster randomized trial. *BMC Health Service Research* 2017. 17. (1):3.
- [16] Lyons VH, Li L, Hughes JP and Rowhani-Rahbar A. Proposed variations of the stepped-wedge design can be used to accommodate multiple interventions. *Journal of Clinical Epidemiology* 2017. 86:160–167.
- [17] Hayes RJ and Moulton LH. Cluster Randomised Trials. 1st ed. USA: Chapman and Hall/CRC, 2009.
- [18] Mhurchu CN, Gorton D, Turley M, Jiang Y, Michie J, Maddison R and Hattie J. Effects of a free school breakfast programme on children’s attendance, academic achievement and short-term hunger: results from a stepped-wedge, cluster randomised controlled trial. *Journal of Epidemiology and Community Health* 2013. 67. (3):257–264.

- [19] Durovni B, Saraceni V, Moulton LH, Pacheco AG, Cavalcante SC, King BS, Cohn S, Efron A, Chaisson RE and Golub JE. 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):852–858.
- [20] Solomon E, Rees T, Ukoumunne OC, Metcalf B and Hillsdon M. The Devon Active Villages Evaluation (DAVE) trial of a community-level physical activity intervention in rural south-west England: a stepped wedge cluster randomised controlled trial. *International Journal of Behaviour Nutrition and Physical Activity* 2014. 11:94.
- [21] Bashour HN, Kanaan M, Kharouf MH, Abdulsalam AA, Tabbaa MA and 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):e002674.
- [22] Gucciardi E, Fortugno M, Horodezny S, Lou W, Sidani S, Espin S, Webster F and Shah BR. Will Mobile Diabetes Education Teams (MDETs) in primary care improve patient care processes and health outcomes? Study protocol for a randomized controlled trial. *Trials* 2012. 13:165.
- [23] Wight D, Raab GM, Henderson M, Abraham C, Buston K, Hart G and Scott S. Limits of teacher delivered sex education: interim behavioural outcomes from randomised trial. *BMJ* 2002. 324. (7351):1430.
- [24] Gambia Hepatitis Study Group. The Gambia hepatitis intervention study. *Cancer Research* 1987. 47. (21):5782–5787.
- [25] Haines T, O’Brien L, McDermott F, Markham D, Mitchell D, Watterson D and Skinner E. 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):144–151.
- [26] Omar RZ and Thompson SG. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Statistics in Medicine* 2000. 19. (19):2675–2688.
- [27] Ukoumunne OC and Thompson SG. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Statistics in Medicine* 2001. 20. (3):417–433.

- [28] Matthews JNS and Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. *Statistics in Medicine* 2017:1–18.
- [29] Grayling MJ, Wason JM and Mander AP. Group sequential designs for stepped-wedge cluster randomised trials. *Clinical Trials* 2017.
- [30] Gruber JS, Arnold BF, Reygadas F, Hubbard AE and Colford John M. J. Estimation of Treatment Efficacy With Complier Average Causal Effects (CACE) in a Randomized Stepped Wedge Trial. *American Journal of Epidemiology* 2014. 179. (9):1134–1142.
- [31] Kirkwood B and Sterne J. Essential Medical Statistics. Wiley, 2010.
- [32] Heagerty PJ and Zeger SL. Marginalized multilevel models and likelihood inference. *Statistical Science* 2000. 15. (1):1–19.
- [33] Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials* 2007. 28. (2):182–191.
- [34] Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, Jewell N, Bruckner T and Satariano WA. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* 2010. 21. (4):467–474.
- [35] De Hoop E. Efficient designs for cluster randomized trials with small numbers of clusters; stepped wedge and other repeated measurements designs. Thesis. Radboud University, 2014.
- [36] Heagerty PJ and Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* 2001. 88. (4):973–985.
- [37] Tully CM, Lambe T, Gilbert SC and Hill AV. Emergency Ebola response: a new approach to the rapid design and development of vaccines against emerging diseases. *Lancet Infectious Diseases* 2015. 15. (3):356–359.
- [38] Piszczek J and Partlow E. Stepped-wedge trial design to evaluate Ebola treatments. *Lancet Infectious Diseases* 2015. 15. (7):762–763.
- [39] Tweel I van der and Graaf R van der. Issues in the Use of Stepped Wedge Cluster and Alternative Designs in the Case of Pandemics. *American Journal of Bioethics* 2013. 13. (9):23–24.

- [40] Bellan SE, Pulliam JR, Pearson CA, Champredon D, Fox SJ, Skrip L, Galvani AP, Gambhir M, Lopman BA, Porco TC, Meyers LA and Dushoff J. Statistical power and validity of Ebola vaccine trials in Sierra Leone: a simulation study of trial design and analysis. *The Lancet Infectious Disease* 2015. 15. (6):703–710.
- [41] Ji X, Fink G, Robyn PJ and Small SS. Randomization inference for stepped-wedge cluster-randomised trials: An application to community-based health insurance. *Annals of Applied Statistics* 2017. 11. (1):1–20.
- [42] Wang R and DeGruttola V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. *Statistics in Medicine* 2017. 36. (18):2831–2843.
- [43] Li P and Redden DT. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology* 2015. 15. (1):38.
- [44] Barker D, D’Este C, Campbell MJ and McElduff P. Minimum number of clusters and comparison of analysis methods for cross sectional stepped wedge cluster randomised trials with binary outcomes: A simulation study. *Trials* 2017. 18. (1):119.
- [45] Zhan Z, Bock GH de, Wiggers T and Heuvel E van den. The analysis of terminal endpoint events in stepped wedge designs. *Statistics in Medicine* 2016. 35:4413–4426.
- [46] Sprent P and Smeeton N. Applied Nonparametric Statistical Methods, Fourth Edition. CRC Press, 2016.
- [47] Flynn TN and Peters TJ. Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results. *BMC Health Service Research* 2004. 4. (1):33.
- [48] Gail MH, Mark SD, Carroll RJ, Green SB and Pee D. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* 1996. 15. (11):1069–1092.
- [49] Hemming K, Taljaard M and Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. *Trials* 2017. 18. (1):101.

- [50] Hooper R, Teerenstra S, Hoop E de and Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine* 2016. 35. (26):4718–4728.
- [51] Turner RM, White IR, Croudace T and Group PIPS. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Statistics in Medicine* 2007. 26. (2):274–289.
- [52] Morgan KE, Forbes AB, Keogh RH, Jairath V and Kahan BC. Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome. *Statistics in Medicine* 2017. 36. (2):318–333.
- [53] Arnup SJ, Forbes AB, Kahan BC, Morgan KE and McKenzie JE. Appropriate statistical methods were infrequently used in cluster-randomized crossover trials. *Journal of Clinical Epidemiology* 2016. 74:40–50.
- [54] Van den Heuvel ER, Zwanenburg RJ and 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* 2017. 26. (2):766–775.
- [55] Twisk JW, Hoogendijk EO, Zwijsen SA and Boer MR de. Different methods to analyze stepped wedge trial designs revealed different aspects of intervention effects. *Journal of Clinical Epidemiology* 2016. 72:75–83.
- [56] Hughes JP, Granston TS and Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. *Contemporary Clinical Trials* 2015. 45:55–60.
- [57] Turner RM, Omar RZ and Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine* 2001. 20. (3):453–472.
- [58] Moulton LH, Golub JE, Durovni B, Cavalcante SC, Pacheco AG, Saraceni V, King B and Chaisson RE. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clinical Trials* 2007. 4. (2):190–199.
- [59] Donner A and Klar N. Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health* 2004. 94. (3):416–422.
- [60] Turner EL, Prague M, Gallis JA, Li F and Murray DM. Review of Recent Methodological Developments in Group-Randomized Trials: Part 2-Analysis. *American Journal of Public Health* 2017:e1–e9.

- [61] Scott JM, deCamp A, Juraska M, Fay MP and Gilbert PB. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. *Statistical Methods in Medical Research* 2017. 26. (2):583–597.
- [62] Pepe MS and Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation* 1994. 23. (4):939–951.
- [63] Sinclair JC and Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology* 1994. 47. (8):881–889.
- [64] Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 1987. 125. (5):761–768.
- [65] Ukoumunne OC, Forbes AB, Carlin JB and Gulliford MC. Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. *Statistics in Medicine* 2008. 27. (25):5143–5155.
- [66] Ukoumunne OC, Carlin JB and Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. *Statistics in Medicine* 2007. 26. (18):3415–3428.
- [67] Eldridge S and Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Wiley, 2012.
- [68] Granston T. *Addressing Lagged Effects and Interval Censoring in the Stepped Wedge Design of Cluster Randomized Clinical Trials*. Thesis. University of Washington, USA, 2014.
- [69] Wei LJ and Johnson WE. Combining dependent tests with incomplete repeated measurements. *Biometrika* 1985. 72. (2):359–364.
- [70] Stram DO, Wei L and Ware JH. Analysis of Repeated Ordered Categorical Outcomes with Possibly Missing Observations and Time-Dependent Covariates. *Journal of the American Statistical Association* 1988. 83. (403):631–637.
- [71] Moulton LH and Zeger SL. Analyzing Repeated Measures on Generalized Linear-Models Via the Bootstrap. *Biometrics* 1989. 45. (2):381–394.
- [72] Zeger SL. Commentary. *Statistics in Medicine* 1988. 7:161–168.

- [73] Pullenayegum EM, Platt RW, Barwick M, Feldman BM, Ofringa M and Thabane L. Knowledge translation in biostatistics: a survey of current practices, preferences, and barriers to the dissemination and uptake of new statistical methods. *Statistics in Medicine* 2016. 35. (6):805–818.
- [74] StataCorp. Stata Statistical Software: Release 14. 2015.
- [75] R Core Team. R: A Language and Environment for Statistical Computing. 2016. URL: <https://www.R-project.org/>.
- [76] Bates D, Maechler M, Bolker B and Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 2015. 67. (1):1–48.
- [77] Scott JM. Vaccine Efficacy Trials Using Stepped Wedge Design. Thesis. University of Washington, 2008.
- [78] Graham N, Arai M and Hagstromer B. multiwayvcov: Multi-Way Standard Error Clustering. 2016. URL: <https://CRAN.R-project.org/package=multiwayvcov>.
- [79] Simpson GL. permute: Functions for Generating Restricted Permutations of Data. 2016. URL: <https://CRAN.R-project.org/package=permute>.
- [80] Kahan BC, Forbes G, Ali Y, Jairath V, Bremner S, Harhay MO, Hooper R, Wright N, Eldridge SM and Leyrat C. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials* 2016. 17. (1):438.
- [81] Taljaard M, Teerenstra S, Ivers NM and Fergusson DA. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clinical Trials* 2016. 13. (4):459–463.
- [82] Hemming K. Sample size calculations for stepped wedge trials using design effects are only approximate in some circumstances. *Trials* 2016. 17. (1):234.
- [83] Hemming K and Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata Journal* 2014. 14. (2):363–380.
- [84] Kish L. Survey sampling. J. Wiley, 1965.
- [85] Hemming K and Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *Journal of Clinical Epidemiology* 2016. 69:137–146.

- [86] Woertman W, Hoop E de, Moerbeek M, Zuidema SU, Gerritsen DL and Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology* 2013. 66. (7):752–758.
- [87] Hemming K and 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):1427–1428.
- [88] Kristunas CA, Smith KL and Gray LJ. An imbalance in cluster sizes does not lead to notable loss of power in cross-sectional, stepped-wedge cluster randomised trials with a continuous outcome. *Trials* 2017. 18. (1):109.
- [89] Hemming K, Girling A, Martin J and 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):1058–1059.
- [90] Kotz D, Spigt M, Arts IC, Crutzen R and 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):1059–1060.
- [91] Hemming K, Haines T, Chilton P, Girling A and Lilford R. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015. 350:h391.
- [92] Lawrie J, Carlin JB and Forbes AB. Optimal stepped wedge designs. *Statistics and Probability Letters* 2015. 99:210–214.
- [93] Heo M, Kim N, Rinke ML and Wylie-Rosett J. Sample size determinations for stepped-wedge clinical trials from a three-level data hierarchy perspective. *Statistical Methods Medical Research* 2016.
- [94] Hughes J. Calculation of power for stepped wedge design. URL: <http://tinyurl.com/hwp5dgr> (visited on 24/07/2017).
- [95] Baio G. SWSamp: simulation-based sample size calculations for a stepped wedge trial (and more). URL: <https://sites.google.com/a/statistica.it/gianluca/swsamp> (visited on 10/07/2017).
- [96] Schulz KF, Altman DG, Moher D and Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010. 340:c332.

4 Aims and objectives

At the start of my PhD, there were many gaps in the SWT methodology literature. For this thesis, I aimed to address the gaps that I felt would have a large impact on improving the use of the SWT design.

The first aim, addressed ways to improve the design of SWTs. This was achieved through the following objective:

1. To identify a more efficient complete SWT design maintaining equal allocation of clusters to sequences through varying the size of the periods before the first and after the final periods switch to the intervention. To identify how this optimised design compares to other designs, namely a CRT, a CRT with baseline observations, and a hybrid design.

The rest of this thesis will be used to address the second aim, to identify improvements to the analysis of SWTs. At the start of my PhD, there was little literature exploring whether the Hussey and Hughes model (model 3.1) was appropriate. To address this aim, I will answer the following objectives:

1. To investigate the effect of misspecifying the correlation structure of a mixed-effect model used to analyse an SWT.
2. To propose an analysis method for SWTs using only the vertical comparisons. In particular, I will propose an analysis using a cluster-summary analysis in each period, combining within-period estimates of the intervention effect using a weighted average. I will assess the use of permutation tests for inference.
3. To develop a Stata command to facilitate the use of permutation tests, and the cluster-summary, within-period analysis described above for SWTs.

5 Methods: Background to Simulation Studies

In chapters 7 and 8 of this thesis I use simulation studies to address the aims of each chapter. The aim of chapter 7 was to explore the performance of model 3.1 when the random effects are misspecified, and to compare this to other potential mixed-effect models. The aim of chapter 8 was to assess the performance of a novel cluster-summary, within-period analysis method and compare this to model 3.1. In this chapter I will provide some details of the simulation studies that I conducted.

Simulation studies aim to replicate the process of conducting a study. When conducting an RCT, a sample is taken from a population. This sample is then randomised to determine whether, or in the case of an SWT when, they receive the intervention. After exposure, an outcome or outcomes are collected from the sampled individuals and the outcomes are then analysed. The principle of a simulation study is that all aspects of this process can be imitated under assumptions about what will happen in the trial. This can be informative about the properties of the trial analysis, such as its accuracy in estimating the intervention effect, the confidence interval coverage, or the power.

The steps in this process are (1) define the study design and generate a set of observations mimicking the process for sampling observations (2) analyse the data (3) repeat steps (1) and (2) many times, and (4) analyse simulation study. The following sections will describe how I implemented these steps in chapters 7 and 8 in more detail.

Chapter 7 simulations were run in R version 3.2, and chapter 8 simulations were run in R version 3.3 [1].

5.1 Data Generation

The first step of a simulation study is to generate some data mimicking data from a particular study. For an SWT, this means deciding on the following design features: selecting a number of clusters, a number of sequences, how many observations are collected in each period and before and after rollout in each cluster. From this, a dataset can be created with a row for each observation within each cluster and columns indicating the cluster that the observation belongs to, the period that the observation was collected in, and whether the observations was made under the control or intervention condition.

The outcomes of each individual will depend on the population from which the individuals have been sampled [2]. The population can be based on general characteristics representing a population where the study is likely to take place. An example is the simulation study by Barker *et al* [3] where a systematic review was used to determine likely values for the ICC and SWT designs. Alternatively, the population can be simulated to represent a specific population in which a study is likely to take place. This was the approach I took in chapters 7 and 8. Any results from a simulation study can only be generalised to the scenarios generated, so it is important that these are well thought through and reflect common situations [4].

In chapter 7, the data were simulated to mimic the deworming trial case study. The simulated study had two periods and three sequences as described in section 2.1. To understand the general population characteristics, without the introduction of the deworming intervention, I studied clusters in the third sequence of the trial, which did not receive the intervention in either year. Using a GLMM version of the random-period model (model 3.4) with a logit link, I determined the odds of school attendance in each year and the covariance matrix relating to these using:

$$\text{logit} \{P(y_{ijk} = 1 \mid t_j, u_i, v_{ij})\} = \mu + (\beta + v_i) t_j + u_i$$

where t_j is zero in the first year and 1 in the second year of the study and

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N \begin{pmatrix} 0 & \sigma_u^2 & \sigma_{u,v}^2 \\ 0 & \sigma_{u,v}^2 & \sigma_v^2 \end{pmatrix}$$

I then substituted these parameters into model 3.4 with a logit link, adding an intervention effect to generate samples of data for the simulated trial.

In chapter 8 I used a similar process. I analysed the NHS health check data to determine the underlying trends in health check acceptance and correlation structure across the two years. Here, I again used a GLMM form of model 3.4 with a logit link. For this data I treated time as continuous and used fractional polynomials [5] to determine the shape of the trends within each year and included a binary covariate for year:

$$\text{logit} \{P(y_{ijk} = 1 \mid t_{1j}, t_{2j}, u_i, v_{1i}, v_{2i})\} = \mu + (\beta_1 + v_{1i}) t_{1j} + (\beta_2 + v_{2i}) t_{2j}^3 + u_i$$

where t_{1j} indicates the year, 0 and 1 in the first year and second year respectively, and t_{2j} indicates the quarter within each year as 0,1,2, or 3.

$$\begin{pmatrix} u_i \\ v_{1i} \\ v_{2i} \end{pmatrix} \sim N \begin{pmatrix} 0 & \sigma_u^2 & \sigma_{u,v_1}^2 & \sigma_{u,v_2}^2 \\ 0 & \sigma_{u,v_1}^2 & \sigma_{v_1}^2 & \sigma_{v_1,v_2}^2 \\ 0 & \sigma_{u,v_2}^2 & \sigma_{v_1,v_2}^2 & \sigma_{v_2}^2 \end{pmatrix}$$

Again, I then substituted the determined values into an identical model with the addition of an intervention effect to generate samples of data for the simulated trial. This generated samples with a specified cluster-specific intervention effect. Other methods are available to generate samples with a specific marginal intervention effect [6].

To widen the generalisability of the results, the simulation study can be repeated using different data generating processes that make different data assumptions. This could include different correlation assumptions, missing data mechanisms, and the size of clusters. When using simulation studies for sample size calculations, it can be useful to explore the spread of power estimates under different assumptions because the form of the data produced during a trial is difficult to predict in advance [2]. When using simulation studies for methodology this gives a wider applicability and understanding of study results.

In Chapter 7, I also generated data with $\sigma_{u,v}^2 = \sigma_v^2 = 0$, and I generated data with a varying intervention effect by adding a random effect $z_i \sim N(0, \sigma_z^2)$, independent of u_i and v_i to the data generating model. This allowed me to assess the impact of different correlation structures on the analysis model performance. I also repeated the simulation study with a larger intervention effect in the second sequence of the trial than in the first sequence of the trial (see figure 2.1 for the trial design). This allowed me to demonstrate the weight given to horizontal and vertical comparisons as only sequence two contributed horizontal comparisons in this design. Further details are given in chapter 7.

In Chapter 8, I generated data with $\sigma_{v_1}^2 = \sigma_{v_2}^2 = \sigma_{u,v_1}^2 = \sigma_{u,v_2}^2 = \sigma_{v_1,v_2}^2 = 0$, and generated data with less variability, by dividing the covariance matrix by a factor of 0.2. This allowed me to assess the performance of the novel analysis in the presence of different correlation structures and different degrees of clustering.

5.2 Analysing Simulated Data

These generated datasets can then be analysed using a chosen analysis model, or several different analysis models can be compared. The results of each analysis model, such as the intervention effect estimate and its standard error, are collected and saved. The extracted results can be used to investigate the properties of the analysis model/s under the data assumptions simulated [4]. Comparing the performance of different analysis models under different data generating processes can help to identify when problems with analysis models occur and when one analysis method may be preferable to another [4].

In Chapter 7, I compared models 3.1, 3.4, and 3.6. For each of these models, there were data scenarios where they were correctly specified (making appropriate assumptions about the data), and incorrectly specified (making incorrect assumptions about the data). Comparing the correctly specified to the incorrectly specified models also helped interpretation of results.

In Chapter 8, I analyse the data using model 3.1 because this the most commonly used analysis method in practice [7]. I compare this to my novel cluster-summary, within-period analysis method.

5.3 Number of Simulations

Using one generated dataset gives very little information about the properties of the analysis model/s because the random variability in the data gives rise to error, known as Monte-Carlo error [8, 9]. Instead, we repeat this process many times. Each time taking a different sample from the statistical distributions to get a different sample of outcomes within the same population, and applying the analysis models. To allow the simulations to be replicable, a random number seed should be provided at the start of the study; this means that the same values are sampled from the distributions each time the simulation study is rerun [4].

The number of repetitions required will depend on how certain one wants to be of a particular property of an analysis model. More details of these properties were given in section 3.2.1. For example, when simulation studies are used as part of a sample size calculation, the power will be of most interest.

In Chapter 7, the primary interest was identifying biased estimation of the intervention effect, so the number of simulations was selected so that an intervention effect odds ratio of 1.5 would be estimated with 5% accuracy in scenarios with the most variance (assumed to be a variance of 0.05). This led to a requirement of 500 simulations, calculated by [4]:

$$N = \left(\frac{(\Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta)) \text{Var}(\theta)}{\theta * \text{accuracy}} \right)^2$$

where N is the number of simulations, $\Phi^{-1}(x)$ is the inverse cumulative standard normal distribution function, α is the type-one error rate, β is the required power to detect a difference at the required level of accuracy, θ is the intervention effect with an expected variance of $\text{Var}(\theta)$.

In chapter 8, the primary interest was maintaining 95% confidence interval coverage. The coverage will follow a binomial distribution, and so we can use the normal approximation to the binomial distribution to decide the required width of the confidence interval for the confidence interval coverage using [10]:

$$CI \text{ width} = 1.96 * \sqrt{\frac{0.95(1 - 0.95)}{N}}$$

Using this formula, 1000 simulations will estimate the confidence interval coverage to within 1.4%.

5.4 Analysing the Simulation Study

Simulation studies can be used to evaluate all of the properties described in subsection 3.2.1.

5.4.1 Bias

Burton *et al* [4] describe the different measures of bias in a simulation study. The parameter estimates generated for each simulated dataset will vary, giving a distribution of estimates with a mean $\bar{\hat{\theta}}$ and variance $\text{var}(\hat{\theta})$. In this context,

the square root of the variance is the empirical standard error of the estimates, and should be similar to the model estimate of the standard error.

The simplest measure of bias is the difference between the mean of the estimates and the true intervention effect (θ_A):

$$\text{absolute bias} = \bar{\hat{\theta}} - \theta_A$$

Bias should be considered in relation to the Monte-Carlo error around the parameter estimates. A small absolute bias in an estimate that has large variability is less important than the same absolute bias in an estimate with a small variability. Mean parameter estimates more than 1/2 a standard deviation away from the true parameter effect have been found have an impact on type 1 error rate and so this is sometimes used to indicate practically important bias [4].

However, this value is dependent on the parameter effect size and the variance of the estimates. The percentage bias removes the dependence on the effect size, but is still dependent on the variance of the estimates:

$$\text{percentage bias} = \frac{\bar{\hat{\theta}} - \theta_A}{\theta_A} * 100$$

Some have used a percentage bias greater than 10% to indicate a concerning amount of bias [11].

Standardised bias removes the dependence on the variance:

$$\text{standardised percentage bias} = \frac{\bar{\hat{\theta}} - \theta_A}{\sqrt{\text{Var}(\hat{\theta})}} * 100$$

A standardised bias of more than 40% has been shown to impact the method's error rates [12].

In Chapter 7, I calculated absolute and the percentage bias, to improve the interpretability of results. In chapter 8, I calculated only absolute bias, but interpreted the results considering the variability of the estimates.

5.4.2 Confidence Interval Coverage and Type-one Error

In a simulation study, we can calculate the percentage of repetitions in which a parameter's confidence interval contains the true effect. If the percentage

is less than 95%, the confidence interval is under covered, if the percentage is greater than 95%, the confidence interval is over covered. Under coverage is of greater concern than over coverage [4].

Both chapters 7 and 8 explored the coverage of all the analysis methods considered. In chapter 7 I also calculated the type-one error rate.

5.4.3 Power

The power is calculated by simulating datasets with an intervention of θ_A and calculating the percentage which have an intervention-effect estimate with $p < \alpha$. For example, a trialist may want to know the power to detect an intervention-effect of θ_A at a significance level of 0.05. If data is simulated to have a true intervention effect of θ_A and 80% of the intervention-effect estimates had $p < 0.05$, the analysis has 80% power to detect an intervention-effect of θ_A at a significance-level of 0.05 under the assumptions made in the data generating process.

When using a simulation study for sample size calculations and when comparing the efficiency of analysis models, the power to detect an effect of size is the property of primary interest. Power calculated in this way is only interpretable if the method has correct confidence interval coverage.

Chapters 7 does not report the power of the analysis models as the simulation study focused on identifying bias and under coverage of confidence intervals. Chapter 8 explored the power of the analysis methods considered, although this was not possible in all scenarios because of under covered confidence intervals.

5.4.4 Presentation of Results

In many previously published simulation studies, the Monte-Carlo error of analysis method properties has not been well reported [9]. This can lead to over interpretation of random variation in each property. To avoid this, in both papers, the parameter mean estimates are displayed with a measure of the spread of the estimates (chapter 7 reports the interquartile range of the estimates, chapter 8 reports the standard deviation of the estimates). Confidence intervals are given to estimates of confidence interval coverage in both chapters, and absolute bias in chapter 7. The variance of each property was calculated assuming normal approximations, which has been shown to be appropriate in most situations [8].

5.5 Summary

Simulation studies are a very flexible tool for exploring the properties of an analysis in a given set of circumstances, and advances in computing power have made simulation studies more feasible. They enable consideration of all aspects of inference, Burton *et al* [4] encourages consideration of several aspects to improve interpretability of results, for example, only investigating power could fail to identify that the analysis is producing biased estimates. However, the results they provide are only as generalisable as the breadth of data-generating processes explored.

In Chapters 7 and 8, I have used simulation studies to explore the properties of several analysis methods. The data generating processes in both chapters were informed by real setting where either an SWT had been conducted, or would be likely to be conducted (the deworming case study and NHS health-check case study respectively). Bias, and confidence interval coverage were explored in both simulations studies. Power was also considered in chapter 8. Careful thought was given to presenting results.

Bibliography

- [1] R Core Team. R: A Language and Environment for Statistical Computing. 2016. URL: <https://www.R-project.org/>.
- [2] Landau S and Stahl D. Sample size and power calculations for medical studies by simulation when closed form expressions are not available. *Statistical Methods in Medical Research* 2013. 22. (3):324–345.
- [3] Barker D, D’Este C, Campbell MJ and McElduff P. Minimum number of clusters and comparison of analysis methods for cross sectional stepped wedge cluster randomised trials with binary outcomes: A simulation study. *Trials* 2017. 18. (1):119.
- [4] Burton A, Altman DG, Royston P and Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006. 25. (24):4279–4292.
- [5] Royston P and Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1994. 43. (3):429–467.

- [6] Ukoumunne OC, Carlin JB and Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. *Statistics in Medicine* 2007. 26. (18):3415–3428.
- [7] Barker D, McElduff P, D’Este C and Campbell MJ. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. *BMC Medical Research Methodology* 2016. 16. (1):69.
- [8] White IR. simsum: Analyses of simulation studies including Monte Carlo error. *Stata Journal* 2010. 10. (3):369.
- [9] Koehler E, Brown E and Haneuse SJ. On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. *Journal of the American Statistical Association* 2009. 63. (2):155–162.
- [10] DÁaz-Emparanza I. Is a small Monte Carlo analysis a good analysis? *Statistical Papers* 2002. 43. (4):567–577.
- [11] Marshall A, Altman DG, Royston P and Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Medical Research Methodology* 2010. 10:7.
- [12] Collins LM, Schafer JL and Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001. 6. (4):330–351.

6 Paper A: The Optimal Design of the Stepped-Wedge Trials with Equal Allocation to Sequences, and a Comparison to Other Trial Designs

In research paper A, I address the first aim of this thesis: to improve the efficiency of the SWT design.

My review of the literature on SWT design suggested that the periods before and after rollout may be inefficient. In this paper I explore this question. I derive a design effect that assumes that data follow model 3.1 (in this chapter referred to as the Hussey and Hughes model). I use differentiation of this new formulation of the design effect to identify the proportion before and after rollout and the number of sequences that minimises the design effect and so gives the most efficient SWT design. I then compare this most efficient SWT design to other cluster randomised designs.

In this paper, k denotes the number of sequences (all other chapters of this thesis have used the letter g). The appendix given here differs to the published version to improve consistency of terminology and format with the rest of this thesis.

The paper has been peer reviewed and is published in *Clinical Trials*, titled “The optimal design of the stepped-wedge trials with equal allocation to sequences, and a comparison to other trial designs” by myself, Katherine Fielding, James Hargreaves, and Andrew Copas. The paper is licensed under CC BY 2.0. It did not require ethics approval because no real data are used. Evidence of retention of copyright is given in Appendix A.



Registry

T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jennifer Thompson
Principal Supervisor	Katherine Fielding
Thesis Title	Improving the design and analysis of stepped-wedge trials

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	Clinical trials		
When was the work published?	11/08/2017		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I derived the design effect and all optimality results that followed from this design effect. I wrote the first draft of the paper and the responses to reviewers comments.
--	---

Student Signature: _____

Date: 12/09/17

Supervisor Signature: _____

Date: 12/9/2017

The optimal design of stepped wedge trials with equal allocation to sequences and a comparison to other trial designs

Clinical Trials

1–9

© The Author(s) 2017



Reprints and permissions:

sagepub.co.uk/journalsPermissions.nav

DOI: 10.1177/1740774517723921

journals.sagepub.com/home/ctj

Jennifer A Thompson^{1,2}, Katherine Fielding¹, James Hargreaves³ and Andrew Copas²

Abstract

Background/Aims: We sought to optimise the design of stepped wedge trials with an equal allocation of clusters to sequences and explored sample size comparisons with alternative trial designs.

Methods: We developed a new expression for the design effect for a stepped wedge trial, assuming that observations are equally correlated within clusters and an equal number of observations in each period between sequences switching to the intervention. We minimised the design effect with respect to (1) the fraction of observations before the first and after the final sequence switches (the periods with all clusters in the control or intervention condition, respectively) and (2) the number of sequences. We compared the design effect of this optimised stepped wedge trial to the design effects of a parallel cluster-randomised trial, a cluster-randomised trial with baseline observations, and a hybrid trial design (a mixture of cluster-randomised trial and stepped wedge trial) with the same total cluster size for all designs.

Results: We found that a stepped wedge trial with an equal allocation to sequences is optimised by obtaining all observations after the first sequence switches and before the final sequence switches to the intervention; this means that the first sequence remains in the control condition and the last sequence remains in the intervention condition for the duration of the trial. With this design, the optimal number of sequences is $1/(1 - \sqrt{R})$, where $R = \rho m / (1 + \rho(m - 1))$ is the cluster-mean correlation, ρ is the intracluster correlation coefficient, and m is the total cluster size. The optimal number of sequences is small when the intracluster correlation coefficient and cluster size are small and large when the intracluster correlation coefficient or cluster size is large. A cluster-randomised trial remains more efficient than the optimised stepped wedge trial when the intracluster correlation coefficient or cluster size is small. A cluster-randomised trial with baseline observations always requires a larger sample size than the optimised stepped wedge trial. The hybrid design can always give an equally or more efficient design, but will be at most 5% more efficient. We provide a strategy for selecting a design if the optimal number of sequences is unfeasible. For a non-optimal number of sequences, the sample size may be reduced by allowing a proportion of observations before the first or after the final sequence has switched.

Conclusion: The standard stepped wedge trial is inefficient. To reduce sample sizes when a hybrid design is unfeasible, stepped wedge trial designs should have no observations before the first sequence switches or after the final sequence switches.

Keywords

Stepped wedge trial, cluster randomised trial, hybrid trial, sample size, design effect, power, study design

Introduction

Stepped wedge trials (SWTs) are growing in popularity, but modification of the design to minimise their sample size have not been fully explored.

In an SWT, clusters are randomised into allocation sequences. Sequences consist of a different number of periods in the control condition, followed by the remaining periods of the trial in the intervention

¹Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK

²London Hub for Trials Methodology Research, MRC Clinical Trials Unit at University College London, London, UK

³Department of Social and Environmental Health Research, London School of Hygiene & Tropical Medicine, London, UK

Corresponding author:

Jennifer A Thompson, Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK.
Email: jennifer.thompson@lshtm.ac.uk

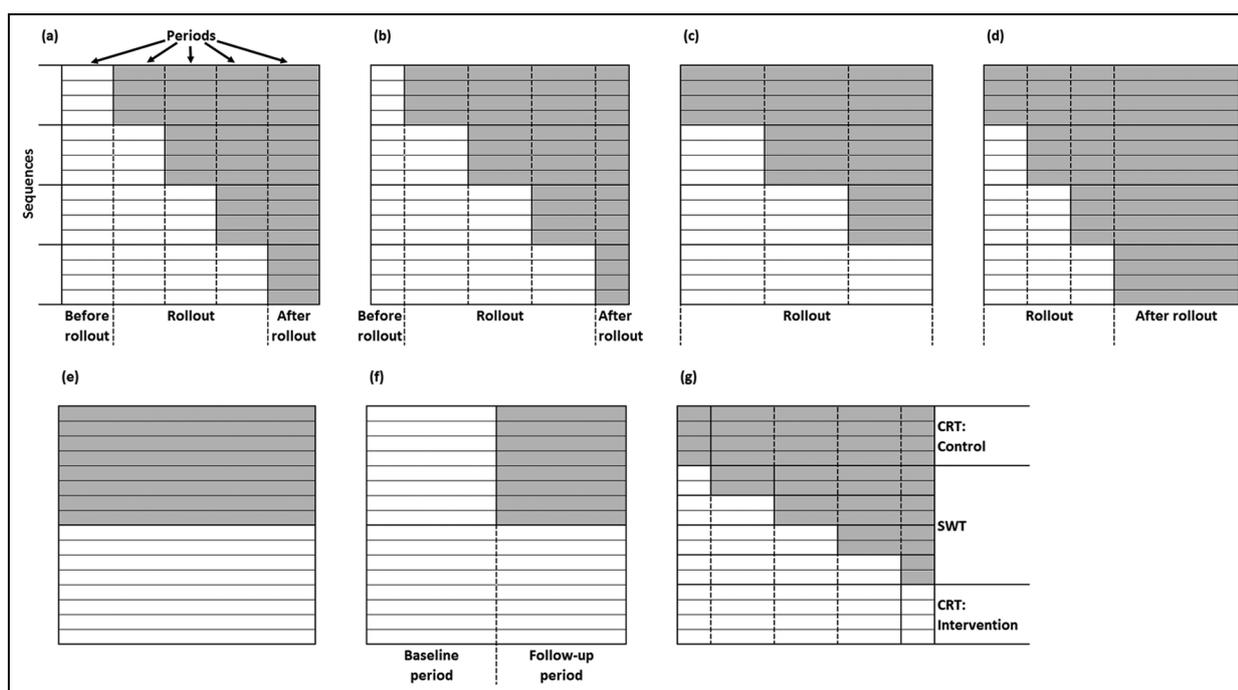


Figure 1. Diagrammatic illustrations of trial designs. Each has the same number of clusters and the same total cluster size. (a)–(d) Stepped wedge cluster-randomised trials (SWTs) with four sequences varying the amount of data before and after rollout. (a) Standard design: the same number of observations before and after rollout and between sequences switching, (b) number of observations before and after rollout is half the number between sequences switching, (c) optimised design: no observations before or after rollout, (d) no observations before rollout, 50% after rollout. (e)–(g) Other designs: (e) parallel cluster-randomised trial: CRT, (f) parallel cluster-randomised trial with baseline observations, (g) hybrid design with 50% CRT, 50% SWT with four sequences and the number of observations before and after rollout equal to half the number between sequences switching.

condition. At the beginning of each period, one of the sequences switches to the intervention, as shown in Figure 1(a). This means that a design with k sequences has $k - 1$ periods between the first sequence switching and the final sequence switching to the intervention condition. We will call this section of the trial ‘rollout’ because the intervention has been introduced to some but not all the clusters.

Before and after rollout, that is, before the first sequence switches to the intervention and after the final sequence switches to the intervention, there can be additional periods of data collection that may be longer or shorter than the periods during rollout (Figure 1(b) and (d)). In a standard SWT design (Figure 1(a)), these periods are the same length as the periods between rollout. Variations of the SWT design could include only the rollout period (Figure 1(c)), a period before rollout but not after or after rollout but not before.

There are many further possible variations but in this article, we only consider designs where the same number of observations is collected in each of the periods during rollout and that the same number of clusters is randomised to each sequence. This implies that the rollout will occur at an even pace, that is, equal numbers of clusters implement the intervention at each time, which we feel is a natural constraint when there are limited

resources to conduct the implementation. We focus on data with equal correlation within each cluster.¹

There are several approaches to sample size calculation for SWTs;^{2–5} the simplest is the design effect approach. Here, a sample size is calculated assuming individual randomisation and is then multiplied by a design effect to increase the sample size appropriately for a different design. Woertman et al. developed what they termed a design effect for an SWT, but this must be multiplied by the number of periods in the trial to give what we define here as the design effect.^{2,6} While their design effect has been a useful contribution to the literature, it is difficult to untangle the effects of each design component on the sample size to examine how to improve the efficiency of SWTs.

One such component which cannot be examined by the design effect of Woertman et al. is the number of periods before and after rollout; changing the number of periods before rollout increases the total cluster size so it is difficult to examine the impact of this change holding the total cluster size constant.² Girling and Hemming found that having half a period before rollout and half a period after rollout produced greater efficiency than the standard design when the total cluster size was held constant,⁷ but it is not known whether even fewer observations before and after rollout would

be more efficient. In addition, although there is a consensus through empirical evidence that the power of a standard SWT increases with an increase in the number of sequences,^{2,3} this has not been explored for variations of the SWT design.

Researchers often cite increased statistical power as a reason for choosing SWTs over other trial designs.⁸ Designs where clusters act as their own control can be more powerful,⁷ but they also require assumptions about changes in the outcome over time. Comparisons have been made between SWTs and parallel cluster-randomised trials (CRTs; Figure 1(e)), CRTs with baseline observations where half of the total cluster size are baseline observations (Figure 1(f)), and more recently with the hybrid design described by Girling and Hemming.⁷ The hybrid design includes sequences that are in the control or intervention conditions for the entire study and allows allocation to those two sequences to differ from allocation to the remaining sequences which form an SWT design, as shown in Figure 1(g). Standard SWT designs have been found to be more efficient than CRTs when the intracluster correlation coefficient (ICC) is high and when the total cluster sizes are large, and a standard SWT with four or more sequences always has more power than a CRT with baseline observations.^{6,9–11} The hybrid design appears to have the highest power as it is most flexible,⁷ but the degree of efficiency gain from allowing unequal allocation has not been established.

In this article, we give a new design effect expression for an SWT that allows the number of observations before and after rollout to vary without increasing the total cluster size but maintains the requirement common to the standard SWT of equal-sized periods between sequences switching to the intervention and the same number of clusters randomised to each sequence. This allowed us to identify the optimal number of sequences and the optimal number of observations before and after rollout to minimise the required number of clusters for a given power, ICC, and total cluster size. We compare the efficiency of our optimised SWT designs to several other common trial designs for a given power, ICC, and total cluster size, and we provide guidance in choosing a trial design. An example is then used to demonstrate the difference in sample size between possible designs.

Methods

SWT

Woertman et al. developed a design effect for an SWT under the assumptions of the Hussey and Hughes analysis model.^{2,3,6} We rewrite this design effect based on similar methodology to that used by Woertman et al.² In our new design effect, the number of observations before and after rollout is specified as proportions of

the total cluster size. For example, one could have half of all observations after rollout and none before rollout, as shown in Figure 1(d).

For simplicity, we assumed that the outcome is normally distributed, clusters are of equal size, and observations are equally correlated within clusters regardless of time or whether from the control or intervention condition. We assume that the intervention effect is constant over time, is fully realised by the first observation after the intervention is implemented, and is common across all clusters. We also require that secular trends are common to all clusters, the same number of clusters is randomised to each sequence, and that there is the same number of observations in all periods between sequences switching to the intervention.

This new design effect will be used to find the combination of number of sequences and proportion of observations before and after rollout that minimise the sample size (number of clusters) for a given power, total cluster size, and ICC. This SWT, derived under the constraint of equal allocation to sequences, will be referred to as an ‘optimised’ SWT.

This optimised SWT will then be compared to other trial designs. We will consider a CRT, a CRT with baseline observations and the hybrid design.⁷ Throughout these comparisons, we fix the power, total cluster size, and ICC.

Parallel CRT

A CRT (Figure 1(e)) is an attractive design because the intervention effect is not confounded with time and so it does not require assumptions about secular trends. The published design effect for a CRT is as follows

$$DE_{CRT} = 1 + (m - 1)\rho \quad (1)$$

where m is the total cluster size, and ρ is the ICC.¹²

Parallel CRT with baseline observations

A CRT with baseline observations (Figure 1(f)) is equivalent to an SWT with two sequences, some proportion of observations before rollout and no observations after rollout.¹³ Making the same assumptions as the SWT, such a design can be analysed with the same model as an SWT,³ and so, the new design effect can also be applied. We used our design effect to find the optimum proportion of observations to have at baseline to minimise the sample size of this design before comparing the required sample size to the optimised SWT.

Hybrid design trial

Girling and Hemming described a trial design where some of the clusters were randomised to a parallel

CRT, while the remaining clusters were randomised to an SWT with half a period before rollout and half a period after rollout (Figure 1(g)).⁷ This hybrid trial design makes the same assumptions as the SWT and can be analysed with the Hussey and Hughes analysis model.³ They found that the optimal proportion of clusters to randomise to the SWT was the cluster-mean correlation defined as follows⁷

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

where $0 \leq R \leq 1$ increases as the ICC or total cluster size increases. So, when the ICC or cluster size increases, the optimal proportion of clusters randomised to the SWT increases and the proportion randomised to the CRT reduces.

The hybrid design is flexible enough that it can simplify to a parallel CRT, it can simplify to a design similar to a standard SWT but with half a period before and half a period after rollout (Figure 1(b)), and it can simplify to a modified SWT design with no period before and after rollout, and the first and final periods are half the size of the other periods, similar to the design considered later in this article. The first two of these simplifications are straightforward to see; all clusters are randomised to the relevant part of the trial. The final simplification requires a proportion of $2/(k+2)$ clusters to be randomised to the parallel CRT and the remaining clusters to be randomised to the SWT with k sequences. Following the recommendations of Girling and Hemming will lead to one of these designs if it is the most efficient option or it will lead to a hybrid design if that is most efficient.⁷

We compared our optimised SWT with k sequences to an optimal hybrid design to see whether the increased flexibility of the hybrid design gave a practically relevant decrease in sample size. The optimal hybrid had a proportion equal to the cluster-mean correlation of clusters randomised to the SWT, and the SWT within the hybrid had as many sequences as there were clusters.

Choosing an SWT design

Finally, we acknowledge that the optimised SWT may not always be a practical design and provide recommendations for how to design an efficient and practical trial. We provide an example to demonstrate the differences in sample size of different designs.

Results

The design effect for an SWT

We define k as the number of sequences, β as the proportion of the total cluster size that is before rollout, and α the proportion of the total cluster size that is

after rollout. For example, in a standard SWT, $\beta = \alpha = 1/(k+1)$ (Figure 1(a)), alternatively one could have no observations before rollout, so $\beta = 0$, but a large period after rollout, say half of the total cluster size, so $\alpha = 0.5$ (Figure 1(d)). The total cluster size remains the same regardless of α and β , and the remaining observations are distributed equally between the periods within rollout.

In Appendix 1, we derive a design effect for an SWT with these characteristics

$$DE_{SWT} = \frac{(1 + (m-1)\rho) \frac{3k(k-1)}{2(k+1)}}{(1-R)} \frac{1}{[1 - (\beta + \alpha)][k(1 - 0.5R[1 - (\beta + \alpha)]) - 1]} \quad (2)$$

The terms α and β only affect the design effect through their sum $\alpha + \beta$ and so it is the combined proportion of observations outside rollout that affects the power, rather than the individual quantities. α and β are also exchangeable in this equation; this means that observations before and after rollout have the same impact on power. This is due to the assumption of observations being equally correlated within each cluster.

Minimising the sample size of an SWT

In Appendix 2, we show that the optimised SWT has no observations outside rollout ($\alpha + \beta = 0$; Figure 1(c)) with the number of sequences depending on the ICC and total cluster size, as shown in equation (3)

$$\text{Optimal number of sequences} = \frac{1}{1 - \sqrt{R}} \quad (3)$$

Equation (3) will give a non-integer number; to find the exact optimal number of sequences, calculate the design effect (equation (2)) for the integers either side of the result given by equation (3), but a rule of thumb is to round the result to the nearest integer.

The optimal number of sequences increases as the cluster-mean correlation increases (i.e. the ICC or total cluster size increase), but for low cluster-mean correlation (low ICC or small total cluster size), a small number of sequences is optimal. For example, with 100 observations per cluster and an ICC = 0.01 ($R = 0.50$), it is optimal to have 3 sequences, but with an ICC = 0.1 ($R = 0.92$), it is optimal to have 24 sequences. Figure 2 shows the optimal number of sequences for different cluster-mean correlations.

Minimising the sample size of a CRT with baseline observations

The design effect (equation (2)) can also give the optimal proportion of baseline observations for a CRT. In

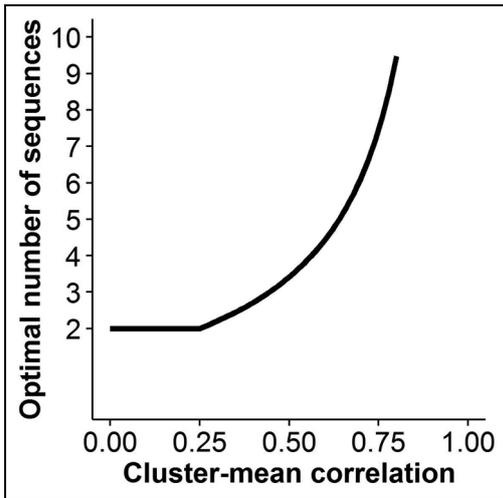


Figure 2. Optimal number of sequences by the cluster-mean correlation. The number of sequences tends to infinity as the cluster-mean correlation tends to 1.

Appendix 4, we show that the proportion of observations at baseline that minimises the sample size of a CRT with baseline observations is as follows

$$\beta = \begin{cases} 1 - \frac{1}{2R} & \text{if } R \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

For low values of the cluster-mean correlation, it is optimal to have no baseline observations, and for higher values, the optimal proportion of baseline measurements increases to a ceiling of 50% of observations.

Comparison of an optimised SWT to a CRT

In Appendix 3, we show that when the optimal number of sequences from equation (3) is < 2.5 , this means that a CRT would require a smaller sample size than any SWT with no observations outside rollout. As a rule of thumb, a CRT will require a smaller sample size when

$$\rho < \frac{1}{\frac{9}{16}m + 1}$$

For example, with 100 observations per cluster, a CRT will require fewer clusters than an SWT with no observations outside rollout if $ICC < 0.005$.

Alternatively, a CRT can be compared to a specific SWT with k sequences and no observations outside rollout. The CRT will require a smaller sample size when

$$\rho < \frac{1}{\frac{(k+1)}{(k-1)}m + 1}$$

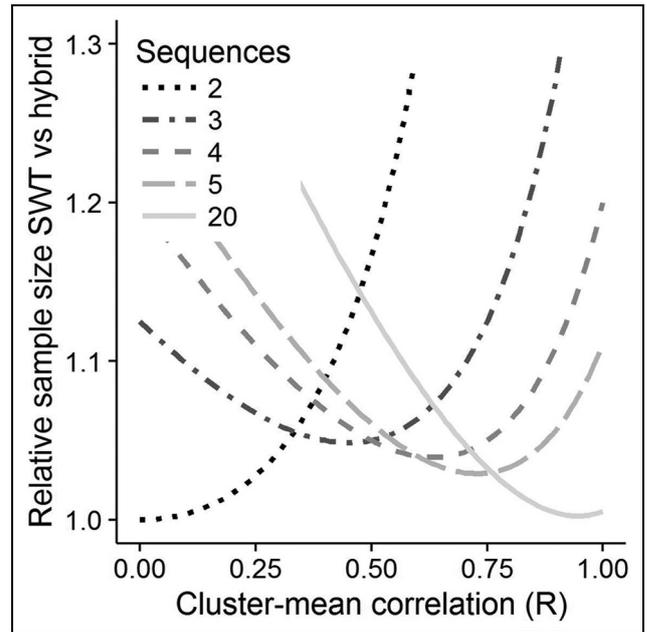


Figure 3. Graph of the sample size of the SWT with no observations outside rollout, relative to the optimised hybrid against the cluster-mean correlation. Darkest and dotted line = 2 sequences, lightest and solid line = 20 sequences. The optimal SWT is the lowest line at any given cluster-mean correlation.

In Appendix 5, we show that a CRT with baseline observations will always require the same or a larger sample size than the optimised SWT.

Comparison to a hybrid design

Figure 3 shows the relative sample size of the SWT with no observations outside rollout and 3, 4, 5, or 20 sequences compared to the optimised hybrid design. The optimal SWT, with the optimal number of sequences, is the lowest line at any value of R . For example, at $R = 0.2$, two sequences are optimal, but at $R = 0.7$, five sequences are optimal. While the hybrid always has the smaller sample size of the two designs, the differences are small when compared to the optimal SWT, and the optimal SWT requires at most a 5% larger sample size than the hybrid design.

Other pragmatic SWT designs: a non-optimal number of sequences and including observations outside rollout

It may not always be practical to use the optimal number of sequences calculated in equation (3) as this may be a large number. The primary constraint on the number of sequences is that it cannot exceed the number of clusters in the trial, and the number of periods in the trial cannot exceed the total cluster size. Furthermore, in many settings, the logistical effort to implement the

intervention at many different time points would be too great.

In such cases, a smaller, feasible number of sequences could be selected and there may then be some gain from obtaining observations outside rollout. For a fixed number of sequences, the optimal proportion of observations outside rollout (see Appendix 2 for derivation) is a function of the number of sequences and the cluster-mean correlation, as shown in equation (5)

$$\alpha + \beta = \begin{cases} 1 - \frac{(k-1)1}{k} \frac{1}{R} & \text{if } R \geq \frac{(k-1)}{k} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

For low values of the cluster-mean correlation, that is, low ICC or small total cluster size, it is optimal to have no observations outside rollout, and for higher values, as the ICC or total cluster size increases, the optimal proportion outside rollout increases up to a proportion of $1/k$.

This proportion varies between 0 (no observations outside rollout) and $1/k$ (equivalent to the same number of observations outside rollout as in one period of the trial). This tells us that the standard SWT, with $1/k$ observations before and $1/k$ observations after rollout, is inefficient.

For an SWT with the proportion of observations outside rollout selected from equation (5), increasing the number of sequences reduces the sample size of the design (see Appendix 7). However, there is little gain from increasing past five sequences, after which there is a maximum 4% further reduction in the sample size. This is only true while $R \geq (k-1)/k$, equivalent to $k < 1/(1-R)$. When the number of sequences passes this threshold, the sample size is smallest with no observations outside rollout and increasing the number of sequences will only continue to reduce the sample size up to the optimal number of sequences from equation (3). Appendix 7 contains comparisons of an SWT with this proportion of the observations outside rollout to the other trial designs considered in this article.

Selecting an SWT design

One strategy for selecting an SWT design with an equal number of clusters in each sequence and an equal number of observations in each period is as follows:

1. Calculate the optimal number of sequences using equation (3).
2. If the number of sequences is feasible, then you have the optimal SWT design by selecting this number of sequences and collecting no observations outside of rollout.
3. If the number of sequences is unfeasibly high, select the number closest to this value that is feasible. Then, compare the cluster-mean correlation to the

chosen number of sequences using equation (5) to see whether there is any gain from including observations outside rollout.

Several iterations of designs may be needed to achieve an equal number of clusters in each sequence, varying the numbers of clusters or sequences, so the former is a multiple of the latter. Iterations may also be required to achieve an equal number of observations in each period, varying the total cluster size and number of periods, so the former is a multiple of the latter. Alternatively, the Stata command by Hemming et al. can be used to calculate the power of an unbalanced design.⁴ Once the most appropriate SWT design has been identified, the sample size can be compared with other potential designs such as the CRT and hybrid design if these are feasible.

Example

Consider a CRT designed to yield 80% power to detect a mean difference of 0.1 in a continuous outcome with a total variance of 1, using a two-sided test at the 5% significance level. The ICC is 0.04, and the total number of observations per cluster is 84. Table 1 shows the number of clusters required to achieve 80% power by several designs. For each design, we give the number of clusters given by the relevant design effect and the number of clusters and power after allowing for an equal number of clusters allocated to each sequence. For the power of some designs, we also made small changes to the total cluster size so that there are an equal number of observations in each period of the trial.

The optimised SWT has eight sequences (and no observations outside rollout). After adjusting the number of clusters to get the same number in each sequence, this design required 88 clusters. Increasing the number of sequences to 88 (one cluster randomised to each sequence) resulted in the design effect giving a larger required number of clusters; this design is an impractical design only given to show that the required number of clusters does not decrease with more sequences. Other SWT designs required between 96 and 99 clusters to achieve 80% power.

A CRT requires almost twice as many clusters as the optimised SWT (162 clusters), and a CRT with baseline observations requires 112 clusters. As expected, the optimised hybrid design, with 78% of clusters randomised to an SWT with 17 sequences, required slightly fewer clusters than the optimised SWT.

Discussion

We have shown that the sample size of an SWT under equal allocation to sequences can be minimised by collecting all observations within rollout. Unlike the standard SWT, in this optimised SWT, the optimal number

Table 1. Illustrative example of the number of clusters required by different designs to achieve 80% power to detect a difference of 0.1 with standard deviation of 1.

Design	Calculated number of clusters	Final design		
		Number of clusters after rounding	Total cluster size ^a	Power (%)
Optimised SWT				
8 sequences, no observations outside rollout	86.1	88	84	81
Other SWT designs				
88 sequences, no observations outside rollout	87.7	88	87	81
8 sequences, 22% outside rollout (standard SWT)	94.0	96	81	80
3 sequences, no observations outside rollout	96.9	99	84	81
3 sequences, optimal outside rollout (14%)	94.2	96	84	81
Other designs				
CRT	161.5	162	84	80
CRT with optimal proportion of observations at baseline (36%)	111.6	112	84	80
Hybrid: 78% 17-sequence SWTs (optimal ^b)	84.8	86: 68 SWTs, 18 CRTs	85	81

CRT: parallel cluster-randomised trial; SWT: stepped wedge trial.

Total cluster size = 84, intracluster correlation coefficient (ICC) = 0.04, 5% significance level and 80% power.

Difference in calculated number of clusters and final number of clusters is due to rounding up and the requirement for an equal number of clusters per sequence.

^aFor power calculations, the total cluster size had to be varied for some of these designs to allow an equal number of observations in each period of the trial.

^bThe optimal number of sequences was 68, which gave a calculated number of clusters of 84.7. For 17 sequences, the calculated number is higher, but the final number of clusters required was the same as for 68 sequences and allowed a total cluster size similar to the other designs being considered.

of sequences depends on the cluster-mean correlation. We have also provided advice on when to consider other trial designs, acknowledging that a hybrid design will be always slightly more efficient.

Our finding that the most efficient SWT design is to have no observations outside rollout, at least if the resulting optimal number of sequences is also feasible, has not been suggested previously. This optimised SWT may, however, be unacceptable because not all the clusters will receive the intervention during the trial. Trialists may want to include some observations after rollout to avoid a ‘disappointment effect’ in the clusters that would not otherwise receive the intervention. Alternatively, the intervention could still be implemented after data collection has been completed.

We found that there were an optimal number of sequences for minimising the sample size of the SWT with no observations outside rollout. The number was large when the cluster-mean correlation was high (high ICC or large total cluster size) but small when the cluster-mean correlation was low (small ICC and small total cluster size). This contrasts with previous research for the standard SWT which showed that the sample size reduced as the number of sequences increased.^{2,3} It is, however, consistent with the consensus in the literature and finding of this study that a CRT requires a smaller sample size than an SWT when the ICC and total cluster size are low.^{7,10}

We examined the optimal proportion of baseline observations in a CRT. We found that when the

cluster-mean correlation is low, there is no benefit for the power of the study from including baseline observations. This is because when the ICC is high, the baseline observations will explain more of the variability in the follow-up measurements than when the ICC is low. Our results differed to much of the current literature that suggests that there is always a benefit to including baseline measurements.^{14,15} In this literature, total cluster size was not held constant – instead, baseline observations were included as additional observations relative to a design with no baseline.

This article is the first to compare the sample size implications of increasing the proportion of observations outside rollout versus increasing the number of sequences. We have found that increasing the number of sequences can have a larger impact on the sample size than increasing the proportion of observations outside rollout. For example, there is a larger reduction in sample size (providing the ICC and total cluster size are large enough) going from a CRT to an SWT with three sequences and no observations outside rollout than adding baseline observations to a CRT.

We found that the optimal number of sequences quickly increased with the ICC and total cluster size to a number that may not be practical. In cases such as this where a non-optimal number of sequences is chosen, we found that observations outside rollout may compensate and provide a reduction in the sample size; however, it is never beneficial to the sample size to have more observations outside rollout than are collected in

one period of the trial, similar to the results from Girling and Hemming.⁷

Some recently published SWTs included a large proportion of data outside of rollout, usually with the justification of investigating the longer-term effect of the intervention.^{1,16} These designs will give a larger variance for the intervention effect than our optimised SWT design with the same number of observations would have done. Trialists should also be aware that with no control observations after rollout, it will be difficult to assess whether changes in the outcome are due to changes in the intervention effect or other reasons. Our design effect assumes that the intervention effect remains constant throughout the trial. If this is not expected to be the case, different methods of sample size calculation, such as simulations,⁵ and more complex analysis methods should be used.

We found that the hybrid design was more efficient than the optimised SWT, as expected, due to its additional flexibility to allow unequal allocation to sequences. However, the gain in efficiency from this flexibility was at most 5%. Therefore, where considerable additional resources would be required to implement the intervention in a larger number of clusters at the start of the trial than at subsequent switches, the hybrid design will be unattractive. This might be the case if, for example, there is only one team available to roll the intervention out. The optimised hybrid design does not, however, always allocate more clusters to implement the intervention immediately than to other sequences, so one approach to design is to first see whether the optimised hybrid is feasible, and if not, then consider the optimised SWT under equal allocation.

We have given comparisons to some alternative designs, but there are many designs that we have not included. We have not explored incomplete designs such as the dog-leg design or unbalanced SWTs.^{17,18} We have compared trial designs fixing the total cluster size, but a further area of research could vary the total cluster size and fix the number of clusters or look to minimise a combination of the two. In some settings, there may be little or no cost associated with collecting observations before or after rollout, for example, with routinely collected data. If this is the case, it may be more informative to compare trial designs for a given cost rather than a fixed total cluster size.

As with all design effects, the assumptions made about the data must hold for the design effect to be valid, such as exchangeability within clusters and time trends that are common to all clusters. If these assumptions do not hold, using the design effect given here may result in an underpowered trial as the assumed analysis model would be inappropriate. These assumptions have sometimes been criticised as being unrealistic, and others have provided design effects where some

assumptions have been relaxed.^{19–21} Baio et al. found the assumption of normality affected sample size calculations for binary outcomes.⁵

Power is only one consideration of many when selecting a trial design. Caution should also be used in designing trials with very few clusters; among other issues, this may reduce generalisability and increase the possibility of chance imbalances.²² The lower sample size requirements of SWT and hybrid designs compared to a CRT come at the cost of requiring assumptions about how the outcome is changing over time because the intervention effect is confounded with time. Care needs to be taken to ensure that these assumptions are appropriate and that the analysis takes this into account adequately.¹⁹

We have identified SWT designs that require fewer clusters than the standard SWT and facilitated comparisons of statistical power between competing trial designs. Following our guidance on selecting a design will result in more efficient trials.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was supported by the Medical Research Council Network of Hubs for Trials Methodology Research (MR/L004933/1-P27) to J.A.T. and by the UK Medical Research Council (MC_UU_12023/29) to A.C. The funders had no involvement in the development of this article.

References

1. Copas AJ, Lewis JJ, Thompson JA, et al. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials* 2015; 16: 352.
2. Woertman W, de Hoop E, Moerbeek M, et al. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol* 2013; 66: 752–758.
3. Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007; 28: 182–191.
4. Hemming K and Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stat Med* 2014; 14: 363–380.
5. Baio G, Copas A, Ambler G, et al. Sample size calculation for a stepped wedge trial. *Trials* 2015; 16: 354.
6. Hemming K and Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *J Clin Epidemiol* 2016; 69: 137–146.
7. Girling AJ and Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med* 2016; 35: 2149–2166.

8. Beard E, Lewis JJ, Copas A, et al. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 2015; 16: 353.
9. De Hoop E, Woertman W and 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. *J Clin Epidemiol* 2013; 66: 1428.
10. Hemming K and Girling A. The efficiency of stepped wedge vs. cluster randomized trials: stepped wedge studies do not always require a smaller sample size. *J Clin Epidemiol* 2013; 66: 1427–1428.
11. Kotz D, Spigt M, Arts IC, et al. The stepped wedge design does not inherently have more power than a cluster randomized controlled trial. *J Clin Epidemiol* 2013; 66: 1059–1060.
12. Kish L. *Survey sampling*. New York: John Wiley & Sons, 1965.
13. Hemming K, Lilford R and Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med* 2015; 34: 181–196.
14. Teerenstra S, Eldridge S, Graff M, et al. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med* 2012; 31: 2169–2178.
15. Borm GF, Fransen J and Lemmens WA. A simple sample size formula for analysis of covariance in randomized clinical trials. *J Clin Epidemiol* 2007; 60: 1234–1238.
16. Martin J, Taljaard M, Girling A, et al. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. *BMJ Open* 2016; 6: e010166.
17. Hooper R and Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ* 2015; 350: h2925.
18. Lawrie J, Carlin JB and Forbes AB. Optimal stepped wedge designs. *Stat Probab Lett* 2015; 99: 210–214.
19. Davey C, Hargreaves J, Thompson JA, et al. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials* 2015; 16: 358.
20. Hayes RJ and Moulton LH. *Cluster randomised trials*. 1st ed. Boca Raton, FL: Chapman and Hall/CRC, 2009.
21. Hooper R, Teerenstra S, de Hoop E, et al. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med* 2016; 35: 4718–4728.
22. Taljaard M, Teerenstra S, Ivers NM, et al. Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clin Trials* 2016; 13: 459–463.

Appendix

Appendix 1: Derivation of the Design Effect

Throughout this appendix we will use the following notation (which is different in some cases to the rest of this thesis):

- i is the number of clusters in each sequence,
- k is the number of sequences,
- $z = ik$ is the number of clusters
- X_{lj} is an indicator variable of whether cluster l is receiving the intervention in period j
- b is the number of observations before rollout relative to the number collected in each period, so if twice as many are collected before rollout this would be 2.
- a is the number of observations are collected after rollout relative to the number collected in each period.
- β is the proportion of the total cluster size that are before rollout
- α is the proportion of the total cluster size that are after rollout
- $\delta = \alpha + \beta$ is the proportion of the total cluster size that is outside rollout
- n is the number of observations collected in each cluster in each period.
- m is the total cluster size.
- ρ is the ICC.

The formula for the variance of the intervention effect from the Hussey and Hughes analysis model [1]:

$$Var(\hat{\theta}) = \frac{I \frac{\sigma_w^2}{n} \left(\frac{\sigma_w^2}{n} + T \sigma_b^2 \right)}{(IU - W) \frac{\sigma_w^2}{n} + (U^2 + ITU - TW - IV) \sigma_b^2}$$

where σ_w^2 is the within cluster variance, and σ_b^2 is the between-cluster variance of the cluster means in a given period, and the other terms are defined as follows:

$$\begin{aligned}
 I &= ik \\
 T &= b + (k - 1) + a \\
 U &= \sum_{lj} X_{lj} \\
 W &= \sum_j (\sum_l X_{lj})^2 \\
 V &= \sum_l (\sum_j X_{lj})^2
 \end{aligned}$$

We want this formula in terms of the total variance σ^2 , so can use the following substitutions (see Woertman *et al* [2] for derivations):

$$\begin{aligned}
 \frac{\sigma_w^2}{n} &= \frac{(1 - \rho) \sigma^2}{n} \\
 \sigma_b^2 &= \rho \sigma^2 = \frac{n \rho \sigma^2}{n}
 \end{aligned}$$

$$\begin{aligned}
 Var(\hat{\theta}) &= \frac{I \frac{(1-\rho)\sigma^2}{n} \left(\frac{(1-\rho)\sigma^2}{n} + T \frac{n\rho\sigma^2}{n} \right)}{(IU - W) \frac{(1-\rho)\sigma^2}{n} + (U^2 + ITU - TW - IV) \frac{n\rho\sigma^2}{n}} \\
 &= \frac{I(1-\rho)\sigma^2}{n} \frac{((1-\rho) + Tn\rho)}{(IU - W)(1-\rho) + (U^2 + ITU - TW - IV)n\rho}
 \end{aligned}$$

We can work out the component parts to substitute into this:

$$\begin{aligned}
 U &= i((1 + 2 + \dots + (k - 1)) + ak) \\
 &= i\left(\frac{1}{2}k^2 - \frac{1}{2}k + ak\right)
 \end{aligned}$$

$$\begin{aligned}
 W &= (1i)^2 + (2i)^2 + \dots + ((k - 1)i)^2 + a(ki)^2 \\
 &= i^2 \left(\frac{k^3}{3} - \frac{k^2}{2} + \frac{k}{6} + ak^2 \right)
 \end{aligned}$$

$$\begin{aligned}
 V &= i(a)^2 + i(a + 1)^2 + i(a + 2)^2 + \dots + i(a + (k - 1))^2 \\
 &= i\left(\frac{k^3}{3} - \frac{k^2}{2} + \frac{k}{6} + a^2k + ak^2 - ak\right)
 \end{aligned}$$

Now we can combine these to get the component parts:

$$\begin{aligned}
IU - W &= ik \left(i \left(\frac{1}{2}k^2 - \frac{1}{2}k + ak \right) \right) - i^2 \left(\frac{k^3}{3} - \frac{k^2}{2} + \frac{k}{6} + ak^2 \right) \\
&= \frac{i^2}{6}k(k^2 - 1) \\
U^2 &= i^2k^2 \left(\frac{1}{2}k - \frac{1}{2} + a \right)^2 \\
&= i^2k \left(\frac{1}{4}k^3 - \frac{1}{2}k^2 + ak^2 - ak + a^2k + \frac{1}{4}k \right) \\
ITU &= ik(b + k - 1 + a) \left(i \left(\frac{1}{2}k^2 - \frac{1}{2}k + ak \right) \right) \\
&= i^2k \left(\frac{1}{2}bk^2 - \frac{1}{2}bk + abk + \frac{1}{2}k^3 - k^2 + \frac{1}{2}k + \frac{3}{2}ak^2 - \frac{3}{2}ak + a^2k \right) \\
TW &= (b + k - 1 + a) \left(i^2 \left(\frac{k^3}{3} - \frac{k^2}{2} + \frac{k}{6} + ak^2 \right) \right) \\
&= i^2k \left(\frac{1}{3}bk^2 - \frac{1}{2}bk + \frac{1}{6}b + abk + \frac{1}{3}k^3 - \frac{5}{6}k^2 + \frac{4}{6}k - \frac{1}{6} + \frac{4}{3}ak^2 \right) \\
&\quad + i^2k \left(-\frac{3}{2}ak + \frac{1}{6}a + a^2k \right) \\
IV &= i^2k \left(\frac{k^3}{3} - \frac{k^2}{2} + \frac{k}{6} + a^2k + ak^2 - ak \right) \\
U^2 - IV &= i^2k \left(\frac{1}{4}k^3 - \frac{1}{2}k^2 + ak^2 - ak + a^2k + \frac{1}{4}k \right) \\
&\quad - i^2k \left(\frac{k^3}{3} - \frac{k^2}{2} + \frac{k}{6} + a^2k + ak^2 - ak \right) \\
&= i^2k \left(-\frac{1}{12}k^3 + \frac{1}{12}k \right) \\
ITU - TW &= i^2k \left(\frac{1}{2}bk^2 - \frac{1}{2}bk + abk + \frac{1}{2}k^3 - k^2 + \frac{1}{2}k + \frac{3}{2}ak^2 - \frac{3}{2}ak + a^2k \right) \\
&\quad - i^2k \left(\frac{1}{3}bk^2 - \frac{1}{2}bk + \frac{1}{6}b + abk + \frac{1}{3}k^3 - \frac{5}{6}k^2 + \frac{4}{6}k - \frac{1}{6} + \frac{4}{3}ak^2 \right) \\
&\quad - i^2k \left(-\frac{3}{2}ak + \frac{1}{6}a + a^2k \right) \\
&= i^2k \left(\frac{1}{6}bk^2 - \frac{1}{6}b + \frac{1}{6}k^3 - \frac{1}{6}k^2 - \frac{1}{6}k + \frac{1}{6} + \frac{1}{6}ak^2 - \frac{1}{6}a \right) \\
U^2 + ITU - TW - IV &= i^2k \left(-\frac{1}{12}k^3 + \frac{1}{12}k + \frac{1}{6}bk^2 - \frac{1}{6}b + \frac{1}{6}k^3 - \frac{1}{6}k^2 \right) \\
&\quad + i^2k \left(-\frac{1}{6}k + \frac{1}{6} + \frac{1}{6}ak^2 - \frac{1}{6}a \right) \\
&= \frac{i^2k(k^2 - 1)}{6} \left(b + \frac{1}{2}k - 1 + a \right)
\end{aligned}$$

Substituting this into the variance formula gives:

$$\begin{aligned}
 Var(\hat{\theta}) &= \frac{I(1-\rho)\sigma_t^2}{n} \frac{((1-\rho) + Tn\rho)}{(IU - W)(1-\rho) + (U^2 + ITU - TW - IV)n\rho} \\
 &= \frac{ik(1-\rho)\sigma_t^2((1-\rho) + (b+k-1+a)n\rho)}{n \left[\left(\frac{i^2}{6}k(k^2-1) \right) (1-\rho) + \left(\frac{i^2}{6}k(k^2-1) \left(b + \frac{1}{2}k - 1 + a \right) \right) n\rho \right]} \\
 &= \frac{(1+\rho)(bn+kn+an-n-1)}{\left(1 + \rho \left(bn + \frac{1}{2}kn + an - n - 1 \right) \right)} \frac{6(1-\rho)\sigma_t^2}{\left(k - \frac{1}{k} \right) nik}
 \end{aligned}$$

To give this formula in terms of the total cluster size m , and proportions before and after rollout β and α respectively, we can use the following substitutions:

$$\begin{aligned}
 m &= bn + nk - n + an \\
 \beta &= \frac{bn}{m} \\
 \alpha &= \frac{an}{m}
 \end{aligned}$$

So the variance becomes

$$\begin{aligned}
 Var(\hat{\theta}) &= 6\sigma^2 \frac{(1+\rho(m-1))}{\left(1 + \rho \left(m - 1 - \frac{1}{2}kn \right) \right)} \frac{(1-\rho)}{\left(k - \frac{1}{k} \right)} \frac{1}{nik} \\
 &= 6\sigma^2 \frac{(1+\rho(m-1))}{\left(1 + \rho \left(m - 1 - \frac{1}{2}kn \right) \right)} \frac{(1-\rho)}{\left(k - \frac{1}{k} \right)} \frac{1}{z \frac{m(1-\beta-\alpha)}{(k-1)}} \\
 &= 6\sigma^2 \frac{(1+\rho(m-1))}{\left(1 + \rho \left(m - 1 - \frac{1}{2}kn \right) \right)} \frac{(1-\rho)}{\left(k - \frac{1}{k} \right)} \frac{k-1}{zm(1-\beta-\alpha)} \\
 &= \frac{6\sigma^2}{mz} (1+\rho(m-1)) \frac{k}{(k+1)} \\
 &\quad \times \frac{(1-\rho)}{(1-\beta-\alpha) \left(1 + (m-1)\rho - \frac{1}{2} \frac{k}{k-1} (1-\beta-\alpha)\rho m \right)} \quad (6.1)
 \end{aligned}$$

The variance of a trial with this sample size if it was individually randomised would be:

$$Var(\theta_{ind}) = \frac{4\sigma^2}{mz}$$

So our design effect becomes:

$$DE = \frac{3}{2} (1+\rho(m-1)) \frac{k}{(k+1)} \frac{(1-\rho)}{(1-\beta-\alpha) \left(1 + (m-1)\rho - \frac{1}{2} \frac{k}{k-1} (1-\beta-\alpha)\rho m \right)}$$

We can rewrite this in terms of the cluster mean correlation defined by Girling

and Hemming [3]:

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

Our design effect becomes:

$$DE = \frac{3}{2} (1 + \rho(m - 1)) \frac{k(k - 1)}{(k + 1)} \frac{(1 - R)}{(1 - \beta - \alpha)(k(1 - 0.5R(1 - \beta - \alpha)) - 1)}$$

In this paper we discuss minimising the design effect in terms of the number of sequences, k , and the proportion outside of rollout $\beta + \alpha$. Note that to do this only the last 2 terms of the DE are manipulated. All other parts of the design effect are held constant. Also note that these last 2 terms are the same as the last 2 terms in the variance formula equation (6.1). So minimising the design effect with respect to k and $\alpha + \beta$ is equivalent to minimising the variance. We can also rearrange equation (6.1) to give a formula for the number of clusters, these last 2 terms remain the same with all other terms remaining constant regardless of the values of k and $\beta + \alpha$. This means that the results we find for the values of k and $\beta + \alpha$ to minimise the design effect give the optimal values to minimise the number of clusters, or the variance.

Appendix 2: Optimal Number of Sequences and Optimal Proportion of Observations Outside Rollout

We want to find the combination of number of sequences k and the proportion outside rollout, $\alpha + \beta$, that minimises the design effect. We can do this by partially differentiating the design effect, firstly in terms of the number of sequences to get an equation for the optimal k , and secondly in terms of $\alpha + \beta$ to get an equation for the optimal $\alpha + \beta$. We will then solve these 2 equations simultaneously to find an optimum design for a given m and ρ .

We have boundaries on these values so we also need to check for optimal values at the boundaries. The boundaries are:

$$k \geq 2$$

$$0 \leq \alpha + \beta \leq 1$$

Let $\delta = \alpha + \beta$ be the proportion of observations that are outside rollout.

Substituting this into the design effect gives:

$$DE = (1 + \rho(m - 1)) \frac{3k}{2(k + 1)} \frac{(1 - \rho)}{(1 - \delta) \left(1 + (m - 1)\rho - \frac{1}{2} \frac{k}{k-1} (1 - \delta) \rho m\right)}$$

Optimising the number of sequences k

Although k will be an integer, we will treat it as continuous and assume that the optimal k will be one of the integers either side of the identified continuous optimal value. This means that we can differentiate the design effect with respect to k to get an equation for the optimal number of sequences for a given δ :

$$\begin{aligned} DE &= (1 + \rho(m - 1)) \frac{3k}{2(k + 1)} \frac{(1 - \rho)}{(1 - \delta) \left(1 + (m - 1)\rho - \frac{1}{2} \frac{k}{k-1} (1 - \delta) \rho m\right)} \\ \frac{d(DE)}{dk} &= \frac{3(1 + \rho(m - 1))(1 - \rho)}{2(1 - \delta)} \\ &\quad \times \frac{d}{dk} \left(\frac{k}{(k + 1) \left(1 + (m - 1)\rho - \frac{1}{2} \frac{k}{k-1} (1 - \delta) \rho m\right)} \right) \\ &= \frac{3(1 + \rho(m - 1))(1 - \rho)}{2(1 - \delta)} \\ &\quad \times \frac{k^2(1 + \rho(m\delta - 1)) - 2k(1 + \rho(m - 1)) + (1 + \rho(m - 1))}{\left((k^2 - 1)(1 + \rho(m - 1)) - \frac{1}{2}(1 - \delta)\rho mk(k + 1)\right)^2} \end{aligned}$$

The optimal number of sequences is when the derivative is equal to zero. So:

$$k^2(1 + \rho(m\delta - 1)) - 2k(1 + \rho(m - 1)) + (1 + \rho(m - 1)) = 0$$

$$\begin{aligned} k &= \frac{2(1 + \rho(m - 1)) \pm \sqrt{(2(1 + \rho(m - 1)))^2 - 4(1 + \rho(m\delta - 1))(1 + \rho(m - 1))}}{2(1 + \rho(m\delta - 1))} \\ &= \frac{(1 + \rho(m - 1)) \pm \sqrt{(1 - \delta)\rho m(1 + \rho(m - 1))}}{(1 + \rho(m\delta - 1))} \end{aligned}$$

There are 2 solutions here. The negative square root will only give a number of sequences greater than 2 when :

$$\rho < \frac{1}{(3m + 1)}$$

This suggests that the negative square root does not give the minimum. Graphical inspection and numerical example support that the positive square root gives a minimum value of the design effect. So the optimal number of sequences fixing all other parameters is:

$$k = \frac{(1 + \rho(m - 1)) + \sqrt{(1 - \delta)\rho m(1 + \rho(m - 1))}}{(1 + \rho(m\delta - 1))}$$

Optimising the proportion outside rollout $\delta = \alpha + \beta$

To find the minimum of the DE with respect to $\delta = \alpha + \beta$, differentiate with respect to δ :

$$\begin{aligned} \frac{d(DE)}{d\delta} &= (1 + \rho(m - 1))(1 - \rho) \frac{3k}{2(k + 1)} \\ &\quad \times \frac{d}{d\delta} \left(\frac{1}{(1 - \delta) \left((1 + \rho(m - 1)) - \frac{1}{2} \frac{k}{k-1} \rho m (1 - \delta) \right)} \right) \\ &= (1 + \rho(m - 1))(1 - \rho) \frac{3k}{2(k + 1)} \\ &\quad \times \left(\frac{(1 + \rho(m - 1)) - \frac{k}{k-1} \rho m (1 - \delta)}{(1 - \delta)^2 \left((1 + \rho(m - 1)) - \frac{1}{2} \frac{k}{k-1} \rho m (1 - \delta) \right)^2} \right) \end{aligned}$$

The turning point is when:

$$\begin{aligned} \frac{d(DE)}{d\delta} &= 0 \\ \Rightarrow \delta &= 1 - \frac{(k - 1)(1 + \rho(m - 1))}{\rho m k} \end{aligned}$$

Taking the second derivative shows that this is a minimum.

δ can get infinitely small but this is not possible under our constraint of $\delta \geq 0$ so we must limit this value to 0 for any value smaller than 0. This happens when:

$$\frac{(k - 1)(1 + \rho(m - 1))}{k \rho m} > 1$$

Or equivalently

$$\frac{\rho m}{(1 + \rho(m - 1))} < \frac{k - 1}{k}$$

We can write this as the proportion outside rollout fixing all other parameters:

$$\delta = \alpha + \beta = \begin{cases} 1 - \frac{(k-1)(1+\rho(m-1))}{\rho m}, & \frac{\rho m}{(1+\rho(m-1))} \geq \frac{k-1}{k} \\ 0 & \textit{otherwise} \end{cases}$$

Since $1 < \frac{(1+\rho(m-1))}{\rho m} < \infty$, this optimal $\alpha + \beta$ has boundaries:

$$0 \leq \alpha + \beta \leq \frac{1}{k}$$

This can also be written in terms of the cluster mean correlation R:

$$\delta = \alpha + \beta = \begin{cases} 1 - \frac{(k-1)}{k} \frac{1}{R}, & R \geq \frac{k-1}{k} \\ 0 & \textit{otherwise} \end{cases}$$

Solving the equations

We now have the simultaneous equations:

$$\begin{aligned} \delta &= 1 - \frac{(k-1)(1+\rho(m-1))}{\rho m k} \\ k &= \frac{(1+\rho(m-1)) + \sqrt{(1-\delta)\rho m(1+\rho(m-1))}}{(1+\rho(m\delta-1))} \end{aligned}$$

There is no solution to these simultaneous equations. This means that there is no minimum value within the boundaries. We must also check for optimal values at the boundaries of k and δ .

Boundary (1) $\delta = 0$

We have a boundary at $\delta = 0$, with all observations during rollout.

The equation for optimal k becomes:

$$k = \frac{(1+\rho(m-1)) + \sqrt{\rho m(1+\rho(m-1))}}{(1-\rho)}$$

Substituting this k and $\delta = 0$ into the design effect we get:

$$DE_{(1)} = 3 \frac{\left(2\rho m(1 + \rho(m-1)) + (1 + \rho(2m-1))\sqrt{\rho m(1 + \rho(m-1))}\right)}{\left((2 + \rho(m-2)) + \sqrt{\rho m(1 + \rho(m-1))}\right)} \\ \times \frac{(1 - \rho)(1 + \rho(m-1))}{\left(\rho m(1 + \rho(m-1)) + (2 + \rho(m-2))\sqrt{\rho m(1 + \rho(m-1))}\right)}$$

Boundary (2) $\delta = 1/k$

Although $0 < \delta < 1$, in our derivation of the optimal value of δ , we found that this had a boundary at $\delta = 1/k$

The design effect in this case is:

$$DE = (1 + \rho(m-1)) \frac{3k^2}{2(k^2-1)} \frac{(1-\rho)}{((1 + \rho(0.5m-1)))}$$

$\frac{k^2}{(k^2-1)}$ decreases as k increases so at this boundary the optimal design is to have a large number of sequences. The design effect here becomes:

$$DE_{(2)} = \frac{3(1-\rho)(1 + \rho(m-1))}{(2 + \rho(m-2))}$$

Boundary (3) $k = 2$

The final boundary to the design effect is the lower limit of k . The optimal value of δ when $k = 2$ is:

$$\delta = \begin{cases} 1 - \frac{1}{2} \frac{(1+\rho(m-1))}{\rho m}, & \frac{\rho m}{(1+\rho(m-1))} \geq \frac{1}{2} \\ 0 & otherwise \end{cases}$$

With these values the design effect becomes:

$$DE_{(3)} = \begin{cases} \frac{4\rho m(1-\rho)}{(1+\rho(m-1))}, & \frac{\rho m}{(1+\rho(m-1))} \geq \frac{1}{2} \\ (1 + \rho(m-1)) & otherwise \end{cases}$$

Which boundary is optimal?

We can work out which of these is optimal by looking at the ratio of these design effects.

$\delta = 0$ **vs** $\delta = 1/k$

Taking the ratio of the design effects when $\delta = 0$ and when $\delta = 1/k$ gives:

$$\frac{DE_{(2)}}{DE_{(1)}} = \frac{\left(2\rho m(1 + \rho(m-1)) + (1 + \rho(2m-1))\sqrt{\rho m(1 + \rho(m-1))}\right)}{\left((2 + \rho(m-2)) + \sqrt{\rho m(1 + \rho(m-1))}\right)} \times \frac{(2 + \rho(m-2))}{\left(\rho m(1 + \rho(m-1)) + (2 + \rho(m-2))\sqrt{\rho m(1 + \rho(m-1))}\right)}$$

We are interested in when this ratio is less than 1. Using Mathematica software [4] shows that this inequality is true for all $0 < \rho < 1$, $m > 0$.

This means that the design with no observations outside rollout and the optimal number of sequences for that design, will always be more efficient than the design with $1/k$ observations outside rollout and the optimal number of sequences for that design.

$\delta = 0$ **and optimal** k **vs** $k = 2$ **and optimal** δ

To compare the design on the $\delta = 0$ boundary to the design on the $k = 2$ boundary we need to split the comparison.

When $\frac{\rho m}{(1 + \rho(m-1))} \geq \frac{1}{2}$ so when $\rho > \frac{1}{(m+1)}$, we can look at the ratio of design effects for these 2 designs which gives:

$$\frac{DE_{(3a)}}{DE_{(1)}} = \frac{4\rho m \left((2 + \rho(m-2)) + \sqrt{\rho m(1 + \rho(m-1))} \right)}{3 \left(2\rho m(1 + \rho(m-1)) + (1 + \rho(2m-1))\sqrt{\rho m(1 + \rho(m-1))} \right)} \times \frac{\left(\rho m(1 + \rho(m-1)) + (2 + \rho(m-2))\sqrt{\rho m(1 + \rho(m-1))} \right)}{(1 + \rho(m-1))^2}$$

We can find out when this ratio of design effects is greater than 1, i.e. when the design effect at $k = 2$ optimal δ is smaller than the design effect at optimal k and $\delta = 0$, and rearrange to get:

$$\begin{aligned} & 2\rho m(1 + \rho(m-1)) \left(\rho^2(m-3)(m+1) + 2\rho(m+3) - 3 \right) \\ & + \left(\rho^3(2m^3 - 5m^2 + 4m + 3) + \rho^2(5m^2 - 8m - 9) \right) \sqrt{\rho m(1 + \rho(m-1))} \\ & + (\rho(4m + 9) - 3) \sqrt{\rho m(1 + \rho(m-1))} > 0 \end{aligned}$$

If each component of the sum is greater than zero, then there sum will also be greater than zero and the inequality will hold.

Using Mathematica software shows that:

$$\left(\rho^2(m-3)(m+1) + 2\rho(m+3) - 3\right) > 0$$

when $\rho > 1/(m+1)$. This is multiplied by 2 positive numbers so this term will be positive if $\rho > 1/(m+1)$.

Secondly:

$$\left(\rho^3(2m^3 - 5m^2 + 4m + 3) + \rho^2(5m^2 - 8m - 9) + \rho(4m + 9) - 3\right) > 0$$

when $\rho > 1/(m+1)$. Since the square root term is also positive, this whole second term will be positive when $\rho > 1/(m+1)$.

So all terms are positive and so the total is positive when $\rho > 1/(m+1)$. This means that, under this conditions, the design with $\delta = 0$ has the smaller design effect.

But, we are currently looking in the region where $\frac{\rho m}{(1+\rho(m-1))} \geq \frac{1}{2}$. In this region $\rho > 1/(m+1)$.

This means that in the region where $\frac{\rho m}{(1+\rho(m-1))} \geq \frac{1}{2}$, $DE_{(3a)} > DE_{(1)}$ and so the design effect is smaller at $\delta = 0$ and optimal k .

Now looking at when $\frac{\rho m}{(1+\rho(m-1))} < \frac{1}{2}$.

In this region, both designs have $\delta = 0$ so we are comparing a design with $\delta = 0$ and an optimal number of sequences that can equal 2 if that is optimal in $DE_{(1)}$, and a design with the number of sequences fixed to 2 and $\delta = 0$ in $DE_{(3b)}$. By definition the design when the number of sequences is allowed to vary will be either the same or more efficient than the design where the number of sequences is fixed to 2, so $DE_{(3b)} > DE_{(1)}$ and the design effect is smaller at $\delta = 0$ and optimal k .

Conclusion

We can conclude that the optimal stepped-wedge design is to have $\delta = 0$, or equivalently $\alpha + \beta = 0$, and k as:

$$k = \frac{(1 + \rho(m-1)) + \sqrt{\rho m(1 + \rho(m-1))}}{(1 - \rho)}$$

The equation for k can be written as:

$$k = \frac{1}{1 - \sqrt{R}}$$

Appendix 3 SWT with No Observations Outside Rollout Compared to a CRT

We will compare an SWT with no observations outside rollout to a CRT in 2 ways. Firstly we can compare the ratio of the design effects of a CRT to a specific SWT design. Secondly we can investigate when the optimal number of sequences in the SWT is < 2.5 so that the SWT becomes equivalent to a CRT.

Comparison of design effects

DE of SWT:

$$DE = (1 + \rho(m - 1)) \frac{3k}{2(k + 1)} \frac{(1 - \rho)}{\left((1 + \rho(m - 1)) - \frac{1}{2} \frac{k}{k-1} \rho m \right)}$$

DE of CRT:

$$DE = (1 + \rho(m - 1))$$

Note that when $k = 2$ the SWT DE cancels to the CRT DE as with no observations outside rollout and $k = 2$ the design is a CRT.

The ratio of the design effects is:

$$\frac{DE_{SWT}}{DE_{CRT}} = \frac{3k}{2(k + 1)} \frac{(1 - \rho)}{\left(1 + \left(m \left(1 - \frac{1}{2} \frac{k}{k-1} \right) - 1 \right) \rho \right)}$$

We are interested in when this ratio is less than 1, so the SWT design effect is smaller than the CRT design effect:

$$\begin{aligned} \frac{3k}{2(k + 1)} \frac{(1 - \rho)}{\left(1 + \left(m \left(1 - \frac{1}{2} \frac{k}{k-1} \right) - 1 \right) \rho \right)} &< 1 \\ \frac{3k(k - 1)(1 - \rho)}{(k - 1 + (m(0.5k - 1) - (k - 1))\rho)} &< 2(k + 1) \end{aligned}$$

Before multiplying by the left hand side denominator we need to check when it is positive:

$$(k - 1 + (m(0.5k - 1) - (k - 1))\rho) > 0$$

This is true when:

$$k > \frac{1 + \rho(m - 1)}{1 + \rho(0.5m - 1)}$$

$1 < \frac{1 + \rho(m - 1)}{1 + \rho(0.5m - 1)} < 2$ because $0 < \rho < 1$ and $m > 1$.

So the left hand side denominator is positive for all $k > 2$ so we can multiply the left hand side denominator to both sides of the inequality:

$$\begin{aligned} 3k(k - 1)(1 - \rho) &< 2(k + 1)(k - 1 + (m(0.5k - 1) - (k - 1))\rho) \\ (k - 1)(k - 2) &< \rho(m(k + 1)(k - 2) + (k - 1)(k - 2)) \end{aligned}$$

Since $k > 2$ we can divide through by $(k - 2)$:

$$\begin{aligned} (k - 1) &< \rho(m(k + 1) + (k - 1)) \\ \rho &> \frac{1}{\frac{(k+1)}{(k-1)}m + 1} \end{aligned}$$

Optimal number of sequences is two

Another way to assess when a CRT will require a larger sample size than an SWT this is to look at when the optimal number of sequences for the SWT is ≥ 2.5 . Note that is only approximately true, the function is not symmetrical so if the optimal number of sequences is 2.5, 2 sequences may be more efficient than 3 sequences in some cases. It should be right for the majority of cases, or good as a rule of thumb:

$$\text{optimal } k = \frac{(1 + \rho(m - 1)) + \sqrt{\rho m (1 + \rho(m - 1))}}{(1 - \rho)}$$

$$\begin{aligned} \frac{(1 + \rho(m - 1)) + \sqrt{\rho m (1 + \rho(m - 1))}}{(1 - \rho)} &\geq 2.5 \\ \sqrt{\rho m (1 + \rho(m - 1))} &\geq 2.5(1 - \rho) - (1 + \rho(m - 1)) \end{aligned}$$

In order to square the left hand side we need to know whether both sides are

positive. $\rho m(1 + \rho(m - 1)) > 0$ for all values but the right hand side could be positive or negative.

If right hand side is negative we cannot square both sides but the inequality will still hold. This is the case when:

$$\begin{aligned} 2.5(1 - \rho) - (1 + \rho(m - 1)) &\leq 0 \\ \rho &\geq \frac{3}{2m + 3} = \frac{9}{6m + 9} \end{aligned}$$

Otherwise, if the right hand side is positive, i.e.:

$$\rho < \frac{3}{2m + 3} = \frac{9}{6m + 9}$$

we can square both sides:

$$\begin{aligned} \rho m(1 + \rho(m - 1)) &\geq \left(\frac{5}{2}(1 - \rho) - (1 + \rho(m - 1))\right)^2 \\ 0 &\geq \rho^2(16m + 9) - \rho(16m + 18) + 9 \end{aligned}$$

$$\begin{aligned} \rho &= \frac{(16m + 18) \pm \sqrt{(16m + 18)^2 - 4 * 9(16m + 9)}}{2(16m + 9)} \\ &= 1 \text{ OR } \frac{9}{16m + 9} \end{aligned}$$

So we have $(\rho - 1)\left(\rho - \frac{9}{16m+9}\right) \leq 0$. Since $\rho < 1$ we are left with:

$$\rho \geq \frac{9}{16m + 9}$$

So our 2 solutions combine to show that the optimal number of sequences in the SWT is greater than 2.5 when:

$$\rho \geq \frac{9}{16m + 9}$$

Appendix 4 Optimal Proportion of Observations at Baseline in a CRT

Our results from Appendix A boundary (3) give the optimal proportion of observations outside rollout for a CRT:

$$\beta = \begin{cases} 1 - \frac{1}{2} \frac{(1+\rho(m-1))}{\rho m}, & \frac{\rho m}{(1+\rho(m-1))} \geq \frac{1}{2} \\ 0 & \textit{otherwise} \end{cases}$$

In terms of the cluster-mean correlation this is:

$$\beta = \begin{cases} 1 - \frac{1}{2R}, & R \geq \frac{1}{2} \\ 0 & \textit{otherwise} \end{cases}$$

Appendix 5 Comparison between the Optimal SWT and a CRT with Baseline Measurements

This comparison has been seen in Appendix 2, “Which boundary is optimal? $\delta = 0$ vs $k = 2$ ”. This was a comparison of an SWT with $\delta = 0$ and an optimised number of sequences (the optimal SWT design), and an SWT with $k = 2$ (or a CRT with and without baseline observations). We saw that in the region where baseline observations would be beneficial, i.e. when:

$$R = \frac{\rho m}{(1 + \rho(m - 1))} \geq \frac{1}{2}$$

the optimised SWT design effect was always smaller than the design effect of the CRT with baseline observations.

Appendix 6 Comparison between the Optimal SWT and the Hybrid Design

Comparison with an SWT with no observations outside rollout

Girling and Hemming 2016 [3] define a hybrid design, which consists of a proportion of the clusters contributing to the design as an SWT with $\delta = \alpha + \beta = 1/k$, and the remaining clusters contributing as a parallel CRT. We will compare this design, with a large number of sequences in the SWT and

an optimal proportion allocated to the SWT, to an SWT with $\delta = \alpha + \beta = 0$ and k sequences.

Let us define γ as the proportion of clusters assigned to the SWT.

The design effect for the hybrid design is [3]:

$$DE_H = \frac{\eta}{4(a_D - b_D R)}$$

where:

$$\begin{aligned} \eta &= 1 - \rho = (1 + \rho(m - 1))(1 - R) \\ 4a_D &= 1 - \frac{\gamma^2}{3} \left(1 + \frac{2}{k^2}\right) \\ 4b_D &= 1 - \frac{\gamma}{3} \left(2 + \frac{1}{k^2}\right) \\ R &= \frac{\rho m}{1 + \rho(m - 1)} \end{aligned}$$

So:

$$DE_H = \frac{(1 + \rho(m - 1))(1 - R)}{1 - \frac{\gamma^2}{3} \left(1 + \frac{2}{k^2}\right) - \left(1 - \frac{\gamma}{3} \left(2 + \frac{1}{k^2}\right)\right) R}$$

For the optimal hybrid design, $\gamma = R$ and there are many sequences so that $1/k^2 \approx 0$. The design effect becomes:

$$DE_H = \frac{3(1 + \rho(m - 1))(1 - R)}{3 - 3R + R^2}$$

Comparing this to our SWT with no observations outside rollout ($\delta = 0$) and k sequences gives:

$$\begin{aligned} \frac{DE_S}{DE_H} &= (1 + \rho(m - 1)) \frac{3k(k - 1)}{2(k + 1)} \frac{(1 - R)}{(k(1 - 0.5R) - 1)} \frac{3 - 3R + R^2}{3(1 + \rho(m - 1))(1 - R)} \\ &= \frac{k(k - 1)(3 - 3R + R^2)}{2(k + 1)(k(1 - 0.5R) - 1)} \end{aligned}$$

This is the function that is shown graphically in figure 3.

Appendix 7 Optimal Design of an SWT with Observations Outside Rollout

In appendix 2 we showed that for a fixed number of sequences k , the optimal proportion of observations outside rollout is:

$$\alpha + \beta = \begin{cases} 1 - \frac{(k-1)(1+\rho(m-1))}{k \rho m}, & \frac{\rho m}{(1+\rho(m-1))} \geq \frac{k-1}{k} \\ 0 & \textit{otherwise} \end{cases}$$

or in terms of R :

$$\alpha + \beta = \begin{cases} 1 - \frac{(k-1)}{k} \frac{1}{R}, & R \geq \frac{k-1}{k} \\ 0 & \textit{otherwise} \end{cases}$$

Substituting this into the design effect gives:

$$DE = \begin{cases} \frac{3k^2}{(k^2-1)} \frac{\rho m(1-\rho)}{(1+\rho(m-1))}, & \frac{\rho m}{(1+\rho(m-1))} \geq \frac{k-1}{k} \\ (1 + \rho(m-1)) \frac{3k}{2(k+1)} \frac{(1-\rho)}{(1+(m-1)\rho - \frac{1}{2} \frac{k}{k-1} \rho m)} & \textit{otherwise} \end{cases}$$

or in terms of R :

$$DE = \begin{cases} (1 + \rho(m-1)) \frac{3k^2}{(k^2-1)} R(1-R), & R \geq \frac{k-1}{k} \\ (1 + \rho(m-1)) \frac{3k(k-1)}{2(k+1)} \frac{(1-R)}{(k(1-0.5R)-1)} & \textit{otherwise} \end{cases}$$

Impact on sample size of increasing the number of sequences in an SWT with the optimal proportion of observations outside rollout

While $\frac{\rho m}{(1+\rho(m-1))} \geq \frac{k-1}{k}$ the number of sequences affects to the design effect by a factor of

$$\frac{k^2}{(k^2-1)}$$

This factor reduces as the number of sequences increases as show in table A1.

Table A1: Sample size of an SWT with the optimal proportion of observations outside rollout and with an increased number of sequences relative to an original number of sequences

	Increased number of sequences							
		2	3	4	5	6	7	8
Original number of sequences	2	1.00	0.84	0.80	0.78	0.77	0.77	0.76
	3		1.00	0.95	0.93	0.91	0.91	0.90
	4			1.00	0.98	0.96	0.96	0.95
	5				1.00	0.99	0.98	0.99
	6					1.00	0.99	0.99
	7						1.00	1.00
	8							1.00

The maximum relative difference in sample size after 5 sequences is

$$\frac{1}{\frac{5^2}{(5^2-1)}} = \frac{(5^2 - 1)}{5^2} = 0.96$$

Comparison with a CRT with baseline observations

We have previously stated that a CRT with baseline observations can be thought of as an SWT with 2 sequences. Since the design effect decreases as the number of sequences increase a CRT with baseline observations (or equivalently an SWT with 2 sequences and the optimal proportion outside rollout) will always require a larger sample size than an SWT with more sequences and an optimal proportion of observations outside rollout.

Comparison with a CRT

When $\frac{\rho m}{(1+\rho(m-1))} \geq \frac{k-1}{k}$ an SWT with k sequences may benefit from some observations being collected outside rollout. Within this region an SWT with k sequences and the optimal proportion of observations outside rollout will always have a smaller sample size than a CRT. We can show this by comparing the design effect from Appendix 7 to the design effect for a CRT:

$$\frac{DE_{SWT}}{DE_{CRT}} = \frac{3k^2}{(k^2 - 1)} \frac{\rho m (1 - \rho)}{(1 + \rho(m - 1))} \frac{1}{(1 + \rho(m - 1))}$$

We are interested in when the design effect of the SWT is smaller than the design effect of the CRT. When this is the case this ratio will be less than 1 so we can look at if and when this is true:

$$\frac{3k^2}{(k^2 - 1)} \frac{\rho m (1 - \rho)}{(1 + \rho(m - 1))} \frac{1}{(1 + \rho(m - 1))} < 1$$

Using Mathematica software shows that this is true when $0 < \rho < 1$, $m > 0$, and $k > 2$.

So when $\frac{\rho m}{(1 + \rho(m - 1))} \geq \frac{k - 1}{k}$ the SWT is more efficient than the CRT.

When $\frac{\rho m}{(1 + \rho(m - 1))} < \frac{k - 1}{k}$ the SWT will have the smallest sample size with no observations outside rollout, a comparison given in the paper.

Appendix Bibliography

- [1] Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials* 2007. 28. (2):182–191.
- [2] Woertman W, Hoop E de, Moerbeek M, Zuidema SU, Gerritsen DL and Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology* 2013. 66. (7):752–758.
- [3] Girling AJ and Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Statistics in Medicine* 2016. 35. (13):2149–2166.
- [4] Wolfram Research Inc. Mathematica, Version 11.1. 2017.

7 Paper B: Bias and Inference from Misspecified Mixed-Effect Models in Stepped-Wedge Trial Analysis

In the next three chapters I will address the second aim of this thesis: improving the robustness of SWT analysis.

This paper addresses objectives one and two of this thesis and investigates how sensitive model 3.1 is to deviations from the model assumptions, and explores alternative analysis models. In this Chapter, model 3.1 is referred to as the standard model.

At the start of this PhD, my literature review found limited published research into the performance of the standard model. Since then, published research has considered this issue, but only within a limited range of within-cluster correlation structures and period effect assumptions. There remains no literature on the impact of the intervention effect varying between clusters.

In order to identify whether or not a more robust analysis method was required for SWTs, I needed to examine whether the standard model was sensitive to the assumptions it made about the data. This paper focuses on the impact of assuming that either the period effect or intervention effect is common to all clusters, when these effects truly vary between clusters. I also demonstrate the weight given to horizontal comparison by the standard model.

Objective two of my second aim was to identify alternative analysis models. To address this objective, I also assessed two alternative analysis models to the standard model; models 3.4 and 3.6. Model 3.4, referred to as the random-period model, relaxes the assumption of a common period effect and exchangeability between periods. Model 3.6, referred to as the random-intervention model, relaxes the assumption of a common intervention effect

and exchangeability between the control and intervention. It is not possible to allow both the period and intervention effects to vary in the same model as this over-parameterises the model. Including these alternative models also aided interpretation of the results of the standard model; for each scenario, a correctly specified model can be directly compared to an incorrectly specified model.

The paper consists of a simulation study and a worked example; the example uses the deworming trial data described in section 2.1 and the simulation study was also based on this trial data. The paper uses the term “groups” to describe what have been called “sequences” elsewhere in this thesis.

Ethics approval is given in Appendix B. The paper has been peer reviewed and is published in *Statistics in Medicine*, titled “Bias and inference from misspecified mixed effect models in stepped-wedge trial analysis”, by myself, Katherine Fielding, Calum Davey, Alexander Aiken, James Hargreaves, and Richard Hayes. The paper is licensed under CC BY 2.0. The licence to reproduce the paper is given in Appendix C. The supporting information has been edited to improve consistency with the formatting of this thesis.

Registry

T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jennifer Thompson
Principal Supervisor	Katherine Fielding
Thesis Title	Improving the design and analysis of stepped-wedge trials

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	Statistics in Medicine		
When was the work published?	29/05/2017		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I designed and conducted the simulation study, and analysed and interpreted results. I wrote the first draft of the manuscript and the responses to reviewers comments.
--	---

Student Signature: _____

Date: 12/09/17

Supervisor Signature: _____

Date: 12/9/2017

Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis

Jennifer A. Thompson,^{a,b*†}  Katherine L. Fielding,^a
Calum Davey,^c Alexander M. Aiken,^a James R. Hargreaves^c and
Richard J. Hayes^a

Many stepped wedge trials (SWTs) are analysed by using a mixed-effect model with a random intercept and fixed effects for the intervention and time periods (referred to here as the standard model). However, it is not known whether this model is robust to misspecification.

We simulated SWTs with three groups of clusters and two time periods; one group received the intervention during the first period and two groups in the second period. We simulated period and intervention effects that were either common-to-all or varied-between clusters. Data were analysed with the standard model or with additional random effects for period effect or intervention effect. In a second simulation study, we explored the weight given to within-cluster comparisons by simulating a larger intervention effect in the group of the trial that experienced both the control and intervention conditions and applying the three analysis models described previously.

Across 500 simulations, we computed bias and confidence interval coverage of the estimated intervention effect.

We found up to 50% bias in intervention effect estimates when period or intervention effects varied between clusters and were treated as fixed effects in the analysis. All misspecified models showed undercoverage of 95% confidence intervals, particularly the standard model. A large weight was given to within-cluster comparisons in the standard model.

In the SWTs simulated here, mixed-effect models were highly sensitive to departures from the model assumptions, which can be explained by the high dependence on within-cluster comparisons. Trialists should consider including a random effect for time period in their SWT analysis model. © 2017 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Keywords: stepped wedge trials; cluster randomised trials; mixed-effect model; model misspecification; simulation study

1. Introduction

Recent reanalysis of a high-profile stepped wedge trial (SWT) has brought into question methods commonly used to analyse these complex studies [1–3]. SWTs are often analysed by using models that make strong assumptions about the clustering in the data [4]. It is currently unknown if estimates from these models are robust to deviations from these assumptions.

An SWT is a type of cluster randomised trial where clusters are randomised into groups. Each group begins to receive the intervention at a different time so that all clusters start the trial in the control condition, and by the end of the trial, all clusters are receiving the intervention.

^a Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, U.K.

^b MRC London Hub for Trials Methodology Research, London, U.K.

^c Department of Social and Environmental Health Research, London School of Hygiene and Tropical Medicine, London, U.K.

*Correspondence to: Jennifer Anne Thompson, Department of Infectious Disease, London School of Hygiene and Tropical Medicine, London, U.K.

†E-mail: jennifer.thompson@lshtm.ac.uk

Funded by MRC hub for trial methodology research (MR/L004933/1-P27) to Jennifer Thompson.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

The control and intervention conditions can, in principle, be compared in two directions known as the vertical and horizontal comparisons [4]. Vertical comparisons compare the outcomes of clusters in the intervention condition with the outcomes of clusters in the control condition within the same time period; because the order of rollout is randomised, each of these comparisons is randomised. Horizontal comparisons compare outcomes from periods in the intervention condition with outcomes from periods in the control condition in the same cluster; these are non-randomised before–after comparisons that are confounded with time period.

In practice, most analysis methods for SWTs incorporate information from both the vertical and horizontal comparisons in the intervention effect estimate and so need some way to adjust for period effects [4]. The most common analysis model (hereafter referred to as the standard model) is a mixed-effect model with a random intercept to account for clustering and adjusting for period effects as a fixed categorical variable; this model is described by Hussey and Hughes [5]. Despite its wide use, guidance for using this analysis model is lacking. The model makes strong assumptions about the correlation structure of the data: The intervention effect and the period effects are assumed to be common to all clusters. It is not currently known whether the intervention effect estimate and its precision are robust to misspecifying these assumptions.

In the context of SWTs, we are most interested in estimation of the intervention effect and how robust this effect is to misspecification of the intervention effect itself as well as misspecification of the period effect. Previous research has found that misspecifying the random effects led to biased effect estimates as well as biased precision of estimates [6]. In parallel cluster randomised trials with baseline measurements and in cluster crossover randomised trials, it has been shown that analyses with hierarchical models should include a random effect for period, sometimes referred to as a cluster-period interaction, to avoid residual confounding [7–10].

The importance of specifying the period effect correctly will depend on how much the horizontal comparisons contribute within the model. This has not been explored in the literature. If a large weight is given to this comparison, any residual confounding of the intervention effect by the period effects could lead to a biased estimate of the intervention effect.

In this paper, we will explore both issues with a simulation study comparing the standard model with other mixed-effect models, focusing on a binary outcome with cross-sectional measurements. We then run a second set of simulations to explore the weight given to horizontal comparisons by each analysis model. Following the simulation studies, we explore the impact of misspecifying analysis models in our motivating example.

2. Motivating example

There has been much debate in recent literature about the results of a reanalysis of a highly cited SWT that investigated the effect on school attendance of a mass deworming intervention for school children in Kenya [1–3]. The trial included 75 schools (clusters) that were randomised into three groups and ran over 2 years. School attendance was measured as a binary outcome with multiple observations for each individual child during each year. There was a geometric mean of 1180 (interquartile range (IQR) 908.5, 1864) observations in each school each year, with the attendance assessed on the same children in year 2 as year 1. Children from schools in the first group began receiving the intervention at the start of the first year. Children from schools in the second group received no intervention during the first year and began receiving the intervention in the second year of the study. Children from schools in the third group did not receive the intervention during these two years (Figure 1).

In the reanalysis of this trial, it was found that the odds ratios (ORs) for school attendance for year 1 and year 2 were both smaller when analysed individually (OR = 1.48 and 1.23 respectively) than the OR given by the standard model when the data were pooled from both years (OR = 1.82) [2]. We hypothesised that this could have been because the analysis model was misspecified and explored two potential types of misspecification:

- (1) The period effects varied between clusters. The standard model assumes that the period effects are common to all clusters. This could lead to a biased estimation of the intervention effect through biased estimation of the period effects.
- (2) The intervention effect varied between clusters. The standard model assumes that the intervention effect is common to all clusters. Treating an effect that truly varies as a fixed effect has been

Group	Cluster	Year 1	Year 2
1	1		
	...		INTERVENTION
	25		
2	26		
	...		
	50		
3	51		
	...	CONTROL	
	75		

Figure 1. Schematic of motivating example: A stepped wedge trial (SWT) with 75 clusters randomised to three groups. The trial consisted of two time periods (years). Group 1 switched to the intervention at the start of period 1. Group 2 switched to the intervention at the start of period 2. Group 3 did not switch to the intervention.

shown to lead to biased estimation of that covariate [6], and so the estimate of the intervention effect could be biased.

In this paper, we first used a simulation study based on the motivating example to explore the effect of ignoring variability between the clusters in the period effect and intervention effect in the analysis of SWTs. Second, we hypothesised that the effect of misspecification would be highly influenced by the weight given to horizontal comparisons in each analysis model and so also performed a further set of simulations to investigate this question. We then analysed the motivating example with different analysis models and compared the results in light of the findings of the simulation studies.

3. Simulation study methods

3.1. Simulation study 1

To investigate the impact of ignoring heterogeneity between clusters in the period effect and intervention effect, we compared analysis models that assumed these effects were common to all clusters (the standard model) to analysis models which allowed these effects to vary between clusters. We performed this with data in which the true underlying period effect and intervention effect were either common to all clusters or varied between clusters. A description of the scenarios we used to compare the analysis models is given, followed by the three analysis models we compared. A summary of the data scenarios simulated is given in Table I.

We used the same trial design as our motivating example with clusters randomised into three groups and followed for two time periods. During the first period, only the first group had received the intervention, and during the second time period, the first and second groups had received the intervention. The third group never received the intervention. This trial design was chosen due to its simplicity; because there are only two time periods, the period effect is simple to model. The horizontal comparison is only possible in one group; this allowed us to explore the weight given to this comparison. To mimic the motivating example and to avoid issues with small sample size, we assigned 25 clusters to each group and the number of observations in a cluster in each time period was drawn from a log-normal distribution ($\mu = 6.9$, $\sigma = 0.74$); this gave a geometric mean number of observations in each cluster in each time period of 1027 (IQR 669, 1798).

The cluster-level distribution of the outcome in the first period and the change from period 1 to period 2 (the period effect) was based on group 3 of the motivating example. This group was chosen because it did not receive the intervention. We modelled the log-odds in the first period and the log-OR period effect from the motivating example as a bivariate normal distribution. This gave mean values for the log-odds in period 1 and log-OR period effect, together with a 2×2 covariance matrix. This distribution described the outcome and how it varied between the clusters in each period. The mean values were used in all the simulation scenarios, but we manipulated the covariance matrix to create four scenarios of how

Table I. Summary of simulation study data scenarios.		
	Description	Similar to motivating example?
Common to all simulations		
Number of groups	3	Yes
Number of time periods	2. In period 1, group 1 received the intervention. In period 2, groups 1 and 2 received the intervention.	Yes
Number of clusters	75	Yes
Cluster size	Log-normal(6.9, 0.74) in each year. Geometric mean = 1027	Yes
Correlation of measurements within clusters	Independent within cluster-periods	No
Mean outcome in year 1	Odds = 6.61	Yes
Mean change in outcome from year 1 to year 2	Odds ratio = 0.32	Yes
Different scenarios		
Period effect	(1) Common period effect, high variability (2) Common period effect, low variability (3) Varying period effect, decreasing variability (4) Varying period effect, stable variability	No No Yes No
Intervention effect	(a) Log(OR) = 0.41 common to all clusters (b) Log(OR) = 0.41, varying between clusters	No No
Intervention effect in group 2		
Simulation study 1	Intervention effect in group 2 the same as group 1 log(OR) = 0.41	No
Simulation study 2	Intervention effect in group 2 is log(OR) = 1.5, and group 1 is log(OR) = 0.41	No

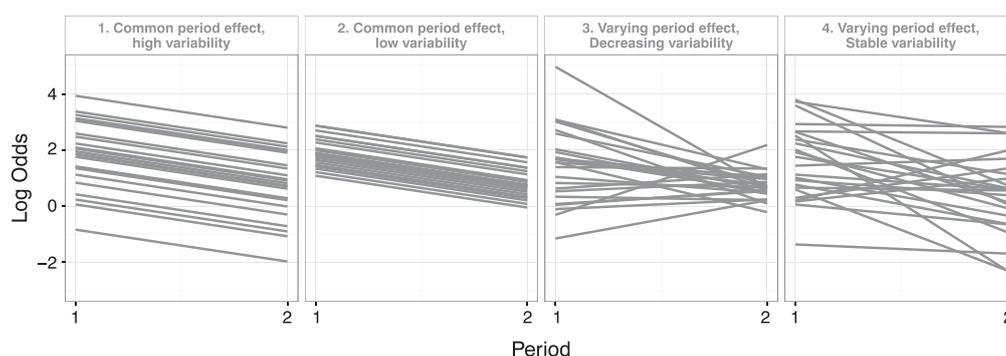


Figure 2. Simulated cluster-level log odds in each period effect scenario. A sample of 25 clusters is shown in time periods 1 and 2. All are in control condition.

the outcome varied between the clusters and periods (Figure 2). The mean odds in the first period was 6.61 (a proportion of 87%), and the mean OR period effect between the second and first period was 0.32, which was equivalent to an odds of 2.12 (proportion of 68%) in the second period. The covariance matrices for each of the four scenarios are given in Data S1 and are described in the succeeding texts:

(1) *Common period effect, high variability:*

The period effect was common to all clusters with between-cluster variance = 1.81. This was the amount of between-cluster variability observed in year 1 of the motivating example. This represents a simple scenario with a large intracluster correlation coefficient (ICC = 0.20), where the standard model would have a correctly specified period effect.

(2) *Common period effect, low variability:*

The period effect was common to all clusters with between-cluster variance = 0.25. This was the amount of between-cluster variability observed in year 2 of the motivating example. This represents a simple scenario with a lower ICC (ICC = 0.05), where, again, the standard model would have a correctly specified period effect.

(3) *Varying period effect, decreasing variability:*

The period effect varied between clusters with the variability between the clusters decreasing from the first period to the second period. The initial between-cluster variance was 1.81, and the period effect variance was 1.89. The decrease in variability from period 1 to period 2 resulted from a negative covariance between the initial value and the period effect of -1.72 . This complex scenario reflects the underlying trends seen in the motivating example. In this scenario, the standard model would have a misspecified period effect.

(4) *Varying period effect, stable variability:*

The period effect varied between the clusters, but the between-cluster variance remained the same for both periods. Here, the initial between-cluster variability and period effect variability remained the same as in scenario (3), but the covariance was reduced to -0.94 . This scenario was chosen to assess the effect of a varying period effect without the additional complication of the between-cluster variation reducing in the second period. In this scenario, the standard model would have a misspecified period effect.

We simulated two scenarios for the intervention effect; these were not based on the motivating example:

(a) An intervention effect that was common to all clusters. We simulated an intervention effect $\log(\text{OR}) = 0.41$ (equivalent to $\text{OR} = 1.5$) for all clusters. We also simulated $\log(\text{OR}) = 0$ to calculate the type I error rate. In these scenarios, the standard model would have a correctly specified intervention effect.

(b) An intervention effect that varied between clusters drawn from the distribution $\log(\text{OR}) \sim N(0.41, 0.3)$. This gave a geometric mean $\text{OR} = 1.5$ with an $\text{IQR} = 1.05\text{--}1.97$. We also simulated a distribution $\log(\text{OR}) \sim N(0, 0.3)$ to calculate the type I error rate. In these scenarios, the standard model would have a misspecified intervention effect.

The variation in the intervention effect was modelled as being independent of the underlying outcome and period effect between-cluster variability. This meant that the intervention effect varying between clusters would lead to increased variability between the clusters in period 2 as more clusters were receiving the intervention in this period.

Each scenario led to the odds of the outcome occurring in each cluster-period. From this, the observations within each cluster-period were sampled from a binomial distribution, assuming independence within each cluster-period. This assumes a cross-sectional design and is a deviation from the motivating example, where children were observed multiple times during the study, chosen for simplicity.

All combinations of these parameters were simulated.

3.2. Simulation study 2

Second, we hypothesised that the horizontal comparisons would depend on the model assumptions more heavily than the vertical comparisons. To aid interpretation of the results of simulation study 1, we sought to investigate the contribution of the horizontal comparisons to each analysis in each scenario.

In the trial design used for this paper, only group 2 contributed horizontal comparisons because groups 1 and 3 remained in the same condition for both periods of the study (Figure 1). This meant that we could investigate the weights given to the horizontal and vertical comparisons by identifying how much weight was given to group 2 relative to groups 1 and 3.

To do this, we reran the simulations but with an intervention effect $\log(\text{OR}) = 1.5$ in group 2 of the trial but kept an intervention effect in group 1 of $\log(\text{OR}) = 0.41$. An unbiased intervention effect estimate from horizontal comparisons alone would have an expectation of $E(\log(\text{OR})) = 1.5$. An unbiased intervention effect estimate from vertical comparisons alone would have an expectation of $0.41 < E(\log(\text{OR})) < 1.5$ depending on the weights given to each cluster and to periods 1 and 2 of the trial. Comparing the intervention effect estimates of each model in each scenario to the horizontal comparison $E(\log(\text{OR})) = 1.5$ allowed us to see how much the horizontal comparisons contributed to the analysis compared with the vertical comparisons. Such a large imbalance in the intervention effect

between groups is, of course, unlikely (although not impossible); this simulation study was designed to investigate the contributions of vertical and horizontal comparisons, rather than to explore a realistic scenario.

4. Analysis models

Each simulated data set was analysed with three analysis models, each making different assumptions about the period effect and intervention effect.

4.1. Standard model

First, we used the standard method of analysis [4,5]: a mixed-effect logistic regression with a random intercept and fixed effects for intervention effect and period effect:

$$y_{ijk} = \mu + \beta Z_j + \theta X_{ij} + u_i \quad (1)$$

where y_{ijk} is the log odds of the outcome in cluster i in year j for observation k , μ is the mean log odds of the outcome in period 1 in the control condition, β is the period effect log-OR comparing the outcome in periods 2 and 1, Z_j is an indicator of year, 0 for the first year and 1 for the second year, θ is the intervention effect log-OR, and X_{ij} is an indicator of whether cluster i received the intervention in year j , $u_i \sim N(0, \sigma_u^2)$ is a random intercept allowing for variability in the outcome between clusters.

This model assumes that the period effect and the intervention effect are common to all clusters so is a misspecified model in scenarios where either the period effect or intervention effect varied between clusters.

4.2. Random period model

Second, we added a random effect for period to the standard model:

$$y_{ijk} = \mu + (\beta + v_i)Z_j + \theta X_{ij} + u_i \quad (2)$$

where $\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim \text{MVN} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{u,v}^2 \\ \sigma_{u,v}^2 & \sigma_v^2 \end{pmatrix} \right)$ are a random intercept and random effect for period respectively.

This model assumes that the intervention effect is common to all clusters but allows the period effect to vary between clusters. It is a misspecified model in scenarios where the intervention effect varies between the clusters.

Sometimes, other literatures have used a different model to allow the period effect to vary between the clusters [11,12]. For details on how these models relate to one another, see Data S2.

4.3. Random intervention model

Third, we added a random effect for the intervention to the standard model:

$$y_{ijk} = \mu + \beta Z_j + (\theta + z_i)X_{ij} + u_i \quad (3)$$

where $\begin{pmatrix} u_i \\ z_i \end{pmatrix} \sim \text{MVN} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{u,z}^2 \\ \sigma_{u,z}^2 & \sigma_z^2 \end{pmatrix} \right)$ are a random intercept and random effect for intervention respectively.

This model assumes that the period effect is common to all clusters but allows the intervention effect to vary between clusters. The model is a misspecified model in scenarios where the period effect varies between the clusters.

Whilst the random period and random intervention models allow for variability in the period and intervention effect respectively, they can estimate a variability of close to zero if the effect is common to all clusters. The random period model is correctly specified in the scenario with common period

effect, and likewise, the random intervention model is correctly specified in the scenario with common intervention effect. Similarly, the random intervention model allows for a covariance between the intervention effect and the intercept ($\sigma_{u,z}^2$) but allows this covariance to be zero, as is the case in our simulation study.

5. Estimands and performance measures

We ran 500 simulations for each combination of parameters. This allowed us to estimate the intervention effect to within 5% accuracy, assuming a variance estimate of 0.05. This variance is conservative as it is larger than the estimated variance we saw in the motivating example.

From the analysis models, we collected the estimated fixed effects, their standard errors, and the estimated between-cluster covariance matrix.

We calculated the mean, standard deviations, 95% confidence intervals (CIs), and the IQR of the intercept, intervention effect, and period effect estimates from the 500 simulations. We calculated percentage bias as

$$\text{percentage bias} = \left(\frac{\bar{\beta} - \beta}{\beta} \right)$$

where β is the true effect and $\bar{\beta}$ is the mean of the effect estimates.

We calculated the coverage of 95% CIs as the proportion of simulations with the true effect contained within the 95% CI of the estimate. We calculated the type 1 error rate as the proportion of simulations with true OR = 1 with $P < 0.05$ against a null of the intervention effect OR = 1.

In the set of simulations with a different intervention effect in group 2 (simulation study 2), we compared the mean of the intervention effect estimates with the horizontal intervention effect of log (OR) = 1.5.

Simulations were run in R version 3.2; the lme4 package was used for mixed-effect models.

6. Results

6.1. Model convergence

The standard model converged in all simulations for both simulation studies. When either the period effect or the intervention effect varied between clusters, the random period and random intervention models also converged in >99% of all simulations. However, when both period effect and intervention effect were common to all clusters, the random period model failed to converge in 3% to 9% of simulations and the random intervention model failed to converge in 4% to 33% of simulations. Estimates from these models were excluded from performance statistics. Further details of convergence of the models are given in Data S3.

6.2. Simulation study 1 results

6.2.1. Bias of fixed-effect estimates. Figure 3 gives the mean and IQR of intervention effect estimates for each scenario. A table of the mean values is given in Data S4.

Where there were common period and intervention effects, all three models performed similarly, with estimation of the intervention effect in line with the true underlying effect.

Where the period effect varied between the clusters, only the random period model gave unbiased estimates of the intervention effect. Depending on the scenario, the standard model had between -20% and -8% bias and the random intervention model between -51% and -8% bias. Bias was larger when the period effect varied with decreasing variability than with stable variability but was similar regardless of whether there was a common or varying intervention effect. We also observed bias in the period effect estimates and intercept estimates from the standard model and random intervention model (Data S5 and S6).

Where the intervention effect varied between the clusters and there was a common period effect, the random intervention model and the random period model gave unbiased estimates of the intervention effect. Only the standard model intervention effect estimates had substantial bias (-9% and -16% bias for common period effects with high and low variabilities respectively).

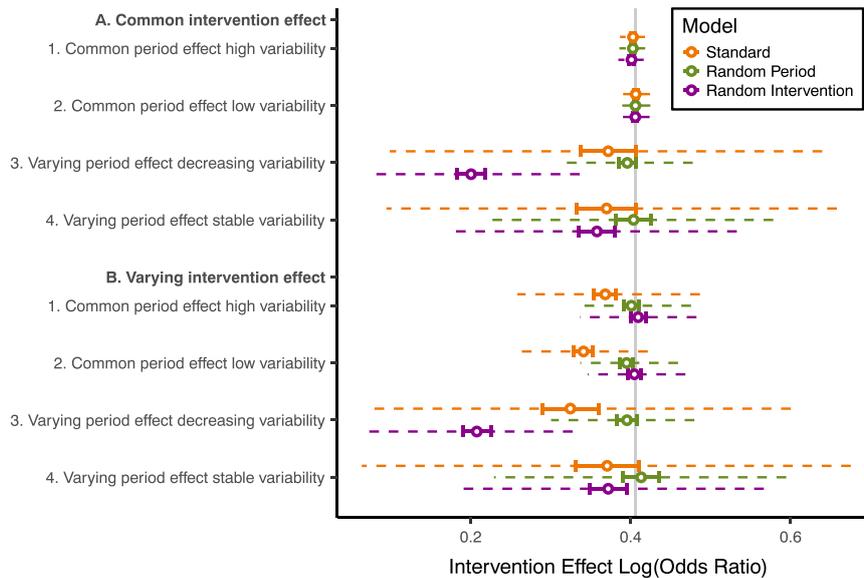


Figure 3. Comparison of intervention effect log(OR) from different analysis models and scenarios with true geometric mean intervention effect log(OR) = 0.41 in all groups. Vertical grey line: true log(OR). Hollow point: mean estimate. Solid barred line: 95% confidence interval. Dashed line: interquartile range (IQR) of estimates.

Where either the period effect or intervention effect varied between clusters, the standard model intervention effect estimates had greater variability compared with the random period model or random intervention model. Differences were larger when the period effect varied between clusters than when the intervention effect varied between clusters. For example, the standard model intervention effect estimates were 3.6 times as variable as the random period model estimates when the period effect varied between clusters with decreasing variability with common intervention effect, whereas the standard model intervention effect estimates were 1.5 times as variable as the random intervention model estimates when the intervention effect varied between clusters with common period effect with high variability.

6.2.2. Standard errors, coverage, and type 1 error. In scenarios with a common period and intervention effect, 95% coverage was maintained regardless of the analysis model and the estimated standard errors were similar across analysis models (Figure 4 and Data S7 and S8).

When period effect or intervention effect varied between clusters, the standard model gave standard errors that were markedly smaller than the random period model and random intervention model. The mean intervention effect standard error from the standard model was less than 0.33 and 0.26 times the mean standard error of the random period model and random intervention model respectively.

The inappropriately small standard errors given by the standard model were in part explained by downward bias in the estimation of between-cluster variability (Data S9). For example, when variability was stable over the two time periods with a variance of 1.79, the standard model estimated the variance as 1.26.

The bias in estimates, standard errors, and increased variability in estimates led to undercoverage of the 95% CIs of the intervention effect estimates (Figure 4). For the standard model, undercoverage was severe when either the intervention effect or the period effect varied between clusters (<25% coverage). Similarly, the random intervention model had undercoverage when the period effect varied between clusters (74% and 88% coverage for decreasing and stable variability respectively) regardless of intervention effect variability. Finally, the random period model had undercoverage of CIs when the intervention effect varied between clusters with a common period effect (86% and 88% coverage for common period effect with high and low variabilities respectively).

Type 1 error rates followed the same patterns as coverage (Data S10).

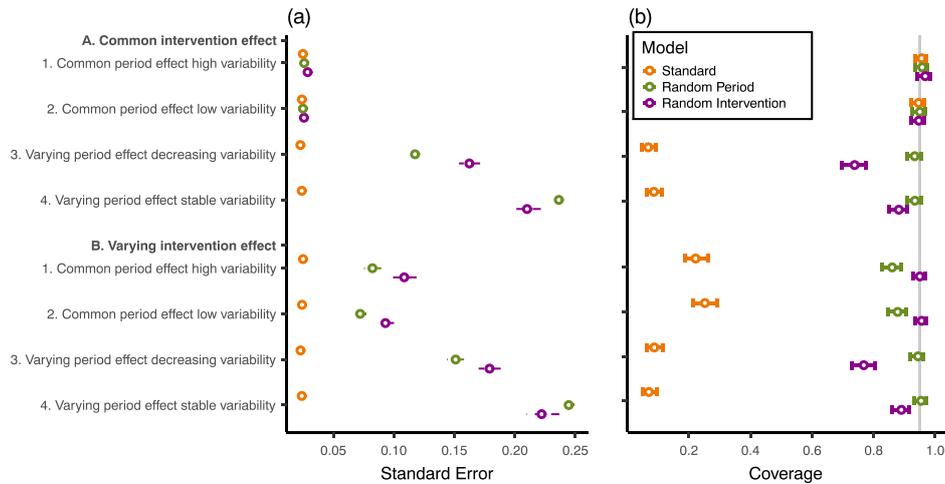


Figure 4. Comparison of estimated intervention effect (a) standard errors and (b) 95% confidence interval coverage for different analysis models and scenarios with a geometric mean intervention effect of log (OR) = 0.41 in all groups. Vertical grey line: 95% coverage. Hollow point: mean estimate. Solid barred line: 95% confidence interval. Dashed line: interquartile range (IQR) of estimates.

6.3. Simulation study 2 results

Figure 5 gives the estimated log(OR) for each scenario where the group 1 and 2 intervention effects differed (log(OR)=0.41 in group 1 and log(OR)=1.5 in group 2).

All analysis models gave a mean estimated intervention effect close to the group 2 effect when there was a common period effect and a common intervention effect; this was the case in the high and low variability scenarios. This suggests that, in these scenarios, the intervention effect is largely estimated from horizontal within-cluster comparisons in group 2; groups 1 and 3 appeared to contribute to estimation of the period effect but had little influence on the intervention effect estimate.

The standard model estimates remained close to the group 2 intervention effect in all scenarios. The downward bias we observed in our first set of simulations suggests that at least some of the movement away from the group 2 effect is because of bias and not because of a reduction in the contribution of the

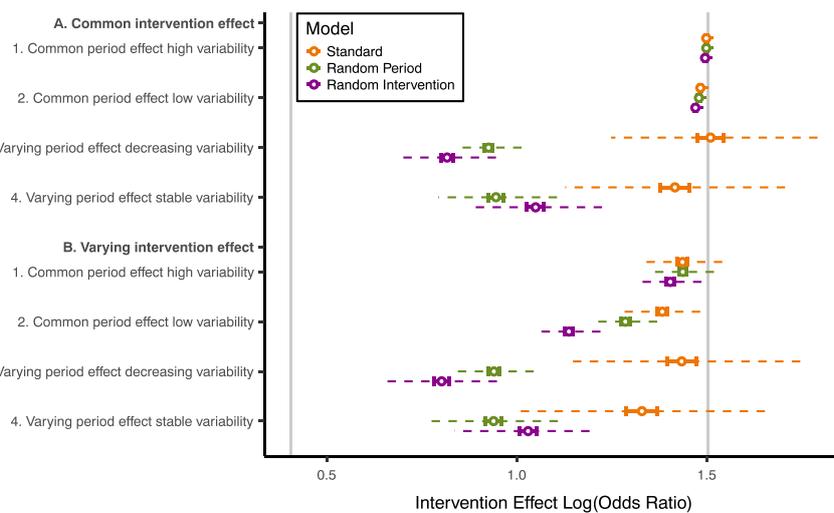


Figure 5. Comparison of intervention effect log odds ratios from different analysis models for all scenarios with the intervention effect larger in group 2 than group 1. Vertical grey lines: true intervention effect in group 1 (log(OR) = 0.41) and group 2 (log(OR) = 1.5). Hollow point: mean estimate. Solid barred line: 95% confidence interval. Dashed line: interquartile range (IQR) of estimates.

horizontal comparisons. This implies that the standard model was continuing to estimate the intervention effect largely from horizontal comparisons in group 2.

In contrast, when the period effect varied between clusters, the random period model gave intervention effect estimates much further from the horizontal comparison estimates. This implies that the horizontal comparisons in group 2 could not contribute as much information to the analysis because there was less certainty about separating the period effect and intervention effect in these comparisons. This was similar in the scenarios where the intervention effect varied between clusters but the period effect was common for both the random period and random intervention models, but to a smaller degree.

7. Example

For our motivating example, we hypothesised that the standard model gave a larger intervention effect than either of the two time periods analysed separately because the standard model was misspecified by ignoring variability in either the period effect or the intervention effect. Our simulation study suggests that this is not the case because we would expect the standard model to underestimate the intervention effect with these types of misspecification, rather than overestimate the effect. However, we also found that the standard model gave a very large weight to the horizontal comparisons. This does help to explain the counterintuitive results seen in the motivating example [2].

We reanalysed the deworming trial by using the three analysis models investigated in the simulation study and additionally looked at years 1 and 2 separately by using a mixed-effect model with a fixed effect for intervention and a random intercept to attain estimates for the intervention effect from vertical comparisons. In line with the published reanalysis of this study, we ignored pupil-level clusters from multiple observations of the same pupils; this is in line with research suggesting that it is sufficient to adjust for the highest level of clustering alone, known as passing the buck [13].

The results in Table II are different from the published reanalysis as we have used a different version of the data (see Data S11 for details) and have not adjusted for covariates other than period [14].

We found that the standard model combining data from both years of the study gave a larger estimate of the intervention effect than either year analysed separately, which is as was found in the reanalysis [4].

Adjusting for variation between clusters in the period effect or intervention effect (i.e. using either the random period model or random intervention model) increased the intervention effect standard error and reduced the intervention effect towards the null. Both approaches gave an intervention effect estimate between the estimated effect in year 1 and year 2. This suggests that the horizontal comparisons are contributing less to these analysis models than to the standard model; this is consistent with the findings of our second simulation study into the contribution of the horizontal comparisons.

The random period model found strong evidence of variability in the period effect ($p < 0.001$), and the random intervention model found strong evidence of variability in the intervention effect ($p < 0.001$). Because the period effect and intervention effect are confounded with one another, evidence of variability in the intervention effect could be caused by variability in the period effect or vice versa. The random period model estimated a between-cluster covariance matrix similar to the simulation study scenario with varying period effect with decreasing variability. The random intervention model

Model	Odds ratio (95% CI)	Standard error	<i>P</i> -value	<i>P</i> -value of random period or intervention effect
Separate year analysis (vertical comparisons)				
Year 1	1.67 (0.90,3.10)	0.32	0.11	
Year 2	1.19 (0.95, 1.50)	0.12	0.13	
Combined analysis				
Standard model	1.74 (1.67, 1.81)	0.02	<0.001	
Random period model	1.26 (1.02, 1.57)	0.11	0.03	<0.001
Random intervention model	1.25 (0.96, 1.62)	0.13	0.09	<0.001

estimated lower variability between clusters in the intervention condition than in the control condition because of the reduced variability in year 2. This is a scenario that we did not consider in our simulation study where we only investigated a scenario with greater variability in the intervention condition. Inspection of the data suggests that the random period model is the most appropriate one. A mixed-effect model with a random effect for period run on observations from group 3, which never received the intervention, finds strong evidence of variability in the period effect ($p < 0.001$). But a mixed-effect model with a random effect for intervention run on observations from groups 1 and 3, where the intervention effect is not confounded with the period effect, finds no evidence of variability in the intervention effect ($p = 0.34$).

The random period model suggests that there is some evidence that the deworming intervention increased school attendance (OR = 1.26, 95% CI 1.02, 1.57; $p = 0.03$). The effect found by using this model is weaker, both in terms of absolute size and level of statistical significance, than the effect found by using the standard model. There are still limitations in these data and this analysis, on which further information has been published elsewhere [1–3].

8. Discussion

We found biased estimates and serious undercoverage of CIs in the SWT scenarios we simulated when the analysis model ignored variability between clusters in the period effect or intervention effect. In these scenarios, results from the standard model were driven largely by the horizontal comparisons.

We have shown that, in the scenarios we considered, misspecifying the random effects of mixed-effect models can result in biased intervention effect estimates. The standard model underestimated the intervention effect when either the period or the intervention effect varied between the clusters. The underestimation when the period effect varied may result from the standard model estimating an intervention effect averaged over the two periods, whereas the true effect for this scenario was a within-period intervention effect. This is analogous to the difference between the population-averaged effect and the cluster-specific effects that are given by different analysis methods. In the presence of intervention effect variability, the standard model also gave biased estimates of the intervention effect. The random intervention model had even larger bias when it was misspecified than the standard model. Conversely, the random period model had only negligible bias in estimates in all scenarios we considered. These results are consistent with previous research into misspecifying mixed-effect models in cluster randomised trials [7,9]. We have built on this literature and shown that these results extend to SWTs. This highlights how sensitive mixed-effect models can be to misspecification of model assumptions.

Caution is needed beyond estimation of the intervention effect itself. In our simulation study, the bias extended to standard errors and between-cluster variability. The latter has implications for reporting the ICC, as recommended by the Consolidated Standards of Reporting Trials guidelines [15]. In addition to the implications for inference, the bias in standard errors has implications for determining the power and sample size of SWTs. Because the standard error from the standard model is used in most current methods of SWT sample size calculations [12,16–18], they should not be applied when the period effect or intervention effect is expected to vary between clusters, at least in relation to the characteristics of the trial exemplar used in this paper. Instead, the method developed by Hooper *et al.* may be more appropriate [11].

The result of these biases was undercoverage of CIs for the intervention effect. If model assumptions do not hold, we risk being overconfident in our conclusions. We found particularly severe undercoverage when using the standard model. This has been seen in previous research into misspecified random effects [6,19] and has recently been seen in the setting of SWTs [20]. This is reflected in our analysis of the motivating example; we see a large increase in the standard error of the intervention effect, and so CIs are much wider when moving from the standard model to the random period model or random intervention model.

The results from our simulation study could be explained by the excessive weight given to the horizontal comparisons, even with a lower ICC = 0.05. Because the horizontal comparisons are within-cluster comparisons, they avoid the additional variability of between-cluster variation. This means that if the period and intervention effects can be separated, the horizontal comparisons will be given more weight than the vertical comparisons by all the analysis models we considered. However, by making the stringent assumption that period and intervention effects are the same in every cluster, the standard model assumes too much certainty in separating the period and intervention effects. The

reason that the standard model performed poorly in the simulation study was because of its reliance on the horizontal comparisons.

In the design we studied, the weight given to horizontal comparisons also meant that greater weight was given to some groups of clusters than others. The implications of this are not well understood. When there is a large difference in the weight given to each group, the intervention effect estimate no longer represents an average effect across the clusters and interpretation becomes more difficult. Further research is needed to explore this issue in more traditional SWT designs with more groups and when all clusters have observations in the control and intervention conditions, and so all clusters contribute through horizontal comparisons.

A criticism of the random intervention model and, to a lesser extent, the random period model is that they sometimes had problems with convergence. This occurred almost exclusively when both the period effect and the intervention effect were common to all clusters; the non-convergence resulted from the models attempting to estimate a true variance of zero, the boundary of the parameter. In this scenario, all the analysis models gave unbiased effect estimates and appropriate CI coverage. We would suggest that an analysis plan gives an alternative, simpler model to use in case of convergence issues due to lack of variability. In our simulation study, this procedure gave good coverage and no bias in the scenarios with common period effect and intervention effect, where convergence was an issue (data not shown).

Given that the mixed-effect model can be so sensitive to model assumptions, other analysis methods should be considered. This choice should be prespecified and prior knowledge used to justify the assumptions made by the chosen analysis method. We found the random period model to be the most robust of the models considered, but there was still undercoverage of CIs in some scenarios. Some have suggested using permutation tests on the standard model [20]. Although this will give correct inference, there is still a risk of biased intervention effect estimation. Alternative analysis methods that make fewer assumptions may be more appropriate. Generalised estimating equations have been suggested for the analysis of SWTs [21] and have been shown to be more robust to misspecification of the correlations in the data in other settings [22], but this robustness has yet to be assessed in the context of SWTs. Analysis methods that only make use of the vertical comparisons are desirable as they require no assumptions about period effects, but there are no such methods currently published, and these analyses are less efficient [23]. Sensitivity analysis could also be used to assess the robustness of results.

We have only considered a limited range of designs in this simulation study. We used a very simple SWT design to make the analyses as transparent as possible; this design only had two steps, and not all clusters received the intervention in the course of the study. Further research is needed to confirm that our findings hold for other SWT designs. In more traditional SWTs, all clusters receive both the control and intervention conditions, and so all clusters contribute horizontal comparisons. Because the problems we highlight arise from the horizontal comparisons, this might exacerbate the problems we identified. We have only considered two values for the ICC when the period effect was common to all clusters and have not assessed the effect of ICC when period effects vary between clusters. In scenarios where these effects varied between clusters, the baseline ICC was 0.20, which, in many contexts, would be considered large. Additionally, there was large variability in the period effect; the effect of a less variable period effect needs further exploration. It is not known how common it is for the period and intervention effects to vary between clusters in practice; however, we have based this simulation on real trial data. Large clusters were used in the simulation study to reflect the motivating deworming trial; however, similar results were seen with a smaller mean cluster size of 250 (data not shown). We used a large number of clusters in each group to avoid small sample issues.

Whilst further research is needed to explore the potential for bias in a wider range of designs and settings, we have demonstrated that there is a potential for the standard model to give biased intervention effect estimates and undercoverage of CIs. These simulations provide clear evidence that the standard model for analysis of SWTs can be both highly sensitive to the data meeting the model assumptions and highly dependent on non-randomised horizontal comparisons. We urge those conducting SWTs to ensure an appropriate analysis is used.

References

1. Miguel E, Kremer M. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 2004; **72**:159–217.
2. Davey C, Aiken AM, Hayes RJ, Hargreaves JR. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. *International Journal of Epidemiology* 2015; **44**:1581–1592.

3. Aiken AM, Davey C, Hargreaves JR, Hayes RJ. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. *International Journal of Epidemiology* 2015; **44**:1572–1580.
4. Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ, Fielding KL. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials* 2015; **16**:358.
5. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials* 2007; **28**:182–191.
6. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* 2001; **88**:973–985.
7. Ukoumunne OC, Thompson SG. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Statistics in Medicine* 2001; **20**:417–433.
8. Koepsell TD, Martin DC, Diehr PH, Psaty BM, Wagner EH, Perrin EB, Cheadle A. Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis of variance approach. *Journal of Clinical Epidemiology* 1991; **44**:701–713.
9. Turner RM, White IR, Croudace T, Group PIPS. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Statistics in Medicine* 2007; **26**:274–289.
10. Morgan KE, Forbes AB, Keogh RH, Jairath V, Kahan BC. Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome. *Statistics in Medicine* 2017; **36**:318–333.
11. 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):4718–4728.
12. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Statistics in Medicine* 2016; **35**:2149.
13. Bottomley C, Kirby MJ, Lindsay SW, Alexander N. Can the buck always be passed to the highest level of clustering? *BMC Medical Research Methodology* 2016; **16**:29.
14. Miguel E, Kremer M. Replication data for: worms: identifying impacts on education and health in the presence of treatment externalities. *Harvard Dataverse* 2014, Version:V2.1. <https://doi.org/10.7910/DVN/28038>.
15. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012; **345**:e5661.
16. 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**:752–758.
17. Hemming K, Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *Journal of Clinical Epidemiology* 2016; **69**:137–146.
18. Hemming K, Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata Journal* 2014; **14**:363–380.
19. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. *Statistical Science* 2000; **15**:1–19.
20. Ji X, Fink G, Robyn PJ, Small SS. Randomization inference for stepped-wedge cluster-randomised trials: an application to community-based health insurance. *Annals of Applied Statistics* 2017; **11**(1):1–20.
21. Scott JM, deCamp A, Juraska M, Fay MP, Gilbert PB. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. *Statistical Methods in Medical Research* 2017; **26**(2):583–597.
22. Liang KY, Zeger SL. Longitudinal data-analysis using generalized linear-models. *Biometrika* 1986; **73**:13–22.
23. Granston T. Addressing lagged effects and interval censoring in the stepped wedge design of cluster randomized clinical trials. University of Washington 2014; Doctor of Philosophy.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article.

Supporting Information

In the tables S3-S10, the geometric mean of the intervention effect log odds ratio is given where the intervention effect varies between clusters. All odds and odds ratios are reported on the log scale.

S1: Covariance Matrices

Below are the covariance matrices used in the data generating process for each scenario. These are given in the format:

$$\begin{pmatrix} \text{var}(\textit{intercept}) & \text{cov}(\textit{period}, \textit{intercept}) \\ \text{cov}(\textit{period}, \textit{intercept}) & \text{var}(\textit{period}) \end{pmatrix}$$

1. Common period effect, high variability

$$\begin{pmatrix} 1.808 & 0 \\ 0 & 0 \end{pmatrix}$$

2. Common period effect, low variability

$$\begin{pmatrix} 0.251 & 0 \\ 0 & 0 \end{pmatrix}$$

3. Varying period effect, decreasing variability

$$\begin{pmatrix} 1.808 & -1.721 \\ -1.721 & 1.885 \end{pmatrix}$$

4. Varying period effect, stable variability

$$\begin{pmatrix} 1.808 & -0.943 \\ -0.943 & 1.885 \end{pmatrix}$$

S2: Comparison of the Random-Period Model and Cluster-Period Interaction Model

Parameterisation A

In this paper, we used the following model to allow the period effect to vary between clusters:

$$y_{ijk} = \mu + (\beta + v_i)t_j + \theta X_{ij} + u_i$$

where y_{ijk} is the log odds of the outcome in cluster i in year j for observation k , μ is the mean log odds of the outcome in period one in the control condition, β is the period effect log odds ratio comparing the outcome in periods two and one, t_j is an indicator of year; 0 for the first year and 1 for the second year, θ is the intervention effect log odds ratio, and X_{ij} is an indicator of whether cluster i received the intervention in year j , and:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{u,v}^2 \\ \sigma_{u,v}^2 & \sigma_v^2 \end{pmatrix} \right)$$

are a random intercept and random effect for period respectively.

Parameterisation B

Other literature has sometimes used an alternative parameterisation:

$$y_{ijk} = \mu + \beta t_j + \theta X_{ij} + v_{ij}^* + u_i^*$$

where now $u_i^* \sim N(0, \sigma_{u^*}^2)$ and $v_{ij}^* \sim N(0, \sigma_{v^*}^2)$.

The parameterisation used in this paper, parameterisation A, is more flexible than the parameterisation sometimes used elsewhere, parameterisation B. Parameterisation A allows the total variability to change between periods. This is necessary to correctly model our motivating example. Parameterisation B assumes that the total variability is the same in each period.

We can add the restraint that the variance is the same in each period to parameterisation A by setting $\sigma_{u,v}^2 = -\frac{1}{2}\sigma_v^2$. When we do this, parameterisations A and B are equivalent and

$$\begin{aligned} \sigma_{u^*}^2 &= \sigma_u^2 - \frac{1}{2}\sigma_v^2 \\ \sigma_{v^*}^2 &= \frac{1}{2}\sigma_v^2 \end{aligned}$$

S3: Table of Convergence of Analysis Models by Simulation Parameters

Simulation study one: Group two intervention effect log odds ratio the same as group one

Group one intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	100	95	67
0.41	Common	Common period effect, low variability	100	96	94
0.41	Common	Varying period effect, Decreasing variability	100	99	100
0.41	Common	Varying period effect, Stable variability	100	100	100
0.41	Varying	Common period effect, high variability	100	100	100
0.41	Varying	Common period effect, low variability	100	100	100
0.41	Varying	Varying period effect, Decreasing variability	100	100	100
0.41	Varying	Varying period effect, Stable variability	100	100	100
0	Common	Common period effect, high variability	100	95	73
0	Common	Common period effect, low variability	100	97	96
0	Common	Varying period effect, Decreasing variability	100	100	100
0	Common	Varying period effect, Stable variability	100	100	100
0	Varying	Common period effect, high variability	100	100	100
0	Varying	Common period effect, low variability	100	100	100
0	Varying	Varying period effect, Decreasing variability	100	100	100
0	Varying	Varying period effect, Stable variability	100	100	100

Simulation study two: Group two intervention effect log odds ratio different to group one

Group one Intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	100	91	73
0.41	Common	Common period effect, low variability	100	99	94
0.41	Common	Varying period effect, Decreasing variability	100	100	100
0.41	Common	Varying period effect, Stable variability	100	100	100
0.41	Varying	Common period effect, high variability	100	100	100
0.41	Varying	Common period effect, low variability	100	100	100
0.41	Varying	Varying period effect, Decreasing variability	100	100	100
0.41	Varying	Varying period effect, Stable variability	100	100	100

S4: Table of Mean of Intervention Effect Log Odds Ratio Estimates by Simulation Parameters

Simulation study one: Group two intervention effect log odds ratio the same as group one

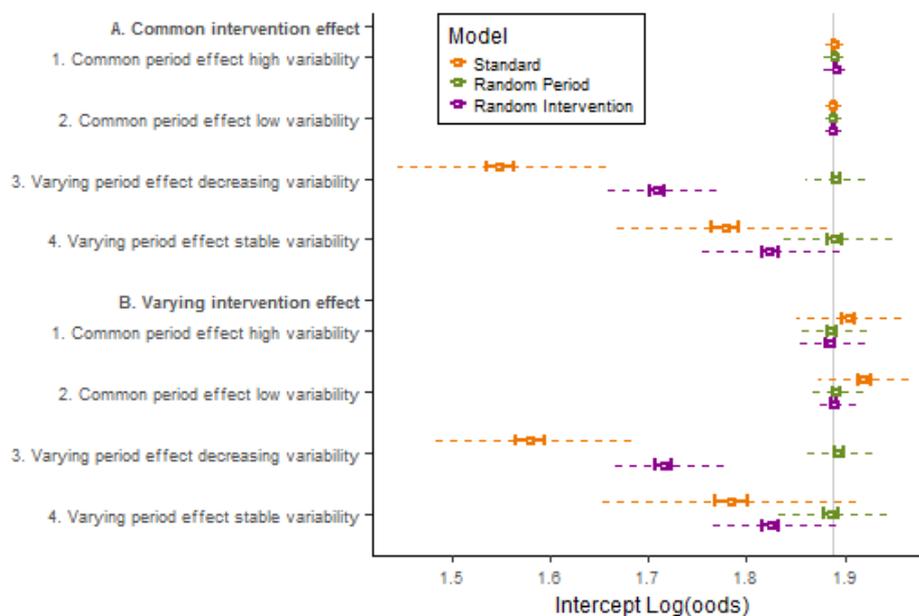
Group one intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	0.4	0.4	0.4
0.41	Common	Common period effect, low variability	0.41	0.41	0.41
0.41	Common	Varying period effect, Decreasing variability	0.37	0.4	0.2
0.41	Common	Varying period effect, Stable variability	0.37	0.4	0.36
0.41	Varying	Common period effect, high variability	0.37	0.4	0.41
0.41	Varying	Common period effect, low variability	0.34	0.39	0.4
0.41	Varying	Varying period effect, Decreasing variability	0.32	0.4	0.21
0.41	Varying	Varying period effect, Stable variability	0.37	0.41	0.37
0	Common	Common period effect, high variability	0	0	0
0	Common	Common period effect, low variability	0	0	0
0	Common	Varying period effect, Decreasing variability	-0.01	-0.01	-0.18
0	Common	Varying period effect, Stable variability	-0.01	-0.01	-0.05
0	Varying	Common period effect, high variability	-0.03	0	0
0	Varying	Common period effect, low variability	-0.07	-0.01	-0.01
0	Varying	Varying period effect, Decreasing variability	-0.07	0	-0.18
0	Varying	Varying period effect, Stable variability	-0.06	-0.02	-0.07

Simulation study two: Group two intervention effect log odds ratio different to group one

Group one Intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	1.5	1.5	1.5
0.41	Common	Common period effect, low variability	1.48	1.48	1.47
0.41	Common	Varying period effect, Decreasing variability	1.51	0.93	0.82
0.41	Common	Varying period effect, Stable variability	1.42	0.95	1.05
0.41	Varying	Common period effect, high variability	1.44	1.44	1.4
0.41	Varying	Common period effect, low variability	1.38	1.29	1.14
0.41	Varying	Varying period effect, Decreasing variability	1.43	0.94	0.8
0.41	Varying	Varying period effect, Stable variability	1.33	0.94	1.03

S5a: Figure of Mean and Spread of Intercept Log Odds Estimates by Simulation Parameters

Hollow point: Mean estimate, solid barred line: 95% confidence interval, dashed line: IQR of estimates.



S5b: Table of Mean of Intercept Log Odds Estimates by Simulation Parameters

True intercept is $\log(6.62)=1.89$.

Simulation study one: Group two intervention effect log odds ratio the same as group one

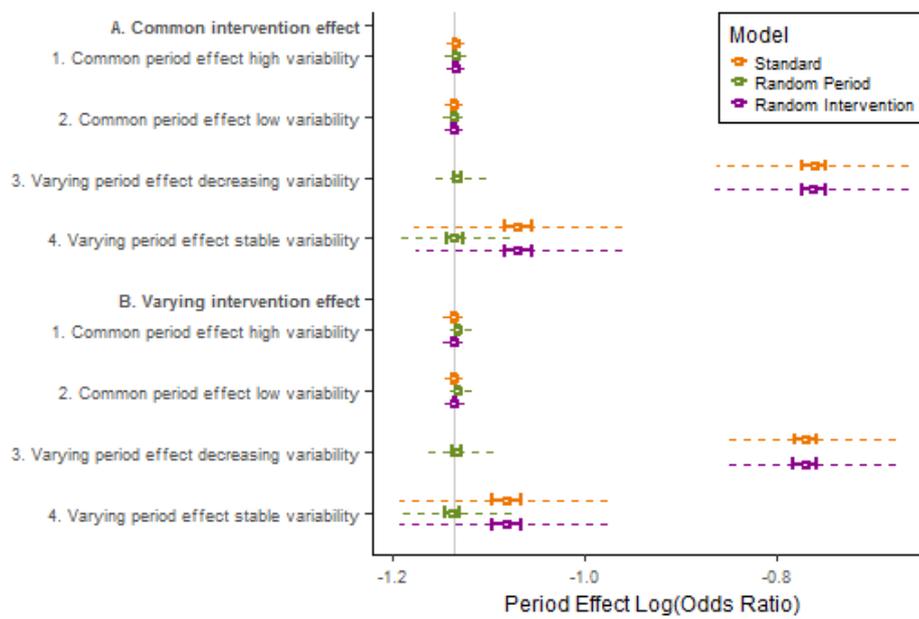
Group one intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	1.89	1.89	1.89
0.41	Common	Common period effect, low variability	1.89	1.89	1.89
0.41	Common	Varying period effect, Decreasing variability	1.55	1.89	1.71
0.41	Common	Varying period effect, Stable variability	1.78	1.89	1.82
0.41	Varying	Common period effect, high variability	1.9	1.89	1.89
0.41	Varying	Common period effect, low variability	1.92	1.89	1.89
0.41	Varying	Varying period effect, Decreasing variability	1.58	1.89	1.72
0.41	Varying	Varying period effect, Stable variability	1.78	1.89	1.82
0	Common	Common period effect, high variability	1.89	1.89	1.89
0	Common	Common period effect, low variability	1.89	1.89	1.89
0	Common	Varying period effect, Decreasing variability	1.55	1.89	1.72
0	Common	Varying period effect, Stable variability	1.77	1.89	1.83
0	Varying	Common period effect, high variability	1.9	1.89	1.89
0	Varying	Common period effect, low variability	1.92	1.89	1.89
0	Varying	Varying period effect, Decreasing variability	1.59	1.89	1.72
0	Varying	Varying period effect, Stable variability	1.8	1.89	1.83

Simulation study two: Group two intervention effect log odds ratio different to group one

Group one Intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	1.52	1.53	1.53
0.41	Common	Common period effect, low variability	1.53	1.53	1.54
0.41	Common	Varying period effect, Decreasing variability	1.19	1.72	1.64
0.41	Common	Varying period effect, Stable variability	1.43	1.71	1.71
0.41	Varying	Common period effect, high variability	1.55	1.54	1.59
0.41	Varying	Common period effect, low variability	1.57	1.6	1.77
0.41	Varying	Varying period effect, Decreasing variability	1.22	1.71	1.66
0.41	Varying	Varying period effect, Stable variability	1.46	1.71	1.72

S6a: Figure of Mean and Spread of Period Effect Log Odds Ratio Estimates by Simulation Parameters

Hollow point: Mean estimate, solid barred line: 95% confidence interval, dashed line: IQR of estimates.



S6b: Table of Mean of Period Effect Log Odds Ratio Estimates by Simulation Parameters

True period odds ratio is $\log(0.32) = -1.14$.

Simulation study one: Group two intervention effect log odds ratio the same as group one

Group one intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	-1.13	-1.13	-1.13
0.41	Common	Common period effect, low variability	-1.14	-1.14	-1.14
0.41	Common	Varying period effect, Decreasing variability	-0.76	-1.13	-0.76
0.41	Common	Varying period effect, Stable variability	-1.07	-1.13	-1.07
0.41	Varying	Common period effect, high variability	-1.14	-1.13	-1.14
0.41	Varying	Common period effect, low variability	-1.14	-1.13	-1.14
0.41	Varying	Varying period effect, Decreasing variability	-0.77	-1.13	-0.77
0.41	Varying	Varying period effect, Stable variability	-1.08	-1.14	-1.08
0	Common	Common period effect, high variability	-1.14	-1.14	-1.14
0	Common	Common period effect, low variability	-1.13	-1.13	-1.13
0	Common	Varying period effect, Decreasing variability	-0.79	-1.13	-0.79
0	Common	Varying period effect, Stable variability	-1.06	-1.13	-1.06
0	Varying	Common period effect, high variability	-1.13	-1.13	-1.13
0	Varying	Common period effect, low variability	-1.14	-1.13	-1.14
0	Varying	Varying period effect, Decreasing variability	-0.79	-1.14	-0.79
0	Varying	Varying period effect, Stable variability	-1.05	-1.13	-1.05

Simulation study two: Group two intervention effect log odds ratio different to group one

Group one Intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	-1.13	-1.13	-1.13
0.41	Common	Common period effect, low variability	-1.13	-1.13	-1.13
0.41	Common	Varying period effect, Decreasing variability	-0.76	-0.94	-0.77
0.41	Common	Varying period effect, Stable variability	-1.06	-0.95	-1.07
0.41	Varying	Common period effect, high variability	-1.13	-1.12	-1.13
0.41	Varying	Common period effect, low variability	-1.13	-1.07	-1.13
0.41	Varying	Varying period effect, Decreasing variability	-0.76	-0.95	-0.76
0.41	Varying	Varying period effect, Stable variability	-1.07	-0.95	-1.07

S7: Table of Mean of Standard Error Estimates by Simulation Parameters

Simulation study one: Group two intervention effect log odds ratio the same as group one

Group one intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	0.02	0.03	0.03
0.41	Common	Common period effect, low variability	0.02	0.02	0.03
0.41	Common	Varying period effect, Decreasing variability	0.02	0.12	0.16
0.41	Common	Varying period effect, Stable variability	0.02	0.24	0.21
0.41	Varying	Common period effect, high variability	0.02	0.08	0.11
0.41	Varying	Common period effect, low variability	0.02	0.07	0.09
0.41	Varying	Varying period effect, Decreasing variability	0.02	0.15	0.18
0.41	Varying	Varying period effect, Stable variability	0.02	0.24	0.22
0	Common	Common period effect, high variability	0.02	0.03	0.03
0	Common	Common period effect, low variability	0.02	0.02	0.02
0	Common	Varying period effect, Decreasing variability	0.02	0.12	0.16
0	Common	Varying period effect, Stable variability	0.02	0.24	0.21
0	Varying	Common period effect, high variability	0.02	0.08	0.11
0	Varying	Common period effect, low variability	0.02	0.07	0.09
0	Varying	Varying period effect, Decreasing variability	0.02	0.15	0.18
0	Varying	Varying period effect, Stable variability	0.02	0.24	0.22

Simulation study two: Group two intervention effect log odds ratio different to group one

Group one Intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	0.03	0.03	0.03
0.41	Common	Common period effect, low variability	0.03	0.03	0.03
0.41	Common	Varying period effect, Decreasing variability	0.03	0.16	0.19
0.41	Common	Varying period effect, Stable variability	0.03	0.26	0.23
0.41	Varying	Common period effect, high variability	0.03	0.08	0.11
0.41	Varying	Common period effect, low variability	0.03	0.09	0.13
0.41	Varying	Varying period effect, Decreasing variability	0.03	0.19	0.2
0.41	Varying	Varying period effect, Stable variability	0.03	0.27	0.24

S8: Table of Coverage of 95% Confidence Intervals by Simulation Parameters

Simulation study one: Group two intervention effect log odds ratio the same as group one

Group one intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	96	96	97
0.41	Common	Common period effect, low variability	95	95	95
0.41	Common	Varying period effect, Decreasing variability	7	93	74
0.41	Common	Varying period effect, Stable variability	9	93	88
0.41	Varying	Common period effect, high variability	22	86	95
0.41	Varying	Common period effect, low variability	25	88	96
0.41	Varying	Varying period effect, Decreasing variability	9	94	77
0.41	Varying	Varying period effect, Stable variability	7	95	89
0	Common	Common period effect, high variability	96	97	96
0	Common	Common period effect, low variability	95	95	95
0	Common	Varying period effect, Decreasing variability	9	96	80
0	Common	Varying period effect, Stable variability	8	95	88
0	Varying	Common period effect, high variability	22	85	95
0	Varying	Common period effect, low variability	26	83	93
0	Varying	Varying period effect, Decreasing variability	10	95	82
0	Varying	Varying period effect, Stable variability	9	95	88

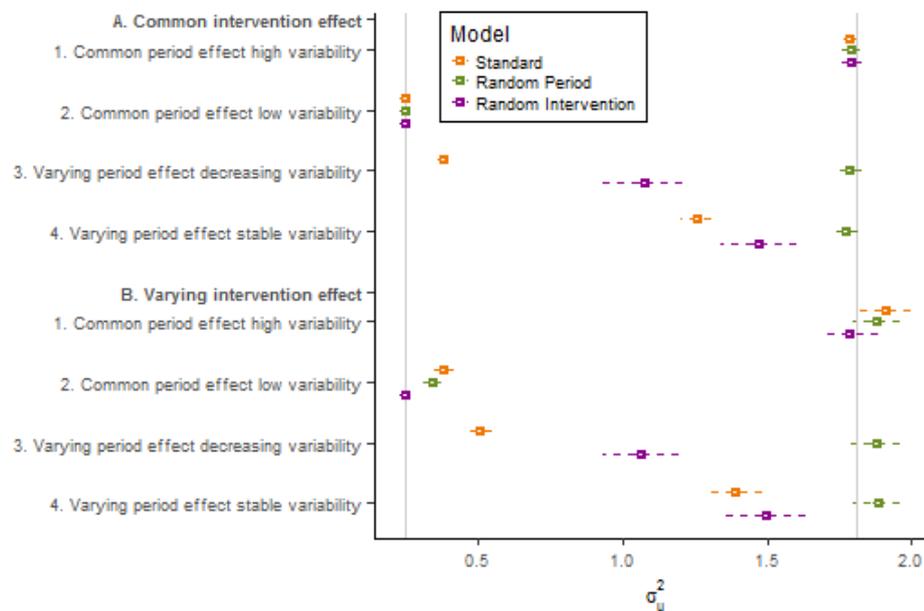
Simulation study two: Group two intervention effect log odds ratio different to group one

This is the percentage of simulations where the 95% confidence interval contained the group 1 true effect

Group one Intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	0	0	0
0.41	Common	Common period effect, low variability	0	0	0
0.41	Common	Varying period effect, Decreasing variability	0	7	41
0.41	Common	Varying period effect, Stable variability	0	43	23
0.41	Varying	Common period effect, high variability	0	0	0
0.41	Varying	Common period effect, low variability	0	0	0
0.41	Varying	Varying period effect, Decreasing variability	1	17	46
0.41	Varying	Varying period effect, Stable variability	0	47	30

S9a: Figure of Mean and Spread of Intercept Between-Cluster Variance Estimates by Simulation Parameters

Hollow point: Mean estimate, dashed line: IQR of estimates.



S9b: Table of Mean of Intercept Between-Cluster Variance Estimates by Simulation Parameters

True variance is $Var(u_i) = 0.25$ for the scenario with common period effect and low variability, and $Var(u_i) = 1.79$ otherwise.

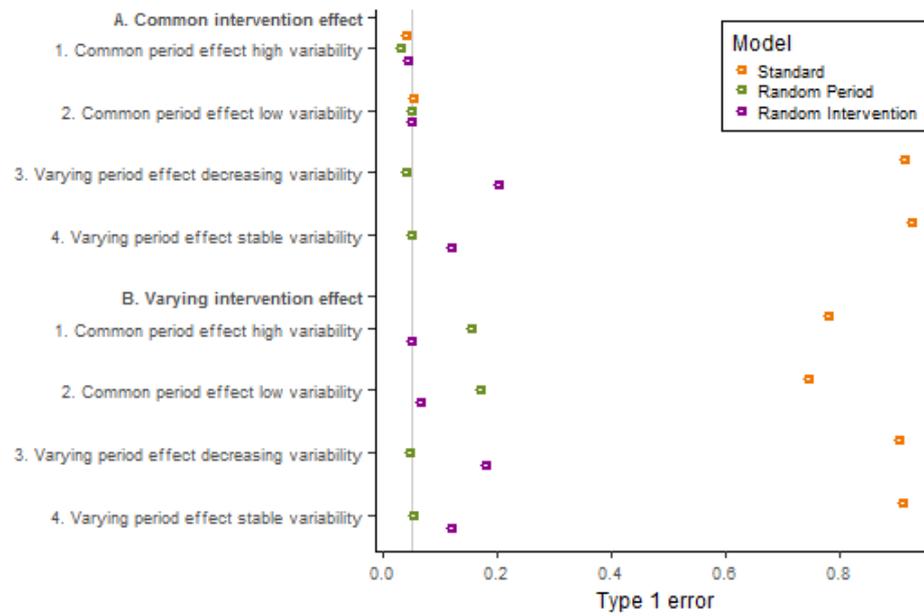
Simulation study one: Group two intervention effect log odds ratio the same as group one

Group one intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	1.79	1.79	1.79
0.41	Common	Common period effect, low variability	0.25	0.25	0.25
0.41	Common	Varying period effect, Decreasing variability	0.38	1.78	1.08
0.41	Common	Varying period effect, Stable variability	1.26	1.77	1.47
0.41	Varying	Common period effect, high variability	1.91	1.88	1.79
0.41	Varying	Common period effect, low variability	0.38	0.34	0.25
0.41	Varying	Varying period effect, Decreasing variability	0.51	1.88	1.06
0.41	Varying	Varying period effect, Stable variability	1.39	1.88	1.5
0	Common	Common period effect, high variability	1.78	1.78	1.79
0	Common	Common period effect, low variability	0.25	0.25	0.25
0	Common	Varying period effect, Decreasing variability	0.37	1.77	1.07
0	Common	Varying period effect, Stable variability	1.26	1.78	1.48
0	Varying	Common period effect, high variability	1.91	1.89	1.78
0	Varying	Common period effect, low variability	0.39	0.35	0.25
0	Varying	Varying period effect, Decreasing variability	0.5	1.88	1.07
0	Varying	Varying period effect, Stable variability	1.37	1.87	1.49

Simulation study two: Group two intervention effect log odds ratio different to group one:

Group one Intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	2.06	2.07	2.09
0.41	Common	Common period effect, low variability	0.51	0.51	0.48
0.41	Common	Varying period effect, Decreasing variability	0.76	1.85	1.08
0.41	Common	Varying period effect, Stable variability	1.53	1.84	1.49
0.41	Varying	Common period effect, high variability	2.13	2.11	1.94
0.41	Varying	Common period effect, low variability	0.59	0.53	0.27
0.41	Varying	Varying period effect, Decreasing variability	0.84	1.94	1.06
0.41	Varying	Varying period effect, Stable variability	1.6	1.95	1.49

S10a: Figure of Type-one Error by Simulation Parameters



S10b: Table of Type-one Error by Simulation Parameters

Simulation study one: Group two intervention effect log odds ratio the same as group one

Group one intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	4	4	3
0.41	Common	Common period effect, low variability	5	5	5
0.41	Common	Varying period effect, Decreasing variability	93	7	26
0.41	Common	Varying period effect, Stable variability	91	7	12
0.41	Varying	Common period effect, high variability	78	14	5
0.41	Varying	Common period effect, low variability	75	12	4
0.41	Varying	Varying period effect, Decreasing variability	91	6	23
0.41	Varying	Varying period effect, Stable variability	93	5	11
0	Common	Common period effect, high variability	4	3	4
0	Common	Common period effect, low variability	5	5	5
0	Common	Varying period effect, Decreasing variability	91	4	20
0	Common	Varying period effect, Stable variability	92	5	12
0	Varying	Common period effect, high variability	78	15	5
0	Varying	Common period effect, low variability	74	17	7
0	Varying	Varying period effect, Decreasing variability	90	5	18
0	Varying	Varying period effect, Stable variability	91	5	12

Simulation study two: Group two intervention effect log odds ratio different to group one

This is the percentage of simulations that rejected at the 5% level the null hypothesis that the intervention effect was equal to the true group 1 intervention effect.

Group one Intervention effect	Intervention effect	Period effect	Standard model	Random Period Model	Random Intervention Model
0.41	Common	Common period effect, high variability	100	100	100
0.41	Common	Common period effect, low variability	100	100	100
0.41	Common	Varying period effect, Decreasing variability	100	93	59
0.41	Common	Varying period effect, Stable variability	100	57	77
0.41	Varying	Common period effect, high variability	100	100	100
0.41	Varying	Common period effect, low variability	100	100	100
0.41	Varying	Varying period effect, Decreasing variability	99	83	54
0.41	Varying	Varying period effect, Stable variability	100	53	70

S11: Deworming Trial Data Cleaning

We performed the same data cleaning steps used for the reanalysis of the trial [1, 2] to the data available at:

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28038>

(data downloaded 19/11/2015) These were as follows:

1. Carry forward missing school ID
2. Remove observations after a pupil moved school
3. Remove observations of pupils who have died, finished school, or moved to secondary school
4. Recode pupil drop out as unattended
5. Remove unscheduled visits.

6. Remove pupils from the data if they were never observed in school during the 2 years
7. Remove visits that had more the 70% missing attendance data for pupils.

Supporting information bibliography

- [1] Aiken AM, Davey C, Hargreaves JR and Hayes RJ. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. *International Journal of Epidemiology* 2015. 44. (5):1572–1580.
- [2] Davey C, Aiken AM, Hayes RJ and Hargreaves JR. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. *International Journal of Epidemiology* 2015. 44. (5):1581–1592.

8 Paper C: Robust Analysis of Stepped-Wedge Trials using Cluster-Level Summaries Within Periods

In this chapter, I address the second objective of my second aim: to propose an alternative analysis method of SWT analysis.

In chapter 7, I identified that model 3.1 (the standard model) is sensitive to violations of model assumptions. Although model 3.4 (the random-period model) had better properties than model 3.1, there was still under coverage of confidence intervals in some scenarios. In addition, model 3.4 becomes increasingly complex as the number of periods in the study increases. Therefore, it will not be applicable to all situations in which SWTs are conducted, and so there is a need for an alternative analysis method.

Here, I describe a novel analysis method that makes no assumptions about the period effect: a cluster-summary, within-period analysis. This method is described with a worked example using data from the TB diagnostic trial described in section 2.2. A simulation study is then used to assess the properties of the method. The simulation study was based on the NHS health-check data described in section 2.3.

In this chapter, model 3.1 is referred to as the standard model, all other terminology is consistent with the rest of this thesis.

Ethics approval for this work is given in Appendix D. This paper is currently under review at Statistics in Medicine.

Registry
 T: +44(0)20 7299 4646
 F: +44(0)20 7299 4656
 E: registry@lshtm.ac.uk

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jennifer Thompson
Principal Supervisor	Katherine Fielding
Thesis Title	Improving the design and analysis of stepped-wedge trials

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Statistics in Medicine
Please list the paper's authors in the intended authorship order:	Jennifer Thompson and Calum Davey, Katherine Fielding, James Hargreaves, Richard Hayes
Stage of publication	Undergoing revision

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	See attached sheet
--	--------------------

Student Signature: 

Date: 12/09/17

Supervisor Signature: 

Date: 12/9/2017



SECTION D – Multi-authored work

For multi-authored with, give full details of your role in the research included in the paper and in the preparation of the paper.

The ideas and the majority of the work for this paper were divided between myself and Calum Davey. I led identifying the methods behind conducting a simulation study and we both decided on the design of the simulation study. We both looked for datasets to use as the basis for the simulation study, and I determined the representation of the chosen data in terms of statistical distributions. Coding the simulation studies was shared between us; I created a first draft of the simulation study whilst Calum Davey created a first draft of the code to conduct the within-period analysis, which we then combined and finalised. We both conducted the analysis and interpretation of results. I conducted the example analysis using the within-period analysis method on a real SWT. Calum Davey drafted a plan for the paper before I wrote the first draft.

Abstract

Cluster-summary analysis methods are one way to account for clustering in parallel cluster-randomised trials that has benefits over other methods [1]. In stepped-wedge trials (SWTs), a cluster-summary analysis is more complex; outcomes cannot be summarised across the cluster because the cluster experiences both the control and intervention, and the intervention is confounded with time. Instead, outcomes can only be summarised within periods of the trial in which clusters remain in the same condition.

We propose a cluster-summary method to analyse SWTs. This method estimates the intervention effect by performing cluster-summary analyses comparing intervention and control arms within each of the trial periods where each cluster is in either the intervention or control arm. The effects from each period are combined with an inverse-variance-weighted average. We use permutation tests to generate a p-value and confidence intervals. We present an example where we apply this method to a previously published SWT. Using simulated data, we compared the cluster-summary method to a commonly used mixed-effect model (the standard model), which has a random effect for cluster and fixed effects for period and the intervention. We simulate scenarios with cluster-specific period effects drawn from a distribution informed by real data, and period effects common to all clusters.

The cluster-summary method provided unbiased estimates of the intervention effect and valid inference for all scenarios. The standard model failed to provide reliable inferential statistics in scenarios with period effects that vary between clusters, but had greater power than the cluster-summary method when period effects were common to all clusters.

The cluster-summary method for analysing SWT is a more robust analysis than the standard model. Our new method provides an alternative method for analysing SWTs when there is uncertainty about the assumptions necessary for the mixed-effects model.

8.1 Introduction

Background

Parallel cluster-randomised trials (CRTs) with sufficient numbers of clusters can be analysed using cluster summaries of observation-level data, or by accounting for correlation in the observation-level data using statistical models. Cluster-summary methods are robust and allow simple computation of risk differences and risk ratios [1, 2]. In contrast, observation-level models with binary outcomes tend to provide an odds ratio, which is more difficult to interpret [3].

Stepped-wedge trials (SWTs) randomise clusters to sequences that start receiving an intervention at different times. The trial is divided into periods between the times when clusters switch from the control to the intervention. Outcomes are typically observed during all periods [4], and they may also be observed before the first sequence starts receiving the intervention and after the last sequence starts to receive the intervention.

The intervention effect can be estimated from two directions of comparison, sometimes referred to as vertical and horizontal comparisons.

Vertical comparisons compare observations corresponding to the control condition with observations corresponding to the intervention condition during each period of the trial, i.e. within periods. These comparisons are randomised comparisons.

Horizontal comparisons compare observations corresponding to control and intervention conditions within each cluster across periods, i.e. between periods. Horizontal comparisons are confounded with changes in the outcome over time (period effects) since the intervention condition is later in time than the control condition.

The most commonly used method to analyse SWTs [5, 6] is a mixed-effect model proposed by Hussey and Hughes [7]. The method combines vertical and horizontal comparisons. The model uses observation-level regression to fit a random effect for the clusters, a fixed effect for the intervention, and fixed effects for periods. Since the model adjusts for period effects with fixed effects for each period, it assumes that the period effects are the same in all clusters, and so it assumes exchangeability of observations within clusters. This model has been shown to severely underestimate the standard error of the intervention effect when period effects vary between clusters, resulting in under coverage of

confidence intervals [8–11].

The correlation structure of the outcomes within clusters is often not known in advance of a trial. To avoid underestimating the required sample size, it may be safer to plan an analysis that does not assume that the period effects are the same in all clusters. There are a few such methods published that use more complex mixed-effect models [10, 12]. However, a recent review of SWTs found that 70% (26/37) of recently conducted SWTs had fewer than 30 clusters in total [13]; in such scenarios mixed-effect models can be unreliable [1], so an alternative method is required.

In this paper, we introduce a simple cluster-summary method for the analysis of SWTs that uses only the vertical, within-period comparisons. Since the method uses cluster summaries instead of random effect to account for clustering, it does not require a large number of clusters to remain reliable [1], and enables the simple calculation of a risk difference for a binary outcome. We evaluate the performance of this method against the commonly used mixed-effects model (referred to as the standard model).

The Cluster-Summary Analysis Method

We extend a cluster-summary analysis method for CRTs to SWTs. We focus on binary outcomes, which are common in SWTs [6], and demonstrate calculation of a risk difference. The principles of the method can be applied to other outcome types (continuous, time-to-event, etc.) and effect measures (risk ratios, odds ratios, etc.).

The method uses vertical comparisons only. Since vertical comparisons are not possible during periods where all clusters are in one condition, the method only uses data collected after the first sequence switches to the intervention and before the last sequence switches to the intervention.

The method uses the following three stages:

1. First, we calculate the risk of the outcome in each cluster i during each period j , as the proportion of individuals with the outcome of interest, p_{ij} .
2. Next, we calculate the period-specific risk difference as the difference between the mean risk in the intervention and the control conditions

during period j :

$$\hat{\theta}_j = \frac{1}{c_{1j}} \sum_{i: X_{ij}=1} p_{ij} - \frac{1}{c_{0j}} \sum_{i: X_{ij}=0} p_{ij}$$

where c_{1j} and c_{0j} are the numbers of clusters in the intervention condition and the control condition respectively during period j , X_{ij} is an indicator of whether cluster i was allocated to receive the intervention or control condition during period j , equal to 1 for intervention and 0 for control.

3. Last, we combine the period-specific risk differences using an inverse-variance weighted average, to give an overall estimated risk difference $\hat{\theta}$:

$$\hat{\theta} = \sum_{j=1}^{j=J} \frac{w_j}{w} \hat{\theta}_j$$

where $w = \sum_{j=1}^{j=J} w_j$.

We use weights based on the pooled variance of the period-specific estimated risk difference allowing for the unequal number of clusters in each condition. These weights are calculated as:

$$w_j = Var(\hat{\theta}_j)^{-1} = \left(\left(\frac{(c_{0j} - 1)s_{0j}^2 + (c_{1j} - 1)s_{1j}^2}{c_{0j} + c_{1j} - 2} \right) \left(\frac{1}{c_{0j}} + \frac{1}{c_{1j}} \right) \right)^{-1}$$

where s_{0j}^2 and s_{1j}^2 are the empirical variances of the risks in the control and intervention conditions respectively at each period j .

To calculate a ratio, such as a risk ratio, take the log of the cluster-period summaries in stage (1). The log risk is not defined for a proportion of 0. To include these clusters, a heuristic adjustment can be used adding a half to both the number of individuals with the outcome of interest and the number of individuals without the outcome, in the affected clusters only [14]. If using log odds, this adjustment will also be required in cluster-periods with a proportion of 1.

We use permutation tests to calculate a p-value and confidence intervals, because they require no assumptions about the correlation between the period-specific effects [15]. We randomly permute the assignment of clusters to sequences, and therefore the time at which clusters switch from control to intervention conditions. We calculate an intervention effect for each permutation, as described above. The p-value against the null hypothesis of no intervention effect is given by the proportion of permutations with an estimated interven-

tion effect the same as or more extreme than that observed. Throughout this paper, we have used 1000 permutations to allow us to calculate p-values to 3 decimal places.

Permutation tests can generate 95% confidence intervals using an iterative process. The following process is conducted iteratively for several given intervention-effect values to find the values that return a one-sided p-value of 0.025: these are then the upper and lower confidence limits.

Continuing with our example using the risk difference, first, we test for evidence against a given risk difference θ_A by subtracting it from the observed risks in the intervention condition:

$$p_{ij}^* = \begin{cases} p_{ij} - \theta_A & \text{if } X_{ij} = 1 \\ p_{ij} & \text{if } X_{ij} = 0 \end{cases}$$

A permutation test is performed (as described above) using the new intervention-condition risks and the original control-condition risks, p_{ij}^* . This provides a p-value testing the null hypothesis that $\theta = \theta_A$.

R code for conducting this analysis, from stage (2) onwards, is given in S1.

The cluster-summary method assumes the clusters are independent of one another within each period and assumes that observations are independent within clusters within each period. No assumptions are made about correlations between observations or cluster summaries over different periods, and no assumptions are made about the period effects. We assume that the intervention effect is the same for all clusters in all periods.

The Standard Model

As a comparison to our novel method we used the mixed-effect model described by Hussey and Hughes [7], herein referred to as the standard model, adapted to a logistic model for a binary outcome as shown below:

$$\text{logit} \{P(y_{ijk} = 1 \mid X_{ij}, u_i)\} = \mu + \beta_j Z_j + \theta X_{ij} + u_i$$

where y_{ijk} is the outcome in patient k at time j in cluster i , μ is the log odds of the outcome in the first period in the control condition, β_j is the change in the outcome from the first period to period j , Z_j is one in period j and zero

otherwise, θ is the intervention effect, X_{ij} is an indicator as to whether cluster i has the intervention during period j , $u_i \sim N(0, \sigma_u^2)$ is a random effect for cluster.

This analysis model assumes that clusters are independent of one another. The model assumes that the intervention effect and period effects are common to all clusters, and that observations are equally correlated within clusters across all periods.

8.2 Application to Tuberculosis Diagnostic Trial

We applied the cluster-summary method to a SWT conducted in Brazil that assessed the effect of a new tuberculosis (TB) diagnostic test on patient outcomes [16].

The new diagnostic test (Xpert MTB/RIF) is more sensitive than the standard smear microscopy method and provides a result for rifampicin drug resistance [17]. The impact of this new test on patient outcomes is unclear, and so this trial aimed to clarify this effect. The trial defined unfavourable outcomes as death, lost-to-follow-up during treatment, transfer out of clinic (including to more specialist centres for those not responding to treatment), or suspected drug resistance. Clusters were defined as laboratories used to diagnose TB in two cities in Brazil. After one month of baseline data collection with all laboratories in the control condition, randomly selected pairs of the 14 trial laboratories started using the new diagnostic technology at the start of each month. Data were collected for one month after all laboratories had received the intervention. Unlike the published analysis of this trial we do not include data from the periods when all laboratories were in the control condition, or when they were all in the intervention condition. We estimated from the data that patients' outcomes had an intra-cluster correlation coefficient (ICC) of 0.003 throughout the trial.

In the published report [16], the authors found no evidence that the new diagnostic test reduced unfavourable outcomes. They used a mixed-effects model with a fixed effect for the intervention, and a random effect for cluster (crude OR=0.92 95% CI 0.79, 1.06; a similar result was found in an analysis adjusting for sex, age, city, HIV status and diagnosis status). This analysis did not adjust for period effects. We found a similar result when we repeated this analysis on our subset of data that excludes data from the period when

all laboratories were in the control or all were in the intervention condition (results not shown).

We applied the cluster-summary method to estimate the intervention effect risk difference and odds ratio. We compared this to the mixed-effects model estimate of the intervention effect odds ratio, which used a fixed effect for period (allowing the outcome to change over time but assuming the same change in all clusters) [7].

Table 8.1: Stages of estimating the risk difference using the cluster-summary method. The mean and variance of cluster-level risks are calculated for each condition in each period. These are used to calculate the risk difference and its variance in each period. An inverse-variance weighted average of these gives an overall effect estimate.

		Period					
Stage in cluster-summary method:		2	3	4	5	6	7
(1)	Control: Mean of cluster-level risks (%)	30.5	26.5	33.9	32.4	40.7	38.4
	Intervention: Mean of cluster-level risk	32.2	23.8	23.9	30.2	33.3	34.6
(2)	Risk difference (%)	+1.7	-2.7	-10.0	-2.2	-7.4	-3.7
(3)	Control						
	Number of clusters	12	10	8	6	4	2
	Variance of cluster-level risks (%)	0.009	0.005	0.006	0.005	0.012	0.036
	Intervention						
	Number of clusters	2	4	6	8	10	12
	Variance of cluster-level risks	0.040	0.015	0.001	0.001	0.007	0.020
	Weight	0.05	0.13	0.27	0.41	0.11	0.03
	Weighted average of risk difference (%)			-4.8			

Table 8.1 shows the three stages in estimating the intervention effect risk difference using the cluster-summary method. The risk difference for the intervention effect ranged from +1.7% to -10.0% across the 6 periods with negative values indicating a reduced risk of an unfavourable outcome in the interven-

tion. The overall estimated risk difference, combining estimates across the 6 periods, was -4.8%.

The cluster-summary method estimated a -4.8% (95% CI -10.0%, -0.3%) reduction in the risk of an unfavourable outcome using the new diagnostic test with some evidence of an effect ($p=0.04$). The cluster-summary method estimated an OR=0.78 (95% CI 0.61, 0.96; $p=0.02$). The standard model with a fixed effect for period gave an odds ratio similar to this (OR= 0.83 95% CI 0.67, 1.03; $p=0.10$).

8.3 Simulation Study

Methods

We performed a simulation study to investigate the performance of the cluster-summary method in scenarios with different period effects and intraclass correlation coefficients (ICCs), and for several SWT designs.

We used local-authority-level data on uptake of NHS health checks in England in 2013-2014, available from Public Health England [18]. Health checks were offered to all adults aged 40-74 every five years by general practices (GPs) and third parties to assess risk of diabetes, heart disease, kidney disease, stroke, and dementia [19]. The mean of the local authority-level percentage of patients accepting health checks when offered was 49% in the first quarter of 2013; this increased to 54% in the last quarter. At the start of 2014, the mean was 46%; this increased to 56% in the last quarter.

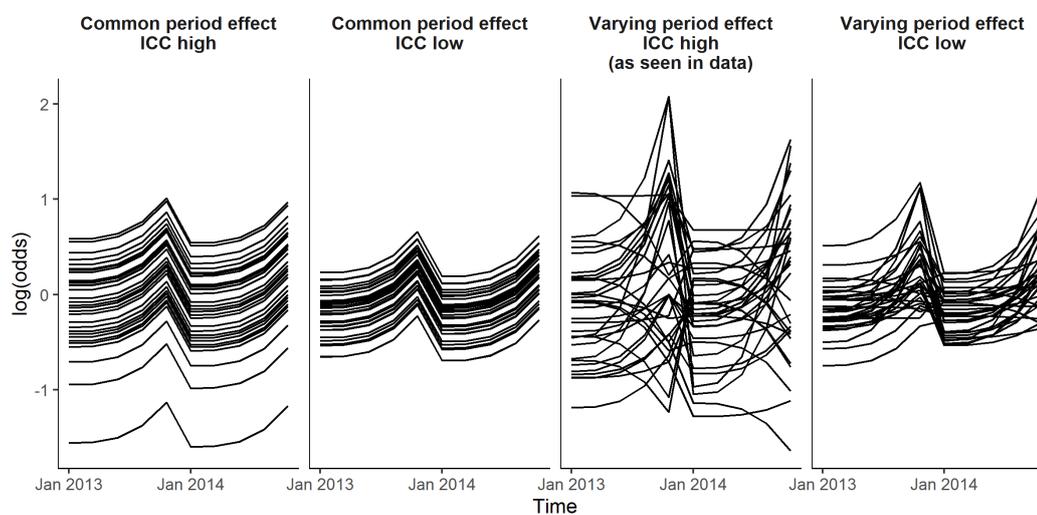
Data generation

We used these health check data to generate four scenarios (Figure 8.1). Details of how we generated the scenarios to be used in the simulation study from these data are given in S2.

We simulated period effects that were common to all clusters, and period effects that varied between clusters to the degree that was observed in the data to check that the cluster-summary method remained unbiased and gave correct confidence interval coverage in a range of scenarios.

We simulated two ICC scenarios to assess the power of the analysis with different values of ICC. For one we used the between-cluster variability observed

Figure 8.1: Simulation study scenarios secular trends and ICC. Based on NHS health-check uptake in England



in the data (ICC=0.08 in the first quarter of 2013, hereafter referred to as ‘high ICC’) and for another we used one-fifth of the observed between-cluster variability (ICC=0.02 in the first quarter of 2013, hereafter referred to as ‘low ICC’).

The four scenarios were therefore: (1) common period effects and high ICC, (2) common period effects and low ICC, (3) varying period effects and high ICC, and (4) varying period effects and low ICC.

When the period effects varied between the clusters, the between-cluster variance changed over time. Therefore, the ICC changed over time. Over the two years the ICC varied between 0.06 and 0.19 for the high ICC and varying period effects scenario, as observed in the data, and between 0.01 and 0.04 for the low ICC and varying period effects scenario.

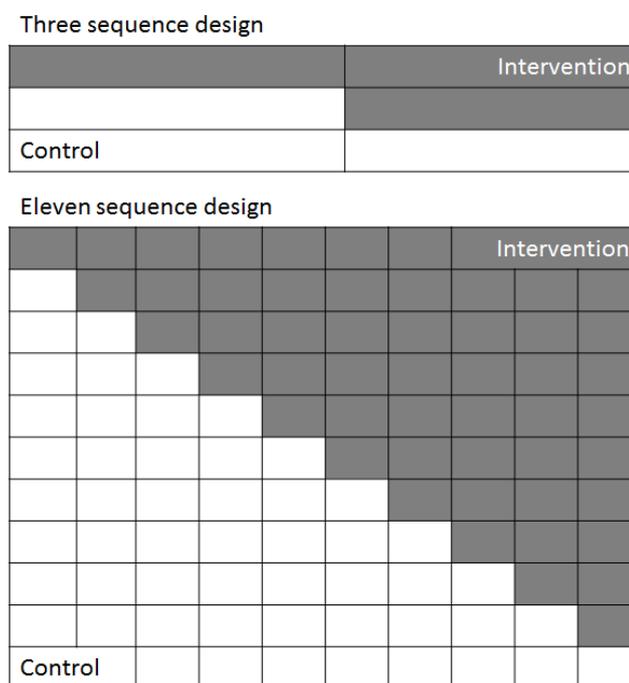
Trial designs

We simulated SWTs in each of these four scenarios that assessed the effect of an intervention designed to increase the acceptance of health checks. The simulated intervention effect had an odds ratio of 1.3 favouring the intervention (log odds ratio=0.26).

We simulated four trial designs for each of the four scenarios to assess how the numbers of sequences, the number of clusters per sequence, and the total number of clusters affected the power of the cluster-summary method. The four

trial designs had either 3 or 11 sequences with either 3 or 11 clusters per sequence (Figure 8.2). This resulted in four trials design with a total of 9 clusters (3 sequences with 3 clusters per sequence), 33 clusters (3 sequences with 11 clusters per sequence, or 11 sequences with 3 clusters per sequence), or 121 clusters (11 sequences with 11 clusters per sequence). Unlike the mixed-effects model, the cluster-summary method requires clusters in both the control and intervention condition at each period so our trial designs began after the first sequence switched to the intervention, and finished before the final sequence switched to the intervention (Figure 8.2).

Figure 8.2: Trial schematics used in simulation study



The total number of observations for each cluster across the trial was selected from a log-normal distribution ($\mu = 5.3, \sigma^2 = 0.25$) regardless of the trial design; this gave a median cluster size of 200 (IQR 143 - 281) with observations evenly distributed across the periods. In scenarios with common period effects and high ICC, this would give the smallest trial (9 clusters) approximately 31% power to detect an odds ratio of 1.3 with the standard model, and the largest trial (121 clusters) approximately 100% power [20].

Each of the four trial designs for each of the four scenarios was simulated 1,000 times, allowing us to estimate coverage of 95% confidence intervals to within 1.4%.

Analysis methods and evaluation

We analysed each simulated trial using the cluster-summary method and the standard model. The cluster-summary method calculated an odds ratio for comparability with the standard model, although this is unlikely to be the measure of choice in practice. We compared the two analysis methods in terms of bias, coverage, and power for each trial design and scenario, in line with recommendations from Burton *et al.* [21].

We calculated the proportion of standard models that converged and the proportion of cluster-summary analyses that required the heuristic adjustment. Bias was calculated as the deviation of the mean of the estimated intervention effect log odds ratio from the true log odds ratio. Effect estimates within half a standard deviation of the true effect were considered unbiased. Below this cut off, bias has been shown to have little effect on the type-one error rate [21, 22]. We compared the variability of the estimates given by each analysis method using the ratio of the variances. The coverage of the 95% confidence intervals was calculated as the proportion of simulations with $p > 0.05$ against the true effect, i.e. the proportion of confidence intervals that contained the true effect. We calculated the power to detect an effect at 5% significance as the proportion of simulations with $p < 0.05$ against no intervention effect.

Simulation Study Results

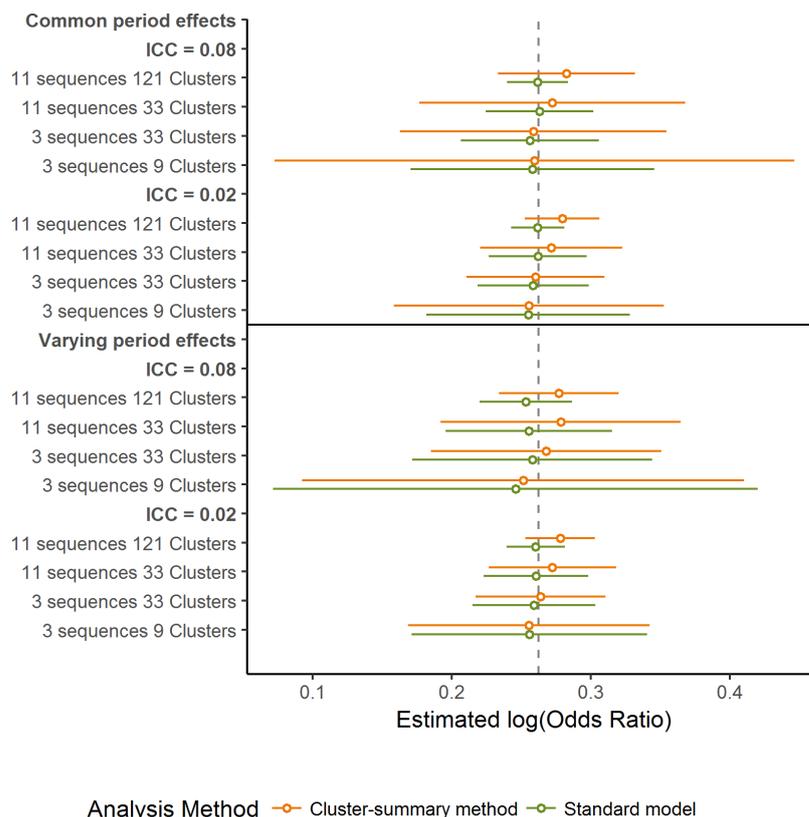
For all simulations, the cluster-summary method provided an effect estimate and p-value and the mixed-effects model converged. The heuristic adjustment was used in 83%-100% of simulations with a trial design of 11 clusters per sequence and 11 sequences (mean of 2-7 times per simulation), 9% - 63% with 3 clusters per sequence and 11 sequences (mean of 0-1 times per simulation), in 1 simulation with 11 cluster per sequence and 3 sequences, and was never used with 3 clusters per sequence and 3 sequences. See S3 for more details.

Bias and variability

Figure 8.3 shows the mean and half a standard deviation either side of the mean of the estimated intervention effects for each analysis method, scenario, and trial design. The mixed-effect model estimates were unbiased. The cluster-summary method estimates were consistently larger than the true effect in

all scenarios with 11 sequences, however this small bias was never larger than half a standard deviation (see S4).

Figure 8.3: Intervention effect estimates in each scenario and trial design, by analysis method. Hollow circle is the mean, solid line is $\pm 1/2$ standard deviation. Grey dotted line is true effect

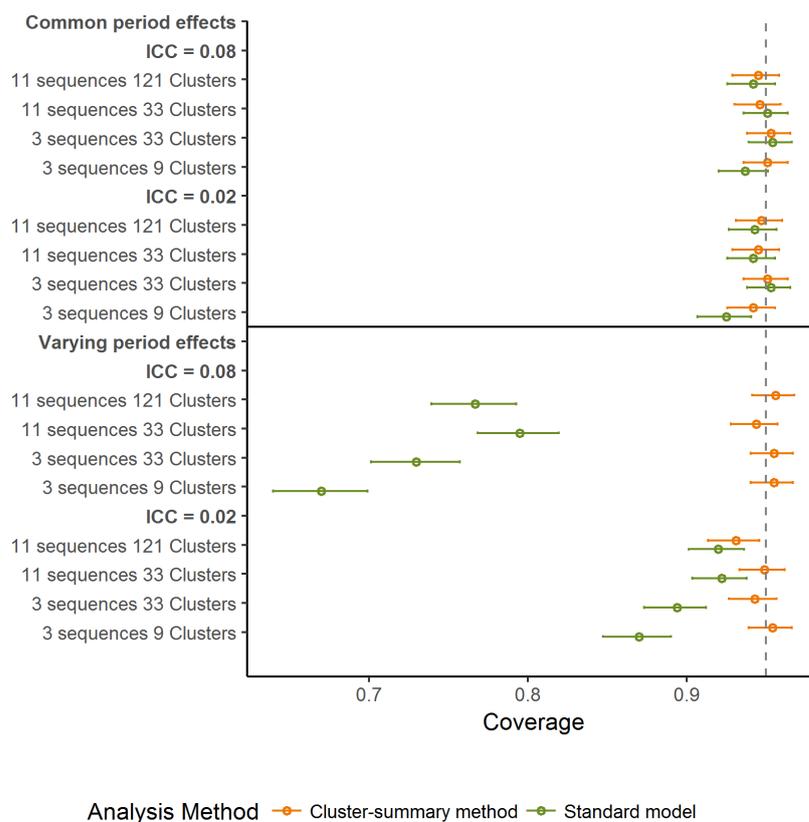


The cluster-summary method estimates were between 1.76 and 6.07 times more variable than the standard model estimates when period effects were common to all clusters. When period effects differed between clusters, the variability of the estimates was more alike between analyses: the cluster-summary method estimates were between 0.83 and 2.08 times as variable as the standard model estimates. This difference between the common-period-effect scenarios and varying-period-effect scenarios was largely driven by an increase in variability of standard model estimates. In all scenarios, the difference in variability was greater when there were more sequences and when the ICC was higher because of a reduction in variability in the standard model estimates and an increase in variability in the cluster-summary analysis estimates respectively.

Coverage

Figure 8.4 shows the coverage of the 95% confidence intervals, and its confidence interval, for each method, scenario, and trial design (see S5 for further details).

Figure 8.4: Coverage of 95% confidence intervals in each scenario and trial design, by analysis method. Hollow circle is the coverage. Barred lines are 95% confidence intervals



When period effects were common to all clusters, shown in the upper half of Figure 8.4, both analysis methods gave reasonable coverage (92%-95%), although there was some suggestion that the standard model gave under-covered confidence intervals with 9 clusters (94% coverage and 92% coverage for high and low ICC respectively).

When period effects varied between clusters, shown in the lower half of Figure 8.4, the cluster-summary analysis achieved 93%-96% coverage for both levels of ICC and all trial designs. In contrast, the standard model produced severe under coverage of 95% confidence intervals for both levels of ICC and most

trial designs. The degree of under coverage was more severe when the ICC was high (67%-80% coverage and 87%-92% coverage for high and low ICC respectively) and with fewer sequences.

Power

Table 8.2 shows the power for each method, scenario, and trial design. When period effects vary, the power for the standard model is not shown because of under coverage of confidence intervals (see Figure 8.4).

Table 8.2: Statistical power in each scenario and trial design, by analysis method. When period effects vary, the power for the mixed-effects model is not shown because of the under coverage of confidence intervals (see Figure 8.4).

Period effect	ICC	Number of sequences	Clusters per sequence	Power (%)		
				Cluster-summary analysis	Mixed-effect model	Difference
Common	0.02	3	3	24	48	22
	0.02	3	11	73	90	17
	0.02	11	3	76	97	21
	0.02	11	11	100	100	0
	0.08	3	3	10	35	25
	0.08	3	11	27	73	46
	0.08	11	3	31	93	62
	0.08	11	11	81	100	19
Varying	0.02	3	3	27		
	0.02	3	11	82		
	0.02	11	3	83		
	0.02	11	11	100		
	0.08	3	3	11		
	0.08	3	11	37		
	0.08	11	3	40		
	0.08	11	11	88		

The power of the cluster-summary method ranged from 11% to 100%. Power

was higher with more clusters and lower ICC. For example, with common period effects, 11 sequences, and 3 clusters per sequence the cluster-summary method had 31% and 76% power for high and low ICC respectively. Whether period effects were common or varying had limited impact on the power of the cluster-summary analysis.

When period effects were common, the cluster-summary analysis had lower power than the standard model for both levels of ICC and all trial designs. The difference ranged from 0% to 63% less power; power loss was greatest when ICC was high and with 11 sequences.

8.4 Discussion

We have presented a cluster-summary approach to analyse SWTs. To the best of our knowledge the analysis of SWTs using cluster-summaries of vertical comparisons has not been used previously, despite many of the trials having a small number of clusters where model based approaches to analysis are less appropriate.

We have shown that the cluster-summary method accounts for period effects by using an example where time was a confounder. The original analysis of the TB diagnostic SWT concluded that there was no evidence of an effect on patient outcomes. In contrast, both the cluster-summary method and the mixed-effects model accounting for period effects found larger effect estimates with some evidence of an intervention effect.

Our simulation study found that the cluster-summary method is more robust than the standard model, which produced confidence intervals with severe under coverage when period effects varied between clusters. Previous research has also identified this under coverage [8–11].

The cluster-summary method, however, had inferior performance to the standard model when period effects were common to all clusters (i.e. when the mixed-effects model assumptions were true). In this scenario, the cluster-summary method had much higher variability in the estimated intervention effects and, therefore, lower power, particularly when the ICC was high. The challenge is knowing if period effects will be common to all clusters when creating an analysis plan and calculating the sample size for a trial. Choosing a method that is robust irrespective of the period effects avoids under-powering the trial if this assumption is not met.

The cluster-summary method required a heuristic for when the prevalence of the outcome in a cluster during a period was either 0 or 1 when calculating an odds ratio. The heuristic was required in the scenarios with 11 sequences because the cluster size in each period was smaller in these scenarios. The use of the heuristic increased with the total number of clusters. The heuristic was not required for calculating risk difference and would not be required for other difference estimates, such as mean differences or rate differences, but it would be required for other ratio estimates.

In our study, there was a small bias away from the null in scenarios with 11 sequences but not in scenarios with 3 sequences. This small bias has been previously observed for estimation of odds ratios when clusters are small [23, 24]. When clusters are small, the cluster-level log-odds are overestimated [24], leading to biased effect estimates as observed. Cox and Snell [24] suggest a simple adjustment of adding a half to the number of individuals with and without the outcome in all clusters to correct this overestimation and we recommend using this adjustment when using a cluster-level analyses to calculate an odds ratio. This correction also negates the need to specifically correct clusters where individuals all have the same outcome.

This study confirms the findings of Wang and DeGruttola [11], that the magnitude of under coverage of confidence intervals from the standard model when period effects vary between clusters is dependent on the degree of clustering. Our study also builds on previous research by showing that the degree of under coverage increases with fewer sequences; this was caused by greater variability in the intervention effect estimates in the scenario with fewer sequences. Our simulation study also suggests that the mixed-effect model may produce under-covered confidence intervals when used with only 9 clusters, similar to the findings of Barker *et al* [25].

We have found that the power of the cluster-summary method is dependent on the total number of clusters, and only marginally on the number of sequences. This contradicts the findings of Moulton *et al* [26] who found that the power of a vertical analysis decreased as the number of sequences increased. The contradiction may be due to Moulton *et al* ignoring clustering. Further research is needed to understand how the number of sequences affects the power of this methods in a wider range of scenarios. Consistent with previous research [7, 27], the statistical power of the standard model was dependent on the number of sequences as well as the number of clusters.

A strength of the simulation study is that we used NHS health check uptake data at the local-authority level to demonstrate the effect of realistic variability in period effects. We explored the performance of the cluster-summary method in a range of realistic settings by varying the ICC and looking at several different trial designs.

The simulation study has limitations. For our results to be most applicable to epidemiological research we explored binary outcomes and to enable comparison with a mixed-effect model estimated an odds ratio. Further research is required to examine whether the standard model is more robust to misspecification with a continuous outcome, and to explore the properties of the cluster-summary method with continuous and rate outcomes. Other features of SWTs, such as delayed intervention, sparse data collection, or lags were not considered. The performance of the cluster-summary analysis in the presence of more extreme ICCs requires further research.

There may be other analysis methods that are able to provide robust estimates whilst including the horizontal comparisons. Research is needed into generalized estimating equations or the addition of robust standard errors in the context of SWTs as these methods have been shown to be robust to misspecification in other settings [28]. However, these methods still rely on having many clusters and generally compute odds ratios.

The weights used for each period-specific effect were chosen to minimise the number of assumptions. The permutation test was chosen as this makes no assumptions about the composition of the data other than that of exchangeability and a common intervention effect [15]. There may be improvements in power available from making assumptions about the data. Granston provided a variance formula for a within-period analysis under the assumptions of the Hussey and Hughes analysis model, which would remove the need for permutation tests [29]. However, the addition of assumptions negates many of the benefits of the cluster-summary method. The method could be extended to incorporate an adjustment for baseline individual-level covariates, as described by Hayes and Moulton [1].

We have shown that the cluster-summary method can be simply adapted to provide estimates of risk differences. The method could also be extended to calculate risk ratios. While there are methods to estimate risk ratios for CRT with at least 50 clusters using a Poisson model [30], which could be adapted for SWT, we are not aware of methods to estimate risk differences.

Presenting absolute as well as relative effect sizes can improve interpretation, as is recommended by the CONSORT statement [31].

A limitation of a cluster-summary method is that power is lost when there is high variability in cluster sizes. Whilst we allowed cluster size to vary, we did not explore the effect of different levels of variability on power. When there is high variability in cluster size the method could be extended to weight clusters in the analysis by their size as described by Hayes and Moulton to improve precision [1].

A cluster-summary approach to the analysis of SWTs is an unbiased and robust alternative to the commonly used mixed-effects modelling approach. Researchers designing and analysing SWTs should consider this as an alternative to mixed-effects models.

Acknowledgements

We would like to thank Professor Anete Trajman, Dr Betina Durovni, Dr Valeria Saraceni, Professor Frank Cobelens, and Dr Susan van den Hof for making available the original data from their study.

8.5 Supporting Information

S1: R Code for Cluster Summary Analysis

```
cluster.summary <- function(data, cluster, sequence, time, rx,
                             summary, null) {
  f.effect <- function(x){
    #Aggregate data by time and intervention condition
    slices.long <- aggregate(summary ~ time + rx,
                              data = x,
                              FUN = function(X)
                                c(mean = mean(X, na.rm = TRUE),
                                  n = sum(!is.na(X)),
                                  var = var(X, na.rm = TRUE)))

    slices.long$mean <- slices.long$summary[,1]
    slices.long$n <- slices.long$summary[,2]
    slices.long$var <- slices.long$summary[,3]

    #reshape so one row per time slice
    slices <- reshape(direction = "wide",
                      data = slices.long[,c("time", "rx", "mean",
                                             "n", "var")],
                      v.names = c("mean", "n", "var"),
                      idvar = c("time"),
                      timevar = "rx")

    #calculate difference
    slices$diff <- slices$mean.1 - slices$mean.0

    #Calculate a weight assuming the same variance in both arms
    slices$wgt <- ( (((slices$n.0 - 1) * slices$var.0 +
                      (slices$n.1 - 1) * slices$var.1) /
                      (slices$n.0 + slices$n.1 - 2)) *
                   (1/slices$n.0 + 1/slices$n.1) )^-1

    weighted.mean(slices$diff, slices$wgt, na.rm = TRUE)
  }

  f.permute <- function(y, design) {

    design$cluster <- sample(design$cluster, replace = FALSE)

    design.long <- reshape(design,
                           direction = "long",
                           idvar = c("cluster", "sequence"))

    permuted.data <- merge(y[, c("cluster", "time", "summary")],
                           design.long,
```

```
        by = c("cluster", "time"))

    return(f.effect(permuted.data))
}

dataset <- data.frame(cluster = eval(substitute(cluster), data),
                     sequence = eval(substitute(sequence), data),
                     time = eval(substitute(time), data),
                     rx = as.numeric(eval(substitute(rx), data)),
                     summary = eval(substitute(summary), data))

dataset$summary <- dataset$summary - null * dataset$rx

#Calculate observed effect
mean.effect <- f.effect(dataset)

#Create a dataset of the design
design <- reshape(direction = "wide",
                 idvar = c("cluster", "sequence"),
                 timevar = "time",
                 v.names = "rx",
                 data = unique(dataset[,c("cluster", "sequence",
                                         "rx", "time")]))

#Calculate p-value under null = 0
permutations <- replicate(10000,
                         f.permute(dataset, design),
                         simplify = TRUE)

p <- sum(abs(permutations) > abs(mean.effect)) / 10000

return(c(mean.effect, p))
}
```

S2: Data Generating Process

We based the simulations on an outcome that might plausibly be the target of an SWT (uptake of NHS health checks) with realistic clusters (English local authorities).

Across England, GP surgeries and third parties offer all adults between the ages of 40-74 a health check to assess the patient's risk of diabetes, heart disease, kidney disease, stroke, and dementia. A recent study found that, while uptake of health checks has improved since their introduction, uptake is still low: in 2012, 30% of patients accepted the offer of a health check [19]. Data on the uptake of these health checks is published on the Public Health England website for each year quarter, by local authority [18]. In this study we used data from 2013-2014 to simulate trials designed to study the effect of an unspecified intervention to improve uptake of health checks.

Local authorities were removed from the dataset if at any time point they recorded 100%, or 0% uptake of health checks, they had no offers of health checks, or they had more cases of health check uptake than offers. We then visually inspected the log odds of health check uptake to identify any outlying local authorities; a further 2 local authorities were removed as they had unusually high uptake in some quarters. This process removed a total of 29/152 local authorities.

In the remaining 123 local authorities we analysed the health check uptake to assess how uptake was changing over time, and how this varied between the local authorities. For each local authority we modelled the acceptance of health checks as:

$$\text{logit}(P(y_{ijk} = 1)) = \mu_i + \beta_{1i}t_{1ij} + \beta_{2i}t_{2ij}^{\gamma_i}$$

Where y_{ijk} is health check acceptance in local authority i at time j , μ_i is the log odds of acceptance in local authority i in the first quarter of 2013, t_{1ij} is an indicator of year; 0 in 2013 and 1 in 2014, β_{1i} is the log odds ratio comparing health check acceptance in 2014 to 2013 in local authority i , $t_{2ij}^{\gamma_i}$ is the quarter 1,2,3 or 4 within each year to a power γ_i selected using fractional polynomials to allow for a non-linear trend (more details below), and β_{2i} is the log odds ratio for quarter in local authority i . This meant that we were assuming the same effect of quarter in both years.

This model was run for each local authority using fractional polynomials to

select a value γ_i for each local authority. The most common value selected was $\gamma_i = 3$.

We then ran a mixed effect model across all the local authorities where:

$$\text{logit}(P(y_{ijk} = 1)) = \mu + u_i + (\beta_1 + v_{1i})t_{1j} + (\beta_2 + v_{2i})t_{2j}^3$$

$$\begin{pmatrix} u_i \\ v_{1i} \\ v_{2i} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u & \text{cov}(u, v_1) & \text{cov}(u, v_2) \\ \text{cov}(u, v_1) & \sigma_{v_1} & \text{cov}(v_1, v_2) \\ \text{cov}(u, v_2) & \text{cov}(v_1, v_2) & \sigma_{v_2} \end{pmatrix} \right)$$

where u_i is a random intercept, v_{1i} is a random effect for year, and v_{2i} is a random effect for quarter.

The resulting model gave coefficient $\mu = -0.14$, $\beta_1 = -0.04$, and $\beta_2 = 0.01$ with the following multivariate normal distribution:

$$\begin{pmatrix} u_i \\ v_{1i} \\ v_{2i} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.306 & -0.150 & -0.002 \\ -0.150 & 0.253 & -0.001 \\ -0.002 & -0.001 & 0.001 \end{pmatrix} \right)$$

We simulated two scenarios for the period effects of quarter and year:

1. Period effect differs between clusters. The coefficients for quarter, year, and the constant for our clusters were sampled from the multivariate normal distribution as seen in the data and given above.
2. Period effect common to all clusters. The mean coefficient for quarter and year were selected for all clusters and the constant was sampled from a normal distribution with mean and variance from the multivariate normal distribution. i.e.

$$\begin{pmatrix} u_i \\ v_{1i} \\ v_{2i} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.306 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right)$$

We wanted to compare our analysis methods in data with a smaller ICC so we used two scenarios for the intra-cluster correlation (ICC).

1. High ICC: corresponding to the original data and the original multivariate normal distribution. When period effects differed between cluster this meant that the ICC varied between 0.06 and 0.15 over the two years. When period effects were common to all clusters this meant that

the ICC remained the same as the ICC in the first quarter of 2013 in the original data, so $ICC=0.08$ throughout.

2. Low ICC: The between-cluster variance was reduced by multiplying the covariance matrix by 0.2. When period effects differed between the cluster this meant that the ICC varied between 0.01 and 0.03 over the two years. When period effects were common to all cluster this meant that the ICC remained at $ICC=0.02$ throughout.

This gave us four data scenarios for how acceptance of health checks varied over time, and how this varied between local authorities.

S3: Table of the Number of Times the Heuristic Adjustment was used in each Scenario and Trial Design

Period effects	ICC	Number of sequences	Clusters per sequence	Number of simulations	Mean times per simulation p=0	Mean times per simulation p=1
Common	0.02	3	3	0	0.0	0.0
Common	0.02	3	11	0	0.0	0.0
Common	0.02	11	3	86	0.0	0.1
Common	0.02	11	11	827	0.5	1.2
Common	0.08	3	3	0	0.0	0.0
Common	0.08	3	11	0	0.0	0.0
Common	0.08	11	3	417	0.2	0.4
Common	0.08	11	11	992	1.8	3.4
Varying	0.02	3	3	0	0.0	0.0
Varying	0.02	3	11	0	0.0	0.0
Varying	0.02	11	3	115	0.0	0.1
Varying	0.02	11	11	881	0.6	1.5
Varying	0.08	3	3	0	0.0	0.0
Varying	0.08	3	11	1	0.0	0.0
Varying	0.08	11	3	627	0.2	0.8
Varying	0.08	11	11	1000	2.0	5.2

S4a: Table of the Mean of Intervention Effect Log Odds Ratio Estimates in each Scenario and Trial Design by Analysis Method

Period effects	ICC	Number of sequences	Clusters per sequence	Standard model	Cluster-summary method
Common	0.02	3	3	0.25	0.26
Common	0.02	3	11	0.26	0.26
Common	0.02	11	3	0.26	0.27
Common	0.02	11	11	0.26	0.28
Common	0.08	3	3	0.26	0.26
Common	0.08	3	11	0.26	0.26
Common	0.08	11	3	0.26	0.27
Common	0.08	11	11	0.26	0.28
Varying	0.02	3	3	0.26	0.26
Varying	0.02	3	11	0.26	0.26
Varying	0.02	11	3	0.26	0.27
Varying	0.02	11	11	0.26	0.28
Varying	0.08	3	3	0.25	0.25
Varying	0.08	3	11	0.26	0.27
Varying	0.08	11	3	0.26	0.28
Varying	0.08	11	11	0.25	0.28

S4b: Table of the Standard Deviation of Intervention Effect Log Odds Ratio Estimates in each Scenario and Trial Design by Analysis Method

Period effects	ICC	Number of sequences	Clusters per sequence	Standard model	Cluster-summary method	Ratio of variance
Common	0.02	3	3	0.15	0.19	1.76
Common	0.02	3	11	0.08	0.10	1.54
Common	0.02	11	3	0.07	0.10	2.10
Common	0.02	11	11	0.04	0.05	1.92
Common	0.08	3	3	0.18	0.37	4.55
Common	0.08	3	11	0.10	0.19	3.73
Common	0.08	11	3	0.08	0.19	6.07
Common	0.08	11	11	0.04	0.10	5.16
Varying	0.02	3	3	0.17	0.17	1.05
Varying	0.02	3	11	0.09	0.09	1.12
Varying	0.02	11	3	0.08	0.09	1.49
Varying	0.02	11	11	0.04	0.05	1.42
Varying	0.08	3	3	0.35	0.32	0.83
Varying	0.08	3	11	0.17	0.17	0.92
Varying	0.08	11	3	0.12	0.17	2.08
Varying	0.08	11	11	0.07	0.09	1.68

S5: Table of Intervention Effect 95% Confidence Interval Coverage in each Scenario and Trial Design by Analysis Method

Period effect	ICC	Number of sequences	Clusters per sequence	Standard model	Cluster-summary method
Common	0.02	3	3	0.92	0.94
Common	0.02	3	11	0.95	0.95
Common	0.02	11	3	0.94	0.94
Common	0.02	11	11	0.94	0.95
Common	0.08	3	3	0.94	0.95
Common	0.08	3	11	0.95	0.95
Common	0.08	11	3	0.95	0.95
Common	0.08	11	11	0.94	0.94
Varying	0.02	3	3	0.87	0.95
Varying	0.02	3	11	0.89	0.94
Varying	0.02	11	3	0.92	0.95
Varying	0.02	11	11	0.92	0.93
Varying	0.08	3	3	0.67	0.96
Varying	0.08	3	11	0.73	0.96
Varying	0.08	11	3	0.80	0.94
Varying	0.08	11	11	0.77	0.96

Bibliography

- [1] Hayes RJ and Moulton LH. Cluster Randomised Trials. 1st ed. USA: Chapman and Hall/CRC, 2009.
- [2] Donner A and Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. Wiley, 2010.
- [3] Sinclair JC and Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology* 1994. 47. (8):881–889.
- [4] Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G and Hargreaves JR. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials* 2015. 16. (1):352.
- [5] Mdege ND, Man MS, Taylor Nee Brown CA and 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):936–948.
- [6] Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ and Fielding KL. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials* 2015. 16. (1):358.
- [7] Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials* 2007. 28. (2):182–191.
- [8] Bellan SE, Pulliam JR, Pearson CA, Champredon D, Fox SJ, Skrip L, Galvani AP, Gambhir M, Lopman BA, Porco TC, Meyers LA and Dushoff J. Statistical power and validity of Ebola vaccine trials in Sierra Leone: a simulation study of trial design and analysis. *The Lancet Infectious Disease* 2015. 15. (6):703–710.
- [9] Ji X, Fink G, Robyn PJ and Small SS. Randomization inference for stepped-wedge cluster-randomised trials: An application to community-based health insurance. *Annals of Applied Statistics* 2017. 11. (1):1–20.
- [10] Thompson JA, Fielding K, C D, Aiken AM, Hargreaves J and Hayes RJ. Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. *Statistics in Medicine* 2017. 36. (23):3670–3682.

- [11] Wang R and DeGruttola V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. *Statistics in Medicine* 2017. 36. (18):2831–2843.
- [12] Hemming K, Taljaard M and Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. *Trials* 2017. 18. (1):101.
- [13] Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, Thompson JA, Fielding KL, Omar RZ, Ononge S, Hargreaves J and Prost A. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 2015. 16. (1):353.
- [14] Sankey SS, Weissfeld LA, Fine MJ and Kapoor W. An assessment of the use of the continuity correction for sparse data in meta-analysis. *Communications in Statistics-Simulation and Computation* 1996. 25. (4):1031–1056.
- [15] Sprent P and Smeeton N. Applied Nonparametric Statistical Methods, Fourth Edition. CRC Press, 2016.
- [16] Trajman A, Durovni B, Saraceni V, Menezes A, Cordeiro-Santos M, Cobelens F and Van den Hof S. Impact on Patients' Treatment Outcomes of XpertMTB/RIF Implementation for the Diagnosis of Tuberculosis: Follow-Up of a Stepped-Wedge Randomized Clinical Trial. *PloS One* 2015. 10. (4):e0123252.
- [17] Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC and Dendukuri N. Xpert(R) MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database of Systematic Reviews* 2014. (1):CD009593.
- [18] Public Health England. Explore NHS Health Check Data. URL: http://www.healthcheck.nhs.uk/commissioners_and_providers/data/ (visited on 20/08/2015).
- [19] Robson J, Dostal I, Sheikh A, Eldridge S, Madurasinghe V, Griffiths C, Coupland C and Hippisley-Cox J. The NHS Health Check in England: an evaluation of the first 4 years. *BMJ Open* 2016. 6. (1):e008840.
- [20] Hemming K and Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata Journal* 2014. 14. (2):363–380.

- [21] Burton A, Altman DG, Royston P and Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006. 25. (24):4279–4292.
- [22] Schafer JL and Graham JW. Missing data: our view of the state of the art. *Psychological methods* 2002. 7. (2):147.
- [23] Ukoumunne OC, Forbes AB, Carlin JB and Gulliford MC. Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. *Statistics in Medicine* 2008. 27. (25):5143–5155.
- [24] Cox DR and Snell EJ. Analysis of Binary Data Second Edition. Chapman & Hall CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989.
- [25] Barker D, D’Este C, Campbell MJ and McElduff P. Minimum number of clusters and comparison of analysis methods for cross sectional stepped wedge cluster randomised trials with binary outcomes: A simulation study. *Trials* 2017. 18. (1):119.
- [26] Moulton LH, Golub JE, Durovni B, Cavalcante SC, Pacheco AG, Saraceni V, King B and Chaisson RE. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clinical Trials* 2007. 4. (2):190–199.
- [27] Woertman W, Hoop E de, Moerbeek M, Zuidema SU, Gerritsen DL and Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology* 2013. 66. (7):752–758.
- [28] Liang KY and Zeger SL. Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* 1986. 73. (1):13–22.
- [29] Granston T. Addressing Lagged Effects and Interval Censoring in the Stepped Wedge Design of Cluster Randomized Clinical Trials. Thesis. University of Washington, USA, 2014.
- [30] Zou GY and Donner A. Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Statistical Methods Medical Research* 2013. 22. (6):661–670.
- [31] Schulz KF, Altman DG, Moher D and Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010. 340:c332.

9 Paper D: A Stata Command for Conducting Permutation Tests for Stepped-Wedge Cluster Randomised Trials

In CRTs, it has been noted that publication of methodological literature is not sufficient to change practice [1]. Further efforts are also required to improve uptake of new methods.

This chapter addresses the final objective of my second aim: to facilitate the use of the identified analysis method.

In chapter 8, I introduced a novel analysis strategy: analyse each period separately, combine these estimates, and use permutation tests for a p-value and confidence interval.

A limitation of this analysis method is the complexity of implementation. Running this analysis would involve manually coding repetitions of the analysis in each period and manually coding the permutation tests. This is likely to limit the uptake of this analysis method by trialists.

To limit this barrier to use, I wrote a Stata command that would perform this analysis for the user. As well as performing the within-period analysis described in chapter 8, the command can also run the analyses described by Ji *et al* [2], Wang and DeGruttola [3], and Bellan *et al* [4] where permutation tests are used for hypothesis testing on model 3.1 estimates.

Stata [5] is commonly used for the analysis for cluster randomised trials [6]. Not all software allows users to contribute to the software's capabilities, however, Stata does allow this. Users are able to write commands using the same language as when using Stata for analysis. Once a user written command is installed, the commands are executed in the same way as the official Stata

commands. User-written commands are published in the Stata Journal, and added to the Statistical Software Components archive (SSC) [7]. Once a command has been listed on SSC, users can install the command using the Stata command “ssc install”. Users can also provide a dialog box so that the command can be executed using Stata’s user interface. Drukker [8] provides details of how to create a command in his series of blogs “Programming an estimation command in Stata”.

The command is written so that it can conduct permutation tests on estimates from a range of analysis models. To accomplish this, I used a prefix command. This means that my command is specified, followed by a colon, followed by a second command giving the analysis required to produce an estimate.

I have tested the command in a wide range of scenarios following the processes described by Gould [9], and the code includes checks that users of the command have provided appropriate input. I have written the command in a way which minimises the time taken for it to run, however, for a large number of permutations it can take a few minutes to complete.

In this chapter, this Stata command is described and worked examples of its use are given using data from the TB diagnostic trial described in section 2.2. This paper will be submitted to The Stata Journal with the title “swpermutate: Permutation tests for stepped-wedge cluster randomised trials”. Ethics approval is given in Appendix E.



Registry

T: +44(0)20 7299 4646

F: +44(0)20 7299 4656

E: registry@lshtm.ac.uk

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Jennifer Thompson
Principal Supervisor	Katherine Fielding
Thesis Title	Improving the design and analysis of stepped-wedge trials

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Stata Journal
Please list the paper's authors in the intended authorship order:	Jennifer Thompson, Calum Davey, Richard Hayes, James Hargreaves, Katherine Fielding
Stage of publication	Not yet submitted

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I had the idea for and coded the Stata command, help file, and dialog box. I wrote the first draft of the manuscript.
--	---

Student Signature: _____

Date: 12/09/17

Supervisor Signature: _____

Date: 12/9/2017

Abstract

Permutation tests are useful in stepped-wedge trials to provide robust statistical tests of intervention-effect estimates. However, the Stata command `permute` does not produce valid tests in this setting because individual observations are not exchangeable. We introduce the `swpermute` command that permutes clusters to sequences to maintain exchangeability. The command provides additional functionality to aid users in performing analyses of stepped-wedge trials. In particular, we include the option “withinperiod” to perform the specified analysis separately in each period of the study and the resulting period-specific intervention-effect estimates are combined as a weighted average. Examples of the application of `swpermute` are given using data from a trial testing the impact of a new tuberculosis diagnostic test on bacterial confirmation of a tuberculosis diagnosis.

9.1 Introduction

Permutation tests are a commonly used non-parametric statistical technique, used to calculate p-values without making distributional assumptions. In individually randomised trials, they are used because they make no distributional assumptions, provide exact confidence intervals, and do not rely on large sample approximations [10]; the Stata command `permute` provides an intuitive and simple way to perform permutation tests in this simple scenario. While the benefits for permutation tests hold for more complex randomised designs, such as parallel and stepped-wedge cluster-randomised trials, `permute` cannot perform a valid test for such designs.

In a cluster-randomised trial, the allocation of clusters of individuals, such as villages or hospital wards, are randomised. In a parallel cluster-randomised trial (CRT), each cluster is randomised to receive either a control, or intervention condition for the duration of the trial, as in figure 9.2a. But, randomisation can be more complex than this. A stepped-wedge trial (SWT) is a cluster-randomised trial run over a number of periods. Clusters are randomised to sequences, where each sequence receives the control condition for a different number of periods, and then receive an intervention condition for the remaining periods of the trial, as in figure 9.2b. For both of these designs but particularly SWTs, it is difficult to assess the assumptions required by para-

metric methods and so permutations tests, which make fewer assumptions, are appealing [2, 3, 11, 12].

Figure 9.1: Schematics of cluster-randomised trial designs. White = time in control condition, Grey = time in intervention condition

Cluster
1
2 CONTROL
3
4
5
6 INTERVENTION
7
8

(a) a parallel cluster-randomised trial

		Period				
Sequence	Cluster	1	2	3	4	5
1	1	CONTROL				
	2					
2	3					
	4					
3	5					
	6					
4	7					
	8					INTERVENTION

(b) a stepped-wedge cluster-randomised trial

SWTs are commonly analysed with parametric models [13], but some have suggested that the intervention-effect estimate from such models may be biased if the correlation structure of the data is misspecified [14]. A “within-period” analysis strategy for SWTs has been suggested that makes fewer assumptions about this correlation structure [12]. First, the trial is separated into the periods between the times when clusters switch from control to intervention. Within each of these periods, an analysis is conducted to provide an intervention-effect estimate and weight for that period. The results are then combined using a weighted average of the within-period estimates, and permutation tests are used to calculate a p-value for evidence of an effect accounting for the correlation of observations in the same cluster.

Here we introduce a new command, `swpermute`. The new command allows specification of clustering and allocation to a time dependent sequence of intervention conditions to enable use with cluster-randomised trials with a particular focus on SWTs. An option is provided to conduct a within-period analysis, and we provide an option to test null hypothesis values other than zero to simplify the process of constructing confidence intervals.

9.2 Technical Details

The `swpermute` command is designed for trials with two treatment conditions; usually this will be a control and intervention condition so we will use this terminology throughout this paper. In this section, we will provide details of the permutation test, and how this is implemented in `swpermute`.

Permutation Tests: Individual Randomisation

In an individually randomised trial, we have some outcome observations y_k , $k = 1, \dots, K$. Half of these observations are collected under a control condition, the other half under an intervention condition, and we are interested in knowing whether the control and intervention conditions lead to different outcomes. As with most frequentist methods, we investigate this by constructing a distribution for the difference between the outcomes in control and intervention under the assumption that there is no difference. The true difference between control and intervention is the intervention effect $\theta = \theta_A$. If there is truly no difference between the two conditions, then assignment of observations to each condition is arbitrary, and for any set of assignments of the y_k to the control and intervention conditions, we can calculate an intervention effect $\theta = \hat{\theta}^*$. By repeating this process for each unique assignment of observations to conditions, we obtain the exact distribution of θ under the null hypothesis that $\theta = 0$. The observed intervention effect can then be compared to this distribution to calculate the p-value as the probability of an intervention effect the same as or more extreme than that observed:

$$p = f/F$$

where F is the number of unique permutations, and f is the number of these that produced an intervention effect the same or more extreme than the observed intervention effect.

It becomes unfeasible to calculate $\hat{\theta}^*$ for all unique permutations when the number of observations is large. In such cases, it is sufficient to randomly sample a number of permutations from all possible permutations, with or without replacement, a process known as Monte-Carlo permutations [10]. Both `permute` and `swpermute` perform Monte-Carlo permutations with replacement. The p-value is now given by:

$$p = f^*/F^*$$

where F^* is the number of permutations, and f^* is the number of these that produced an intervention effect the same as or more extreme than the observed intervention effect. Because Monte-Carlo permutations involve randomly selecting permutations, the p-value calculated may differ when the process is repeated with a different set of permutations.

With Monte-Carlo permutations, the number of permutations is chosen by the user and affects the accuracy with which a p-value can be computed. Good [15] recommends using between 100 and 1,600 permutations, increasing this number if the uncertainty around the p-value makes interpretation difficult. A confidence interval can be constructed for the p-value from a binomial distribution where the number of permutations is the number of Bernoulli trials, and the p-value is the probability of a success.

Two assumptions are required for permutation tests to be valid. Firstly, permutation tests test a strong null hypothesis, meaning they test if $\theta = 0$ for every observation rather than $\theta = 0$ on average. So, permutation tests are not valid where the effect of an intervention is expected to vary between observations. Secondly, permutations tests assume exchangeability of observations. This means that any assignment of observations to the conditions is equally likely. This second assumption does not hold for an SWT; in the next section we will discuss how to meet the assumption of exchangeability in these studies.

Extending Permutation Tests to Stepped-Wedge Trials

In the context of cluster-randomised trials, exchangeability holds at the unit of the cluster, but will not hold at the individual observation level. Because of this, the assignment of clusters must be permuted rather than assignment of individual observations. In SWTs, it is the assignment of clusters to a sequence that must be permuted to maintain the assumption of exchangeability. The `permute` command permutes at the level of the individual observations, so is not a valid test for SWTs. The `swpermute` command permutes clusters to sequences of allocations observed in the data and so is valid for SWTs.

Selecting an Intervention Effect Estimator

So far we have only discussed tests related to an intervention effect but we have not described an appropriate method for estimating an intervention effect. In this section, we will discuss the analyses currently recommended in the literature for SWTs.

A key design feature of all SWTs is that the intervention effect is confounded with time. This can be accounted for either by adjusting for period effects, or by conditioning on periods.

Adjusting for period

The most common analysis model used for SWTs, introduced by Hussey and Hughes [16], is an example of an analysis that adjusts for period effects:

$$y_{ijk} = \mu + \beta_j + \theta X_{ij} + u_i + e_{ijk} \quad (9.1)$$

where y_{ijk} is the outcome of individual k in period j from cluster i , μ is the mean outcome in the first period, β_j is the difference between period j and the first period with $\beta_1=0$, θ is the intervention effect, X_{ij} is 1 if cluster i received the intervention in period j and 0 otherwise, $u_i \sim N(0, \sigma_u^2)$ is a random effect for cluster, and $e_{ijk} \sim N(0, \sigma_e^2)$ is the within-cluster variability. This model can be extended to a generalised linear mixed model for outcomes that are not normally distributed.

Despite the popularity of this analysis model, it has been criticised for inflated type-one error rates if the period effects differ between the clusters. One solution that has been suggested is to use this analysis model to estimate an intervention effect but to use permutation tests to calculate a p-value and confidence intervals [2–4].

Other analysis models that adjust for time can also be used with permutation tests.

Conditioning on period: a within-period analysis

Using model 3.1 to estimate an intervention effect requires that this model gives an unbiased estimate; Thompson *et al* [14] showed that this is not always the case.

As an alternative, SWTs can be analysed with what is known as a within-period analysis, sometimes known as a vertical analysis because of how the trial schematics of these designs are drawn as shown in figure 9.2b. The general process is as follows:

1. An intervention effect is estimated for each period of the study using any analysis method that would be valid for a CRT.
2. These within-period estimates are given weights.
3. The within-period estimates are combined by taking a weighted average to get an overall intervention-effect estimate.

4. The permutation test is used to perform hypothesis tests on the overall intervention-effect estimate.

This procedure conditions on period, and so removes any confounding between the intervention effect and the period effects without making assumptions about the period effect or about the correlation of observations between periods. Any analysis that can be used for a CRT could be used within each period, for example Thompson and Davey *et al* [12] suggested using a cluster-level analysis in each period allowing for a simple, intuitive calculation of an intervention-effect estimate, but an appropriate individual-level analysis could also be used and either type of analyses could be extended to adjust for covariates.

The overall intervention effect estimate variability can be reduced by using appropriate weights for each period [11]. This is an area of ongoing research. Matthews and Forbes assume that the variance of the observations does not change over time and so suggest weighting periods by the imbalance in the number of clusters in the control and intervention [17]. Others have suggested that the weights given to each period should reflect the uncertainty around the within-period estimates [12], so that within-period estimates that are estimated with greater precision are given greater weight. The estimate's precision will depend on the variability of the observations and on the imbalance in the number of clusters in the control and intervention conditions in a particular period. In periods with more variability between the observations, the intervention effect estimate will have lower precision than periods with less variability between observations. Periods with a larger imbalance in the number of clusters in control and intervention will have an intervention effect estimate with lower precision than periods with an equal number in each condition. The estimated variance of the intervention effect estimate will incorporate both of these factors but, since it is only an estimate of the variance, can introduce additional variability into the overall intervention effect estimate. Therefore, if the total variability of the observations is not expected to change between periods, it may be more efficient to weight the periods only by the imbalance between the number of clusters in each condition.

We introduce a within-period option for `swpermute` to perform this procedure, thus making such an analysis easier to perform.

Constructing Confidence Intervals

Confidence intervals are created by finding the set of values for the true intervention effect that are not rejected at the α level, i.e., hypothesised intervention effects for which $p > \alpha$. One way to identify this set of values is to test many values. Values with $p \geq 0.05$ fall within the 95% confidence interval and $p < 0.05$ fall outside this interval.

If the effect of the intervention was to increase the outcome by θ_A , then after subtracting θ_A from the observations collected under the intervention condition, there should be no difference between the control and intervention conditions. Therefore, a null hypothesis of $\theta = \theta_A$ is tested by first subtracting θ_A from observations collected in the intervention condition, then running the permutation test as described above to get a p-value.

This process is simple for the case of an absolute difference in a continuous outcome. To test a relative difference, take the log of the observations before subtracting $\log(\theta_A)$. Similarly, the process is simple if we have summarised the outcome for each cluster-period: we might calculate the risk, log risk, odds, log odds, rate, or log rate for each cluster-period and can subtract the null value from these summaries.

However, when the outcome is on a different scale to the intervention effect, the outcome observations must be transformed before subtracting the null value. Take as an example, testing a null hypothesis of an odds ratio for an individual-level binary outcome. We start with the outcomes on the binary scale, either a 0 or 1 for each observation. First, we would need to group these to the binomial scale in the format d cases out of D observations in each cluster-period. From this, the log odds of the outcome in each group can be calculated and the null log odds ratio subtracted from these log odds. This would then need to be back-transformed to a number of d^* cases out of D observations.

9.3 The `swpermute` Command

The permutation test for SWT, described in section 9.2, is implemented in `swpermute`. `swpermute` runs a permutation test for SWTs on any analysis specified by the user. The algorithm identifies sequences in the observed data and permutes clusters between these sequences for each permutation. The

specified analysis can be run either across all periods in the study or within each period with results combined as a weighted average. Users can specify several null hypothesis values to be tested in order to construct confidence intervals if they have cluster-level or continuous outcomes.

Data Requirements

The `swpermute` command requires specification of a clustering variable identified in `cluster()`, a period variable identified in `period()`, and an intervention variable identified in `intervention()`. These three variables define the design of the study; each cluster is assigned to an intervention status in each period. The data should be in long format, with observations in each period given in different rows of the dataset. All observations within a cluster must follow the same sequence.

Syntax

The syntax of the `swpermute` command is as follows:

```
swpermute exp , cluster(varname) period(varname)
           intervention(varname) [ reps(num) left|right
           strata(varlist) saving(filename, ...) null(numlist)
           outcome(varname)withinperiod weightperiod(weightperiod)
           nodots level(num) seed(num) ] : command
```

exp specifies the result to be collected from results stored by the execution of *command*. Examples are `r(mu_1)-r(mu_2)` the mean difference estimated by `ttest`, or `_b[varname]` a coefficient estimate from a regression model .

`cluster(varname)` specifies the variable identifying the clusters. `cluster()` is required and must be a numeric variable. Observations with `cluster()` missing will be excluded from the analysis.

`period(varname)` specifies the variable identifying the periods. `period()` is required and must be a numeric variable. Observations with `period()` missing will be excluded from the analysis.

`intervention(varname)` specifies the variable identifying the intervention assignment. `intervention()` is required and must be a binary variable where 0 and 1 represent the control and intervention conditions, respectively. All observations within each value of `cluster()` must have the same value of `intervention()` in each `period()` or the command will return an error. If all values of `intervention()` are missing for a `cluster()` in a `period()`, this is assumed to be part of the sequence, for example as a washout period, and the missing value will be permuted. Otherwise, observations with `intervention()` missing will be excluded from the analysis.

`reps(num)` specifies the number of permutations to perform. The default is `reps(500)`.

`left|right` requests that one-sided p-values be computed. If `left` is specified, the p-value reported is the proportion of permutations where *exp* gives a value less than or equal to the observed value. If `right` is specified, the p-value reported is the proportion of permutations where *exp* gives a value greater than or equal to the observed value. The default is two-sided p-values, where the p-value reported is the proportion of permutations where *exp* is the same or further from zero than the observed value.

`strata(varlist)` specifies that the permutations be performed within each stratum defined by the values of *varlist*. This option should be used if randomisation of clusters was stratified [10].

`saving(filename, ...)` creates a Stata file (.dta file) consisting of a row for each permutation for each value in `null()`. The file consists of three variable containing the `null()` value being tested, the observed value of *exp* for that null value, and values of *exp* for each permutation. A new filename is required unless `replace` is specified. The option `double` specifies that results should be stored in *double* precision, the default is to store results as *float*. The option `every(num)` writes results to file every *num* permutations. This will allow recovery of partial results should the command not complete running.

`null(numlist)` specifies a list of values to test as the null hypo-

thesis. For each value specified, the value will be subtracted from the variable specified in `outcome()` if the variable defined in `intervention()` is equal to 1. The permutation test is run on this modified dataset to calculate a p-value. This option should only be used with cluster-level or continuous outcomes. The null values are assumed to be on the same scale as the outcome (e.g. risk differences if the outcomes are cluster-period risks). Ratios such as risk ratios, or odds ratios should be given on the log scale. The default is `null(0)`. When values other than the default are specified the option `outcome(varname)` is required.

`outcome(varname)` specifies the variable identifying the outcome. This option is only required when `null(numlist)` is specified with `numlist!= 0`. `outcome()` is assumed to be on the same scale as the values specified in `null()`. For example, `outcome()` should contain risks if `null()` gives risk differences, or log risks if `null()` gives log risk ratios.

`withinperiod` specifies that a within-period analysis should be performed. `command` is run within each unique value of the variable specified in `period()` and the resulting values of `exp` are combined as a weighted average using the weights specified in `weightperiod()`.

`weightperiod(weightperiod)` specifies the weights to be used if `withinperiod` is specified. This option is only required when `withinperiod` is specified and is one of the following:

`weightperiod(none)`: each period is given equal weight, so the weight $w_j = 1$ for all periods j .

`weightperiod(N)`: periods are weighted by the number of clusters in the control and intervention conditions as:

$$w_j = \left(\frac{1}{c_{0j}} + \frac{1}{c_{1j}} \right)^{-1}$$

where c_{0j} and c_{1j} are the number of clusters in the control condition and intervention condition

respectively in period j . It is equivalent to weighting by the variance if the variance of the outcome does not change over time.

`weightperiod(variance exp_2)`: each period is weighted by the inverse of the statistic exp_2 stored by the execution of *command*. That is

$$w_j = \left(\frac{1}{exp_{2j}} \right)$$

exp_2 is assumed to be the variance of each within-period estimate and will be different to the exp defined earlier in the command. This specification is suggested by Thompson and Davey *et al* [12].

`nodots` suppresses display of the dots at the completion of each permutation. By default, one `.` is displayed for each successful permutation. A red `x` is displayed if *command* returns an error or if the statistic in exp is missing for a permutation.

`level(num)` specifies the confidence level, as a percentage, for confidence intervals of the p-value. The default is `level(95)` or as set by `set level`; see [u] **20.7 Specifying the width of confidence intervals**.

`seed(num)` sets the random-number seed. Specifying this option is equivalent to typing:

```
set seed num
```

prior to calling `swpermute`. If no seed is specified, `swpermute` will return different results each time it is run due to the random selection of permutations.

Stored Results

`swpermute` stores the following in `r()`:

Scalars

<code>r(N_cluster)</code>	number of clusters
<code>r(N_strata)</code>	number of strata if strata option has been used
<code>r(obs_value)</code>	value of <i>exp</i> observed in the data
<code>r(N_reps)</code>	number of permutations

Matrices

<code>r(design)</code>	a matrix of 0 and 1 values showing the design of the SWT
<code>r(obs_period)</code>	value of <i>exp</i> observed in the original data within each period if a within-period analysis is specified
<code>r(p)</code>	p-values with their confidence intervals for each null value

9.4 The Dialog Box

The `swpermute` command can be used both as a coded command, and through a drop-down dialog box. To install the dialog box run the follow commands:

```
. window menu append submenu "stUser" "&Cluster RCTs"  
. window menu append item "Cluster RCTs" "Permute for stepped-wedge  
  trials (&swpermute)" "db swpermute"  
. window menu refresh
```

Running these commands from within Stata will only install the dialog box for the current session of Stata. To install the menus permanently, place the above commands into your profile.do file. See [u] [GSW] B.3, [u] [GSM] B.1, or [u] [GSWU] B.1 **Executing commands every time Stata is started** for more details on how to do this.

The supporting information S1 shows the dialog box for this command.

9.5 Example

To demonstrate the use of `swpermute` we will use data from an SWT conducted in Brazil that assessed whether switching to a new tuberculosis (TB) diagnostic test increased the proportion of patients with a bacterially confirmed TB diagnosis [18]. The real data could not be shared with the command. Instead, a simulated dataset is included that closely mimics the characteristics of this trial data.

The standard diagnostic test for TB, sputum smear microscopy, is quick and cheap but has low sensitivity [19]. Because of this, many patients are diagnosed with TB based on clinical symptoms alone, even if their test comes back negative [19]. A new TB diagnostic test (Xpert MTB/RIF) is known to be more sensitive than the standard smear microscopy method of diagnosis. It also provides a result for rifampicin drug resistance at the time of diagnosis [20].

This SWT in Brazil sought to explore the impact of switching from smear microscopy to the Xpert test. Here, we focus on a secondary outcome of the trial: whether patients had their TB diagnosis bacterially confirmed by either a smear microscopy test or the Xpert test [18].

The trial included 14 laboratories that covered most diagnoses in the cities of Rio de Janeiro and Manaus in Brazil. At initiation of the study, all laboratories were using sputum smear microscopy to diagnose TB. Following a month of baseline data collection, the Xpert test was rolled out to two randomly assigned laboratories each month, so that 7 months later, all laboratories were using the Xpert test.

Our dataset contains 3,924 patients diagnosed with TB during the study in the 14 laboratories; their type of diagnosis was recorded as either clinical (with a negative test or no test done), or bacterially confirmed. 2,147 (55%) patients were diagnosed with the Xpert test, and 2,833 (72%) had a confirmed TB diagnosis.

The output below describes the dataset we will use.

```

. use $DataDir\TBdiagnostic, clear
.
. describe
Contains data from T:\XpertMTBRIF\Mar2017\TBdiagnostic.dta
  obs:      3,924
  vars:      4                               18 May 2017 14:06
  size:     15,696

```

variable name	storage type	display format	value label	variable label
lab	byte	%8.0g		Laboratory
study_month	byte	%8.0g		Study period
arm	byte	%8.0g	lbl_arm	Smear microscopy (0) or Xpert (1)
confirmed	byte	%9.0g	lbl_conf	Clinical (0) or bacterially confirmed (1)

```

Sorted by: lab study_month
.
. list in 1/5

```

	lab	study_-h	arm	confirmed
1.	1	1	Smear	Confirmed
2.	1	1	Smear	Confirmed
3.	1	1	Smear	Clinical
4.	1	1	Smear	Confirmed
5.	1	1	Smear	Clinical

Each row gives the diagnosis type, `confirmed`, of a patient. `lab` identifies which laboratory they were diagnosed in and so assigns the patient to a cluster. `study_month` identifies which month of the study they were diagnosed in, and `arm` identifies whether the laboratory was using smear microscopy or the Xpert diagnostic at the time of diagnosis. The Xpert diagnostic was rolled out to laboratories as follows, where 0 assigns the laboratory to using smear microscopy and 1 assigns the laboratory to the Xpert diagnostic:

```

. table lab study_month , c(max arm)

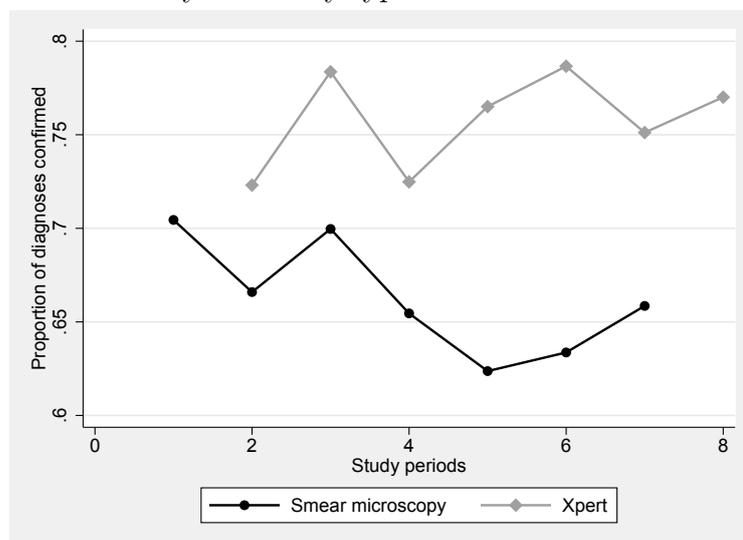
```

Laboratory	Study periods							
	1	2	3	4	5	6	7	8
1	0	1	1	1	1	1	1	1
2	0	0	0	1	1	1	1	1
3	0	0	0	0	1	1	1	1
4	0	0	0	0	1	1	1	1
5	0	0	0	0	0	1	1	1
6	0	0	0	0	0	0	0	1
7	0	0	1	1	1	1	1	1
8	0	0	0	0	0	0	1	1
9	0	0	0	0	0	1	1	1
10	0	0	0	0	0	0	1	1
11	0	0	0	0	0	0	0	1
12	0	0	0	1	1	1	1	1
13	0	1	1	1	1	1	1	1
14	0	0	1	1	1	1	1	1

Figure 9.2 shows the proportion of patients with a confirmed diagnosis in each study month by whether they were diagnosed with a smear microscopy test, or an Xpert test.

We will explore four analyses with permutation tests; the first will use the analysis model 9.2 with a permutation test, the last three analyses will demonstrate different within-period analyses.

Figure 9.2: Proportion of patients in the TB diagnostic trial with a confirmed diagnosis in each study month by type of test used



Adjusting for Period

Analysing the data using model 9.2 gives the following results:

```
. melogit confirmed i.study_month arm || lab : , nolog
Mixed-effects logistic regression      Number of obs   =   3,924
Group variable: lab                   Number of groups =    14
                                      Obs per group:
                                      min =         92
                                      avg =       280.3
                                      max =         559

Integration method: mvaghermite      Integration pts. =     7
Log likelihood = -2270.7069           Wald chi2(8)    =   27.02
                                      Prob > chi2     =   0.0007
```

confirmed	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
study_month					
2	-.1908554	.1400958	-1.36	0.173	-.4654381 .0837273
3	-.0421501	.1525528	-0.28	0.782	-.3411482 .256848
4	-.3054589	.1540779	-1.98	0.047	-.6074459 -.0034718
5	-.2372221	.1587384	-1.49	0.135	-.5483436 .0738994
6	-.0743544	.1729088	-0.43	0.667	-.4132494 .2645406
7	-.2004482	.1844925	-1.09	0.277	-.562047 .1611505
8	-.1249807	.1947377	-0.64	0.521	-.5066596 .2566983
arm	.4155579	.1235129	3.36	0.001	.1734771 .6576386
_cons	.8810099	.1506117	5.85	0.000	.5858163 1.176203
lab					
var(_cons)	.1636317	.0759324			.065898 .4063148

LR test vs. logistic model: chibar2(01) = 51.72 Prob >= chibar2 = 0.0000

There is strong evidence that the use of the Xpert test increased the odds that a patient's TB diagnosis would be confirmed (odds ratio (OR)=1.52 96% CI

1.19, 1.93; $p=0.001$)

This analysis can also be run using the above model to calculate the intervention effect, but using permutation tests to calculate the p-value. Stata stores all regression coefficients in the system variables `_b[*]`, and stores the standard errors of these coefficients in `_se[*]`. In this example, we want to use the model coefficient for the variable `arm` in the permutation test, and so we set the `exp` to `_b[arm]` in the `swpermute` command:

```
. swpermute _b[arm], cluster(lab) period(study_month) intervention(arm) /*
>          */ reps(1000) seed(20255) saving(example1_results, replace) nodots: /*
>          */ melogit confirmed i.study_month arm || lab :
```

(note: file example1_results.dta not found)

Monte Carlo permutation results

```
command: melogit confirmed i.study_month arm || lab :
statistic: _b[arm]
design:
```

freq	1	2	3	4	5	6	7	8
2	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	1	1
2	0	0	0	0	0	1	1	1
2	0	0	0	0	1	1	1	1
2	0	0	0	1	1	1	1	1
2	0	0	1	1	1	1	1	1
2	0	1	1	1	1	1	1	1

statistic	obs_value	null	c	n	p	[95% Conf. Interval]
<code>_b[arm]</code>	.4155579	0	2	1000	0.0020	.0002423 .0072058

Note: confidence interval is with respect to p
p-value is two-sided

`swpermute` first of all repeats `command` and `exp`. It then shows the design-pattern matrix. Each row represents a unique sequence of allocations observed within the data, each column represents a period. For each sequence and period, a 0 or 1 is shown representing the intervention condition of clusters in that sequence in that period. The left most column shows the number of clusters assigned to each sequence.

The table below this gives the results of the permutation test. Notice that the observed value of `_b[arm]`, shown in column 1 of the table, is identical to the value from the model run without the permutation tests; this is because the permutation test does not affect the estimate itself. The second column of this table shows the null hypothesis being tested, in this example we only test a null hypothesis of no difference. The third column gives the number of permutations with a value of `exp` the same or more extreme than the observed value and the fourth column gives the total number of permutations successfully completed. Here, we found that only 2/1000 permutation gave a result

the same or more extreme than that observed; this gives the p-value shown in column 5. The last 2 columns give a two-sided 95% confidence interval for the p-value. For this dataset, the permutation test gave a similar result to the parametric model.

In running this command, we specified that the results were saved to a Stata dataset. We can look at this dataset to see more details of the permutation test. A sample of the dataset is given below:

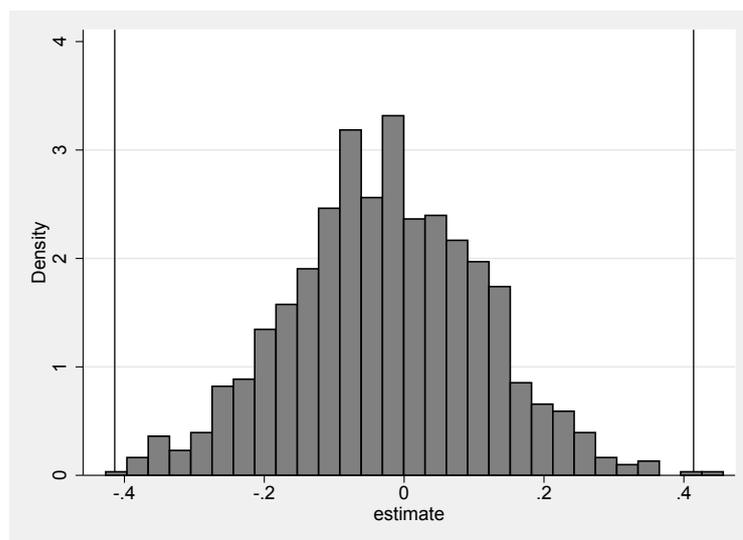
```
. use example1_results, clear
.
. list in 1/5
```

	null	observed	estimate
1.	0	.4155579	-.1569259
2.	0	.4155579	-.1907562
3.	0	.4155579	.0071681
4.	0	.4155579	-.0969895
5.	0	.4155579	.1421856

```
.
```

The estimates can be plotted in a histogram to show the distribution of *exp* under the null hypothesis, as shown in figure 9.3. The vertical lines show the observed values of *exp* and *-exp*. These are at the tails of the distribution, hence only 2 permutations had values of *exp* more extreme than that observed.

Figure 9.3: Distribution of the permutation test log odds ratio under the null hypothesis of no effect



Within-Period Analysis and Specifying Null Values

In this example, we will use a within-period analysis to calculate the difference in the risk (the proportion) of a confirmed diagnosis and show how to construct confidence intervals. In order to calculate a risk difference (the difference in proportions), we will use a cluster-level analysis within each period.

First, we calculate the proportion of confirmed diagnoses in each cluster-period. We run `swpermute` with `regress` as the *command* to calculate a risk difference and its variance. We select the `withinperiod` option to run the regression within each period, and set the period weights as variance weights. The `null()` option is specified with several null values using trial and error to identify the boundaries of the 95% confidence interval.

```
. swpermute _b[arm], cluster(lab) period(study_month) intervention(arm) /*
>      */ seed(9845) withinperiod weightperiod(variance _se[arm]^2) nodots /*
>      */ reps(5000) null(0 -0.004 -0.005 0.200 0.201) outcome(risk_confirmed): /*
>      */ regress risk_confirmed arm
```

Warning: study_month = 1 not included in analysis. Clusters all in one condition

Warning: study_month = 8 not included in analysis. Clusters all in one condition

Monte Carlo permutation results

```
command: regress risk_confirmed arm
statistic: _b[arm]
design:
      freq | 1  2  3  4  5  6  7  8
-----+-----
      2    | 0  0  0  0  0  0  0  1
      2    | 0  0  0  0  0  0  1  1
      2    | 0  0  0  0  0  1  1  1
      2    | 0  0  0  0  1  1  1  1
      2    | 0  0  0  1  1  1  1  1
      2    | 0  0  1  1  1  1  1  1
      2    | 0  1  1  1  1  1  1  1
```

Within period Estimates and Weights:

```
Estimate  Weight
study_month:2  0.0401  0.0581
study_month:3  0.0998  0.1787
study_month:4  0.0334  0.1329
study_month:5  0.1512  0.1835
study_month:6  0.1362  0.2714
study_month:7  0.0909  0.1753
```

statistic	obs_value	null	c	n	p [95% Conf. Interval]		
_b[arm]	.1052611	0	289	5000	0.0578	.0514913	.0646309
		-.004	250	5000	0.0500	.0441232	.0564089
		-.005	242	5000	0.0484	.0426166	.0547176
		.2	249	5000	0.0498	.0439348	.0561976
		.201	246	5000	0.0492	.0433697	.0555635

Note: confidence interval is with respect to p
p-value is two-sided

We see a warning that study month 1 and study month 8 are not included in this analysis; this is because all clusters are in the same condition during these periods so an intervention effect cannot be calculated. As well as the table of results we discussed in the previous example, we also see a list of effect

estimates and weights for each period in the study. Since we specified variance weights, the weights are the inverse of the variance of each effect estimate. Greatest weight is given to study month 6, despite the imbalance in clusters in the control and intervention conditions, because there was less variability in the cluster-level outcomes during this period.

The observed value in the table of results is the weighted average of these period-specific estimates. This method estimates that there is a 10.5% increase in the proportion of patients with a confirmed diagnosis using the Xpert test compared to using smear microscopy. The permutation test gives $p=0.06$ against a null hypothesis of no difference between the diagnostic tests.

As well as testing a null hypothesis of no difference, we have also tested several other values to construct a confidence interval. The boundaries for rejecting the null hypothesis at the 5% level are -0.4% and 20.0%, hence the 95% confidence interval is (-0.4%, 20.0%).

Within-Period Analysis Adjusting for a Cluster-Level Confounder

We saw in the previous example that the within-period analysis does not use observations from the first period of the study. We might want to incorporate this information by adjusting for the proportion of confirmed diagnoses in each cluster during this period.

The data must be reshaped so that the proportion in the first period is given as a variable, rather than as rows that contribute as outcomes.

```
. use $DataDir\TBdiagnostic, clear
. collapse (mean) confirmed , by(lab study_month arm)
. qui reshape wide confirmed arm, i(lab) j(study_month)
. rename confirmed1 baseconfirmed
. drop arm1
. qui reshape long
```

We then add this variable as a covariate in *command*. We have demonstrated this analysis through use of the dialog box in Supporting Information S1.

Inputting these setting runs the following command:

```
. swpermute _b[arm], cluster(lab) period(study_month) intervention(arm) /*
>      */ seed(9845) withinperiod weightperiod(variance _se[arm]^2) nodots /*
>      */ reps(1000): regress confirmed baseconfirmed arm
```

Warning: study_month = 8 not included in analysis. Clusters all in one condition

Monte Carlo permutation results

```
command: regress confirmed baseconfirmed arm
statistic: _b[arm]
design:
```

freq	2	3	4	5	6	7	8
2	0	0	0	0	0	0	1
2	0	0	0	0	0	1	1
2	0	0	0	0	1	1	1
2	0	0	0	1	1	1	1
2	0	0	1	1	1	1	1
2	0	1	1	1	1	1	1
2	1	1	1	1	1	1	1

Within period Estimates and Weights:

	Estimate	Weight
study_month:2	0.0160	0.0230
study_month:3	0.0847	0.1025
study_month:4	0.0840	0.2250
study_month:5	0.1519	0.0772
study_month:6	0.1112	0.4425
study_month:7	0.1016	0.1299

statistic	obs_value	null	c	n	p	[95% Conf. Interval]
_b[arm]	.1020699	0	6	1000	0.0060	.002205 .0130134

Note: confidence interval is with respect to p
p-value is two-sided

The intervention effect is similar to when we did not adjust for baseline, but the evidence of an effect is now stronger.

9.6 Concluding Remarks

swpermute is an extension of the Stata command permute that permutes clusters between sequences and can perform within-period analyses. We also incorporated functionality to test non-zero null hypotheses to facilitate the construction of confidence intervals. Although this command has been designed for use with SWTs, it can also be used with other trial designs such as CRTs and crossover cluster-randomised trials.

The command does, however, have limitations. Testing non-zero null hypothesis values is only available for continuous outcomes and cluster-level analyses. For other outcome types, the process would involve manipulating the

dataset to a such degree that we felt that it was safer for the user to perform this themselves. Whilst we have incorporated stratification of randomisation by a list of variables, some randomisation strategies such as restricted randomisation cannot be captured in this way [21]. It is a limitation of permutation tests generally that confidence intervals have to be constructed by the user.

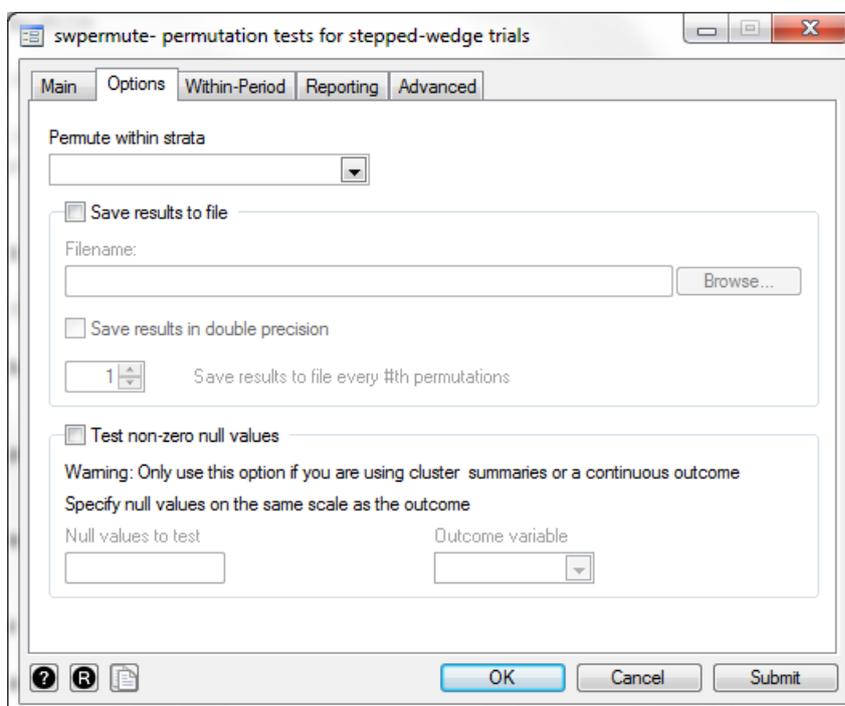
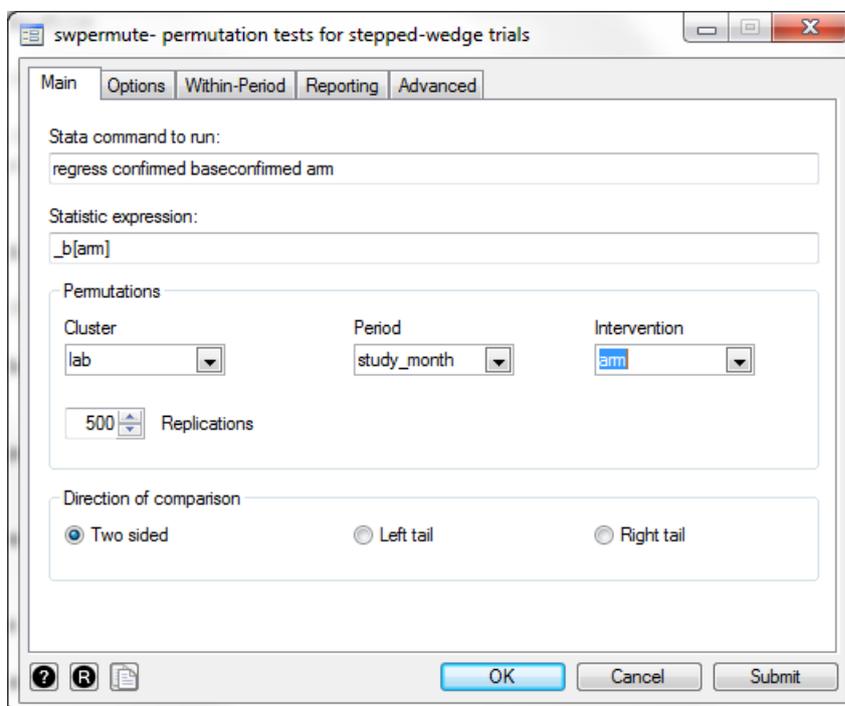
`swpermute` facilitates the use of robust analysis methods for an SWT, making complex analysis easier to perform.

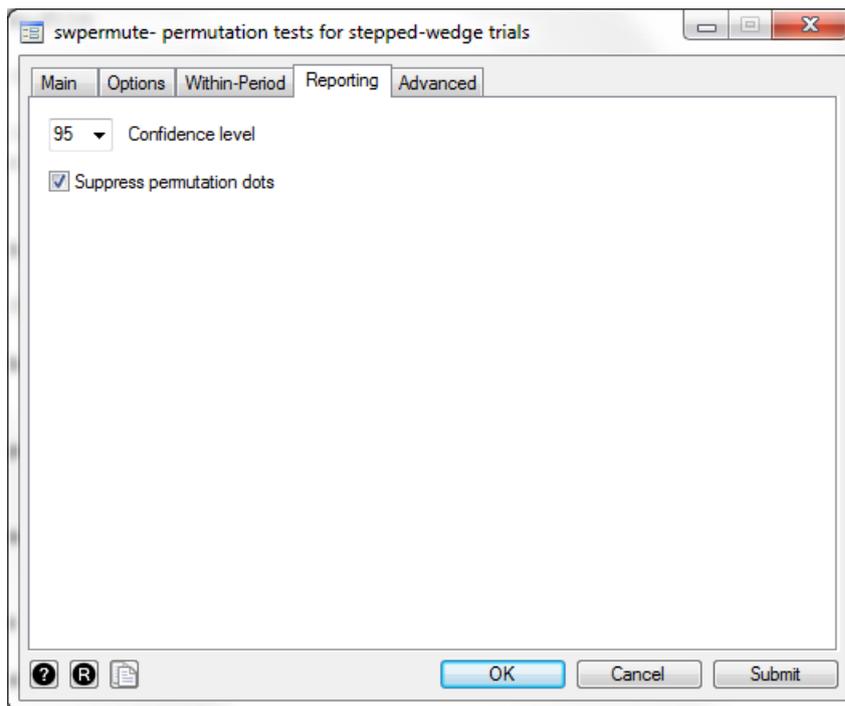
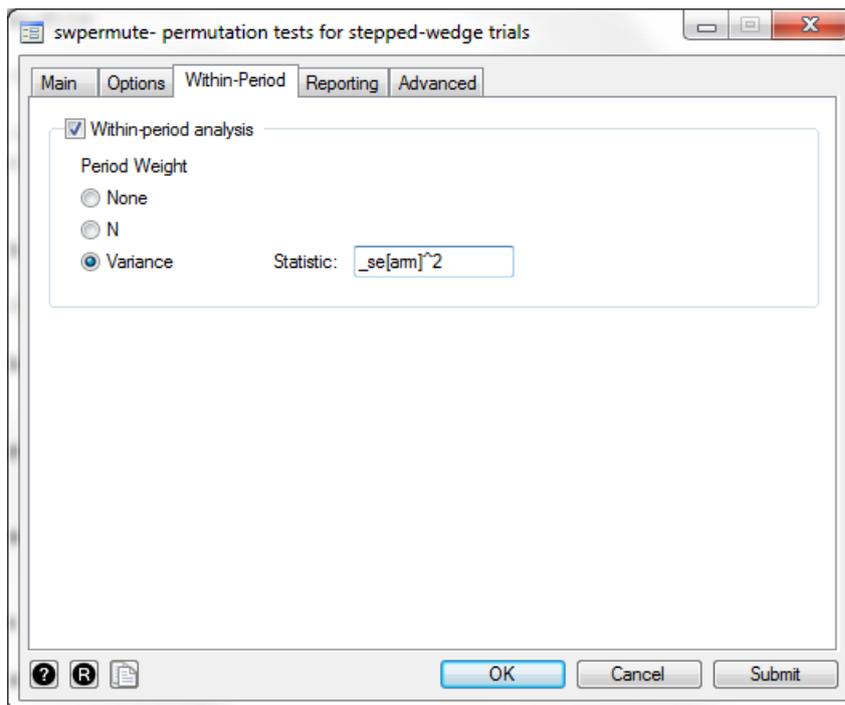
Acknowledgements

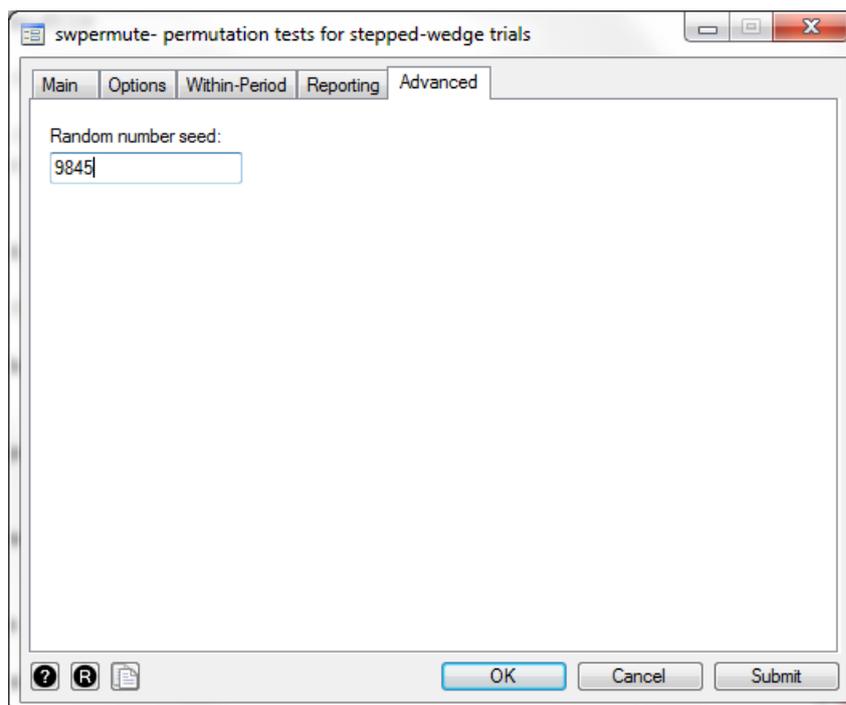
We would like to thank Professor Anete Trajman, Dr Betina Durovni, Dr Valeria Saraceni, Professor Frank Cobelens, and Dr Susan van den Hof for making available the original data from their study.

9.7 Supporting Information

S1: Screen Shots of Dialog Boxes that run Example 5.3







Bibliography

- [1] Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, Skea Z, Brehaut JC, Boruch RF, Eccles MP, Grimshaw JM, Weijer C, Zwarenstein M and Donner A. 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.
- [2] Ji X, Fink G, Robyn PJ and Small SS. Randomization inference for stepped-wedge cluster-randomised trials: An application to community-based health insurance. *Annals of Applied Statistics* 2017. 11. (1):1–20.
- [3] Wang R and DeGruttola V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. *Statistics in Medicine* 2017. 36. (18):2831–2843.
- [4] Bellan SE, Pulliam JR, Pearson CA, Champredon D, Fox SJ, Skrip L, Galvani AP, Gambhir M, Lopman BA, Porco TC, Meyers LA and Dushoff J. Statistical power and validity of Ebola vaccine trials in Sierra Leone: a simulation study of trial design and analysis. *The Lancet Infectious Disease* 2015. 15. (6):703–710.
- [5] StataCorp. Stata Statistical Software: Release 14. 2015.
- [6] Conference Paper. 2010.
- [7] Statistical software Component archive. URL: <https://ideas.repec.org/s/boc/bocode.html>.
- [8] Drukker DM. Programming an estimation command in stata a map to posted entries. URL: <http://blog.stata.com/2016/01/15/programming-an-estimation-command-in-stata-a-map-to-posted-entries/>.
- [9] Gould W. Statistical software certification. *Stata Journal* 2001. 1. (1):29–50.
- [10] Ernst MD. Permutation methods: A basis for exact inference. *Statistical Science* 2004. 19. (4):676–685.
- [11] Hayes RJ and Moulton LH. Cluster Randomised Trials. 1st ed. USA: Chapman and Hall/CRC, 2009.
- [12] Thompson JA, Davey C, Fielding K, Hargreaves J and Hayes RJ. Robust analysis of stepped wedge trials using cluster-level summaries within periods. Under Review.

- [13] Martin J, Taljaard M, Girling A and 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.
- [14] Thompson JA, Fielding K, C D, Aiken AM, Hargreaves J and Hayes RJ. Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. *Statistics in Medicine* 2017. 36. (23):3670–3682.
- [15] Good P. Permutation, Parametric, and Bootstrap Tests of Hypotheses. Springer, 2006.
- [16] Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials* 2007. 28. (2):182–191.
- [17] Matthews JNS and Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. *Statistics in Medicine* 2017:1–18.
- [18] Trajman A, Durovni B, Saraceni V, Menezes A, Cordeiro-Santos M, Cobelens F and Van den Hof S. Impact on Patients' Treatment Outcomes of XpertMTB/RIF Implementation for the Diagnosis of Tuberculosis: Follow-Up of a Stepped-Wedge Randomized Clinical Trial. *PloS One* 2015. 10. (4):e0123252.
- [19] Siddiqi K, Lambert ML and Walley J. Clinical diagnosis of smear-negative pulmonary tuberculosis in low-income countries: the current evidence. *Lancet Infectious Diseases* 2003. 3. (5):288–296.
- [20] Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC and Dendukuri N. Xpert(R) MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database of Systematic Reviews* 2014. (1):CD009593.
- [21] Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clinical Trials* 2004. 1. (3):297–305.

10 Discussion and Conclusions

Since the start of this PhD, the landscape of the literature on SWTs has evolved. Before I began, almost all literature on the design and analysis of SWTs focused on model 3.1, and in this setting I identified gaps in the literature on efficient SWT design and robust analysis.

In chapter 6, I identify a more efficient SWT design. I developed a new formulation of a design effect. This allowed me to algebraically explore the impact of the size of periods outside rollout and the number of sequences on the required sample size whilst holding all other parameters, such as the cluster size, ICC, and power, constant. In chapters 7, 8, and 9, I explore methods of robust analysis. This involved identifying problems with model 3.1, suggesting an alternative analysis method, and facilitating use of that method.

During the last three years, the analysis literature is now starting to place a greater emphasis on identifying alternative analyses to model 3.1. My work in chapters 7, 8, and 9 has contributed to this, along with the work of others notably Heuvel *et al* [1], Ji *et al* [2], and Wang and DeGruttola [3].

The sample size literature has also begun to move away from model 3.1, and now focuses on model 3.2. Whilst this is an improvement given my findings in chapter 7 that confidence intervals from model 3.1 sometimes have low coverage, model 3.2 is yet to be verified as an appropriate analysis model in typical SWT settings. Understanding of efficient SWT design has also improved over the past three years. The work in this thesis, along with that of Lawrie *et al* [4] and Girling and Hemming [5], has improved the understanding of how to increase the power of SWTs.

In this final chapter, I will summarise my findings and bring them together in the context of the current literature. Section 10.1 will give a discussion of the findings from the novel work of this thesis. Section 10.2 gives details of how I have disseminated these results and worked to increase their impact. Section 10.3 explores the strengths and limitations of my work, and section

10.4 provides ideas for the direction of future research to build on this work. Lastly, I will provide the overall conclusions from the thesis in section 10.5.

10.1 Synthesis of Findings

In this section I will give the main findings of each paper and discuss how they relate to one another and to the wider literature. These are important findings for the SWT literature and will have practical implications for researchers designing SWTs in the future.

10.1.1 Summary of Results

In chapter 6, I found that, in a complete SWT with an equal allocation of clusters to sequences, it is inefficient to collect observations outside the rollout period. With this modified design, the optimal number of sequences increases as the cluster-mean correlation increases. When a non-optimal number of sequences is used, there may be some efficiency gains from including some observations outside rollout, and I provide a formula for the size of the period outside rollout.

The findings of this chapter complement the recent results of Girling and Hemming [5] who found that the sample size could be reduced by allowing a sequence of clusters to remain in the control condition and a sequence to remain in the intervention condition for the duration of the trial. However, my findings contradict the results of several other papers that found that the statistical power increased with the number of sequences. A more in depth discussion of this contradiction is given later in this section.

In chapter 7, I conducted a large simulation study to assess how robust model 3.1 is to misspecification of the random effects. I found that, for all types of misspecification I considered, model 3.1 (the standard-model) produced confidence intervals with low coverage and, when the intervention effect varied between the clusters gave biased intervention-effect estimates. I identified that the random-period model (model 3.4) had better properties, but still produced confidence intervals with low coverage when the intervention effect varied between the clusters.

The bias in the intervention effect estimates have not been observed in the context of SWTs prior to this work, but have been observed for longitudinal data

with binary outcomes [6]. With a binary outcome and logit link, the relationship between the cluster-specific and marginal effect estimates are dependent on the correlation structure of the data. When the correlation structure of the data is misspecified, this relationship is misspecified and can lead to bias in the cluster-specific estimates [6]. For linear mixed effect models, correct specification of the correlation structure is required for estimation of the standard errors, but does not affect the consistency of the parameter estimates [6].

This paper also explored the weight each model gave to horizontal comparisons. I found that, compared to the Hussey and Hughes model (model 3.1), the random-period model and random-intervention model gave less weight to the horizontal comparisons when the period or intervention effects varied between clusters. This partly explains why the Hussey and Hughes model is so sensitive to misspecified random effects.

This chapter has confirmed that model 3.1 is prone to confidence intervals with low coverage [2, 3, 7]. The investigation of the impact of the intervention effect varying between the clusters and the resulting identification of potential bias of the intervention-effect estimates are novel. The use of the random-period and random-intervention models are also novel in the context of SWTs. My finding that model 3.1 gives a large weight to horizontal comparisons is expected, but this is the first work that explicitly shows this to be the case; this finding has since been confirmed by others [8].

To address the need for a more robust analysis that was demonstrated in chapter 7, chapter 8 introduces a novel analysis method that makes very few assumptions. Each period of the study is analysed separately to calculate a period-specific intervention-effect estimate and a variance for that estimate. The estimates are then combined using an inverse-variance weighted average and permutation tests are used to calculate a p-value and confidence interval. I used a cluster-summary analysis within each period as this is most widely applicable, but any analysis valid for a CRT, such as a random-effect model on individual level data could potentially be used here. A cluster-summary analysis has the benefit that the analysis is reliable even with a small number of clusters, and, for binary outcomes, risk differences can be calculated with ease [9]. This is a simple, transparent method of analysis that makes only two assumptions about the data: clusters are exchangeable between sequences and the intervention effect is common to all clusters. However, this comes at the cost of reduced power compared to methods that make stronger assumptions. The cluster-summary, within-period analysis performed well in my simulation

study, regardless of the form of the period-effect between-cluster variability.

Similar methods have been suggested in the analysis of longitudinal data [10, 11], but these are likely to require many clusters per sequence to reliably estimate a covariance matrix for the within-period intervention effect estimates. A within-period method has also been suggested for SWTs [12], but this method made strong assumptions about the correlation of observations between clusters. The flexible method I have provided is novel in the context of SWTs. The reduction in power was expected and has been noted by others [11, 12].

In chapter 9, I describe a Stata command that I developed. The command has the primary purpose of facilitating the use of the within-period analysis method described in chapter 8. The command performs permutation tests for SWTs by permuting clusters between sequences. It allows specification of any chosen analysis method for estimating the intervention effect, using cluster summaries or individual level data, to be run across all periods of the trial [2, 3, 7], or within-periods as I described in chapter 8. Functionality is included to allow users to test different null hypotheses for the intervention effect in order to construct confidence intervals.

With the use of this command, the within-period analysis becomes almost as simple to execute as a mixed-effect model, removing a barrier to the use of this robust analysis method. Prior to the development of this command, the only way to conduct a valid permutation test for an SWT was to manually code the tests oneself. The inbuilt `permute` command in Stata can only permute individual observations between intervention conditions. In Chapter 8, R code is also provided for conducting a within-period analysis and using permutation tests.

10.1.2 Horizontal and Vertical Comparisons

Chapters 6-8 all contribute to understanding the use of vertical and horizontal comparisons in SWT analyses.

In chapters 7 and 8, I explored the weight given to vertical and horizontal comparisons and demonstrate a dependence on the cluster-mean correlation, which has recently been confirmed by Matthews and Forbes [8].

In chapter 7, the example I used had a large cluster-mean correlation because the clusters had a geometric-mean size of 1000 observations; this led to

a cluster-mean correlation of 0.98 for the scenario with lower ICC=0.05 and common period and intervention effects. In this high cluster-mean correlation scenario with common period effects, model 3.1 retrieved almost all the intervention effect information from horizontal comparisons.

In chapter 8, I also found that there was a large difference between the power of the vertical analysis and model 3.1 in scenarios with common period effects, confirming that model 3.1 obtained a large amount of information from horizontal comparisons. Again, this paper used scenarios with a high cluster-mean correlation because of the large cluster sizes (median 200 observations per cluster); the low ICC scenario had a cluster-mean correlation of 0.80, while the high ICC scenario had a cluster-mean correlation of 0.95. Even with this small increase in cluster-mean correlation, there was a clear increase in the difference in power between the within-period analysis and model 3.1. This demonstrates that the use of horizontal comparisons depends on the ICC and cluster size.

This explains why in chapter 6 I found that there was an optimal number of sequences. The optimal SWT design must balance the precision of the vertical and horizontal comparisons. In a design with all observations within rollout, increasing the number of sequences increases the imbalance between the conditions in each period and so reduces the precision of vertical comparisons [9, 13]. Conversely, increasing the number of sequences improves data usage for horizontal comparisons so increases the precision of horizontal comparisons, despite increasing the number of period effects estimated by the model.

Previous research found that increasing the number of sequences increased the power of the standard design [5, 14, 15]. With such a design, the proportion of observations in the periods outside rollout is determined by the number of sequences. Therefore, the horizontal precision continues to increase with sequences, and now the vertical precision increases as well because a larger proportion of the observations can contribute to these comparisons.

Adding periods outside of the rollout period and increasing the number of sequences both have the potential to increase the precision of the horizontal comparisons when the cluster-mean correlation is high. My finding that it is optimal to have no observations outside rollout suggests that increasing the number of sequences has a greater benefit to the precision of the horizontal comparison than introducing periods outside rollout.

This thesis has clarified some of the roles horizontal comparisons play in SWTs.

They increase precision, but the high dependence on these comparison when the cluster-mean correlation is high can be the cause of bias and under coverage of confidence intervals if the incorrect assumptions are made in the analysis.

10.1.3 Utility of the Hussey and Hughes Model

Chapters 7 and 8 both showed that model 3.1 has poor confidence interval coverage in a range of scenarios. This brings the use of this analysis model into question without a strong justification for why it is appropriate. The finding of under coverage when the period effect varies between the clusters confirms previous research [2, 3, 7], and is consistent with findings from similar trial designs such as CRXO and CRT with baseline observations [16, 17]. In chapter 8, I show that the degree under coverage is dependent on the strength of clustering in the data in line with findings by Wang and DeGruttola [3]. I demonstrate the novel finding that the under coverage depends on the number of sequences. Under coverage of confidence intervals also occurred in scenarios with only 9 clusters, confirming that mixed effect models are unreliable with a small number of clusters [9]. In chapter 7, I also showed that model 3.1 produces biased intervention-effect estimates when the intervention effect varies between clusters. Together, these findings show that model 3.1 is not appropriate for analysing SWTs in a large range of settings.

This confirms that the results of chapter 6 will not be valid when model 3.1 is misspecified, as suggested by Hooper *et al* [18]. In particular, the design effect I derived will underestimate the sample size required to achieve a given power if the period effects vary between clusters. However, it seems likely that removing periods before and after rollout will continue to improve efficiency when this assumption is relaxed, such as in the random-period model, because less information is obtained from horizontal comparisons. This is supported by Girling and Hemming [5] who found that the Hybrid design, with some clusters in the control or intervention conditions for the duration of the trial, is more efficient than a standard SWT under the assumptions of the less restrictive model 3.2. Likewise, although the formula for the number of sequences will not hold exactly, it is likely that the general relationship between the cluster-mean correlation and the number of sequences will be similar.

10.1.4 Selecting an Analysis

After identifying that model 3.1 is not a robust analysis method, I have identified two potential alternatives: the random-period model, and the cluster-summary method. The choice between these methods will depend on the setting in which they are being used.

The random-period model will have more power than a within-period analysis. However, the model may have problems converging if the period effects are similar in all clusters and this model suffered low confidence interval coverage when the intervention effect varied between clusters. The model also grows in complexity as the number of sequences increases.

Alternatively, the cluster-summary method is transparent, robust, and allows for simple calculation of the risks difference. However, this comes at the cost of power.

Other alternative analysis models that I have not considered were detailed in chapter 3.

These include GEE, which have been shown to be robust to misspecification of correlation structures in other settings [19], and model 3.2. These are methods which require further research.

Some of my results bring the utility of other alternative methods into question. In chapter 7, I found that model 3.1 produced biased intervention-effect estimates when the intervention effect varied between clusters. This challenges the validity of the analysis suggested by Bellan *et al* [7], Wang and DeGruttola [3] and Ji *et al* [2] in this scenario because it uses the estimate from model 3.1.

Greater use of sensitivity analyses would increase readers' confidence in SWT results and has been recommended [20]. The within-period analysis would be particularly useful as a sensitivity analysis because it makes so few assumptions.

10.1.5 Issues with Prespecifying an Analysis

The choice of analysis methods has implication from the planning stages of a trial and so must be prespecified, ideally in an analysis plan. Campbell and Walters give advice on writing an analysis plan [21]. This allows the design of the study to reflect the chosen type of analysis, and prevents trialist from running several analysis models and selecting one based on the results.

Some analysis methods have more power to detect an effect than others, so sample size calculations should reflect the analysis choice.

The cluster-summary method will require a larger sample size to achieve the same power as the random-period model. Moulton *et al* [13] provide advice on sample size calculation for a within-period analysis for time to event outcomes but until further research is conducted, simulations are required to calculate the sample size of a within-period analysis.

Others provide advice on sample size calculations for methods which incorporate the horizontal comparisons (and adjust for period effects) allowing for random-period effects [5, 18]. These methods require assumptions about period effects and changes to the correlation over time and there is little guidance available in the literature to guide these assumptions. A conservative approach is to assume lower than expected correlation between observations in different periods (auto-correlation) in these formula.

I also found that models with more complex variance structures, such as the random-period model and random-intervention model, failed to converge when the variance of the random effects were truly zero; this has been previously noted for mixed-effect models with binary outcomes [22]. An analysis plan should specify what action will be taken if this is a problem in the final analysis. In the discussion in chapter 7, I advised that a simpler model is prespecified in an analysis plan that can be used if the more complex model fails to converge. An alternative solution is to identify methods that improve convergence.

10.1.6 Implications for the Sample Size of Stepped-Wedge Trials

A consequence of using these more robust analysis methods is that SWTs will need to be larger to achieve the same power than previously thought.

If the cluster-summary method is planned, a CRT will have more power than an SWT with the same sample size because of the imbalance in the number of clusters in each condition in each period [13].

If a random-period model analysis is planned, the difference in sample size compared to a CRT may not be as large as the model 3.1 standard error suggests. This may have implications for trialists choice of design if sample size is a primary concern: the additional complications of the SWT design may outweigh a small reduction (or even increase) in sample size. However,

implementing my design alterations from chapter 6 may counteract some of the sample size increases required for the random-period model.

10.1.7 Implications for Reporting

This PhD has raised awareness about the complexity of designing and analysing SWTs. In addition to the considerations required by all RCTS and by CRTs, SWTs require that appropriate consideration has been given to period effects and correlations over time.

In particular, my research has highlighted the high risk of bias from SWT analysis. In order for readers to assess the validity of trial results, results must be reported in a way that allows readers to assess the appropriateness of analysis assumptions. This means that both the assumptions themselves and an assessment of those assumptions should be clearly reported. Improved reporting of the observed correlation structure of the data will also help to inform the planning of future trials in similar settings.

Several reviews have found such reporting to be lacking [23–26]. A lack of guidance may be responsible as there is currently no CONSORT guidelines available for SWTs. The development of guidelines are in progress [27], and I am a member of the expert panel involved in their development. By providing guidance on reporting before SWTs become commonly used, we hope trialists will foster good reporting practices. A downside to developing guidelines while the methodology literature is still developing is that guidelines may lag behind developments in the methodological literature.

10.2 Dissemination and Increasing Impact

Throughout this PhD, I have placed an emphasis of disseminating findings in a timely manner to increase their impact. Chapters 6 and 7 have been published [28, 29], chapter 8 is currently under review at *Statistics in Medicine*, and chapter 9 will be submitted to the *Stata* journal. The research in this thesis has also been presented at the following conferences and meetings.

In 2014, I presented the aims of this PhD at the London School of Hygiene & Tropical Medicine research-degree poster day. This received considerable interest and I was awarded best pre-upgrading poster in my faculty.

In 2015, at the International Conference for Clinical Trials Methodology in Glasgow, UK, I gave oral presentations of preliminary findings of chapter 6 and chapter 8. The abstracts from these presentations are published in *Trials* [30, 31].

In 2016, I presented the results of chapter 7 as a poster at the First International Conference on Stepped-Wedge Trial Design in York, UK; this won a prize for the best poster and the abstract is published in *Trials* [32]. In 2017, I presented this poster again at the School's research-degree poster day, and the poster won a prize as one of the best posters in the faculty.

Also in 2016, I presented the examples given in chapter 8 as a demonstration of the cluster-summary, within-period analysis at the Meeting on Current Developments in Cluster Randomised Trials and Stepped-Wedge Design in London, UK.

In 2017, I gave an oral presentation of the final results from chapter 6 as part of an invited session on SWT design at the International Conference for Clinical Trials Methodology in Liverpool, UK.

As well as presenting the findings of this thesis at conferences, I have also become involved in other projects that allow me to influence the future methods used for SWT designs. I am part of an expert panel involved in developing CONSORT guidelines for reporting SWTs. The development of the Stata command in chapter 9 was done with the purpose of improving the uptake of permutation tests and within-period analyses in SWTs. I am also involved in a project to create a website providing advice on the design and analysis of cluster randomised trials and am leading the sections on designing and analysing SWTs.

10.3 Strengths and Limitations

In this section, I will describe the strengths and limitations of the original work in this thesis.

The guidance I have provided is practical and simple for trialists to implement. This should improve the uptake of my suggestions leading to a greater impact. The design suggestions in chapter 6 are simple adaptations of the standard SWT design that have a minimal impact on the ease of rolling out the intervention. The adaptations discussed in the literature involve allowing

more clusters to be randomised to some sequences than to others [4, 5]: this could make rollout more difficult where resources for rollout are limited. I also provided a process for how to choose a trial design. This process acknowledges that the most efficient design might be infeasible to implement and guides trialists in choosing the most efficient, practical design given the restraints of their setting. However, this advice focuses on settings where the data is being collected for the purpose of the trial. In settings with routinely collected data, inclusion of baseline data may be possible at no additional cost and will then be beneficial to reducing the intervention effect variance.

Likewise, chapters 7 and 8 had a pragmatic focus. The simulation studies focused on realistic scenarios, and I have made practical recommendations for trial analysis. The Stata command described in chapter 9 was developed in a way to make it flexible but easy to use, and the help file and journal article give clear instructions. This command makes my suggested with-period analysis much simpler to conduct in Stata, and the R code supplied in Chapter 8 makes the analysis much simpler in R.

Another strength of the work in this thesis is the choice of methods.

The results of chapter 6 were derived algebraically, which means that they hold wherever the assumptions of the design effect holds. However, these are strong assumptions.

Chapters 7 and 8 used simulation studies following the advice of Burton *et al* [33]. The simulation studies used to assess the analysis methods covered a range of settings. Across both chapters, the scenarios covered different formulations of between-cluster variability. I studied ICCs ranging from 0.02 to 0.2, different forms of variability in the period effect, and variability in the intervention effect. Both simulation studies were motivated by real data from two very different settings. This, along with the findings of previous research, gives strength to my conclusions that model 3.1 is sensitive to deviations from the model assumptions across a range of settings.

In developing the Stata command described in chapter 9 I have performed rigorous testing and incorporated many checks into the code that the user has correctly specified the command.

However, there are limits to the generalisability of my findings.

Whilst I did consider a range of ICC values, these were all relatively high for health research where the median ICC is approximately 0.01 [34], and

the cluster sizes were relatively large (a recent review of SWTs found a median cluster size of 55 [25]) leading to high cluster-mean correlations in all scenarios. I have also made several simplifying assumptions. For example, throughout this thesis I have assumed that the intervention is constant over time, so that there is no lag or waning of the intervention effect. Assuming a constant effect when there is really a lag or waning of effect is likely to lead to underestimation of the intervention effect. This is a topic that requires further research.

Throughout, I have only considered cross-sectional, complete SWT designs. Although these designs are common in practice [26], my results are not generalisable beyond these designs.

There have also been issues reported with the distributional assumptions of outcomes. Chapter 6 assumes a normally distributed outcome, which has been shown to be inappropriate for binary outcomes [35]. Conversely, chapters 7 and 8 focus only on binary outcomes because of their common use in practice [25]. My simulation studies focused on estimation of log odds ratios because the logit link transforms probabilities onto the real line. However, interpretation of odds ratios is made difficult by non-collapsibility. This is an issue in chapter 7, where there are issues of comparability of effects from different analysis models. The log odds ratio calculated by a cluster summary analyses has been shown to be biased [36], although I did not find any such issues in the analyses presented in chapter 8. This discrepancy is because I compared the estimates to a cluster-specific effect, whereas Ukoumunne *et al* [36] used a marginal effect.

A limitation of chapter 6 is the number of assumptions and restraints placed on the designs that I considered. The results are only valid when model 3.1 is valid, which later chapters of this thesis show often is not the case. My design effect required an equal allocation of clusters to each sequence and the same number of observations in each cluster-period within rollout. Therefore, my optimised SWT is only optimal within these restraints.

A limitation of my conclusions in chapters 7 and 8 is that both of the analysis methods I have recommended make the assumption that the intervention effect is common to all clusters. Chapter 7 shows that the random-period model is sensitive to this assumption, but I did not explore how sensitive the within-period method will be. It is not known how common it is for intervention effects to vary in practice. In CRTs, permutation tests have been shown to be robust to intervention effect variability when there is the same number of clusters

in each condition, and are conservative when the assumption is violated [37]. Further research is needed to explore how this relates to the SWT design where there is overall balance between the number of observations in each condition but an imbalance in most periods.

10.4 Future Research Directions

The limitations of this work naturally leads on to topics for future research.

The later part of this thesis advises against the use of model 3.1. Instead, I have suggested the use of the random-period model or the cluster-summary, within-period analysis. Efficient design when using the random-period model requires further research. Whilst my findings about the use of vertical and horizontal comparisons in the random-period model suggest that the general results of removing the periods before and after rollout and selecting a number of sequences depending on the cluster-mean correlation are likely to hold, this needs to be confirmed and the specific relationship needs to be identified. The design effects provided by Hooper *et al* [18] and Girling and Hemming [5] may be able to shed light on this question assuming that the total variance does not change between period and that there is the same correlation between all cluster-periods. Alternatively, simulation studies could be used to explore this topic.

Further research is required to see if there are practically important gains in efficiency from removing the constraints placed on my optimisation. The hybrid design shows us that there are gains from allowing an imbalance in the allocations, but further gains may be possible through allowing all the periods to vary in length as I did for the periods before and after rollout. There may be efficiency gains from allowing more observations to be collected when there is an equal number of clusters in the control and intervention conditions compared to periods where there is a large imbalance. It should be possible to explore this concept algebraically using similar methods to those used in chapter 6. Zhan *et al* [38] found that such a design reduced the mean-square error in simulation studies that incorporated no clustering, but no other research has further explored this issue.

One of my important findings was that the intervention effect estimate from model 3.1 may be biased if the intervention effect varies between the clusters. This has not been shown previously and so needs confirming in a variety of

other scenarios including in data with a lower ICC, smaller clusters, and smaller variability in the intervention effect to demonstrate when bias may be an issue.

The random-period model is a promising parametric analysis for SWTs. The form that I suggested in chapter 7 is very flexible: it places no constraints on how the variability changes over time, or how the periods relate to one another. A disadvantage of this degree of flexibility is that the model becomes increasingly complex as the number of periods increases because each additional period in the study increases the number of variance and covariance parameters that are estimated by the model. For the majority of scenarios, a simpler covariance structure may suffice. The model 3.2 is a simpler model recommended for CRXO and CRT with baseline observations [16, 17]. It assumes that the total variability in the data remains the same throughout the trial and assumes that there is the same relationship between all periods. While this model performs well for CRXO and CRT with baseline observations, the larger number of periods in SWTs may make this model less suitable. Future work must consider if and when this simpler alternative will suffice, and when a more complex correlation structure is necessary.

Further work is also needed to aid in prespecification of analysis models, such as improving identification of appropriate correlation structures before a trial begins. In addition, improved convergence of models in the presence of low variability between clusters would allow trialists to prespecify a flexible analysis model when they are uncertain about the correlation structure of their data. Bayesian hierarchical models may be one method of achieving this [39].

There is little work to date on the number of clusters required for mixed-effect models to be reliable in the context of SWTs. In CRTs, it is recommended that there are at least 15 clusters per arm for mixed-effect models [9], but this may be different for an SWT. The model 3.4 is likely to require more clusters than model 3.1 to remain reliable because of the additional parameters being estimated. Simulation studies would be useful to explore the estimates and error rates of the random-period model when varying the cluster size, number of clusters, ICC, and number of sequences. The utility of small sample corrections have been shown to improve the performance of mixed effect models for CRTs with continuous outcomes [40], so should be explored for SWTs with different outcome types.

GEE are largely unexplored in SWTs. They are generally thought to give correct parameter estimates, even when the working correlation structure is

misspecified [41], and have been shown to give nominal coverage with as few as 10 clusters if small sample adjustments are applied [42]. However, they may produce biased estimates in models with time varying covariates [43], so further research is required to explore this analysis option.

There is also much development possible for the cluster-summary method. Permutation tests require that the intervention effect is common to all clusters, but I have not assessed how sensitive the tests are to this assumption in the context of SWTs. If permutation tests are found to be very sensitive, an alternative method, such as bootstraps may be required. Whilst bootstraps in SWTs have been unsuccessful to date [7], this is an area that warrants further investigation. Future work should also look at ways to improve the power of the within-period analysis, for example by using different weights or adjusting of covariates. The methods developed by Wei, Ware, Moulton and Stram [10, 11, 44] that utilise the covariance matrix for weights (described in section 3.2.7) is worth further consideration, particularly for trials with a large number of clusters. A benefit of this method over my within-period method is that confidence intervals and p-values can be calculated from a normal distribution.

There are also possible improvements to be made to the Stata command. Some SWTs use a constrained randomisation where the randomisation scheme is selected from the set of schemes that minimise the imbalance of important baseline covariates [45]. For such a trial to use permutation tests, the set of permutations would need to be restricted to the set of schemes considered. The Stata command cannot currently be used for such a design. A future version of the Stata command would incorporate such functionality.

10.5 Concluding Remarks

In this thesis, I have shown that the most commonly used design is inefficient. Simple analysis methods make strong, sometimes inappropriate assumptions about the data. Trialists should give careful thought to how their choice of design and analysis can affect the power of the study as well as the reliability of the conclusions.

Designing trials in an efficient way is important to reduce the cost of conducting a trial and I have shown that SWTs can be run more efficiently by removing the periods before and after rollout and carefully choosing the number of sequences. A conservative approach to sample size calculations would

be to assume that there is less correlation between observations in different periods than expected, as this will reduce the chance of under powering the study.

Using the correct analysis for a trial is clearly of vital importance; after the time and energy spent conducting a study, it would be disastrous to draw the wrong conclusions. Given the results of this thesis, I would discourage the use of model 3.1 for the analysis of SWTs. Instead, I have suggested two methods of analysis that will be appropriate in a wider range of settings because they both make fewer and more realistic assumptions about the data. I recommend that trialist use either the random-period model if power is of primary concern, or the within-period analysis with cluster summaries if there is a high degree of uncertainty about the correlation structure of the data, or few clusters.

The wide adoption of the findings from this thesis would help to overcome some of the disadvantages of the SWT design as described in chapter 1. When using an analysis method that utilises horizontal comparisons, removing or shortening the periods before and after rollout will lead to trials that require fewer observations, potentially shortening the trials and leading to more timely impact of trial results. Alternatively, using a within-period analysis with permutation tests will lead to intervention-effect estimates that are robust to misspecification. The results from such an analysis will accurately reflect the uncertainty around the intervention-effect estimates. Consequently, policy makers will be better informed about how much weight to place on a particular trial result.

Bibliography

- [1] Van den Heuvel ER, Zwanenburg RJ and 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* 2017. 26. (2):766–775.
- [2] Ji X, Fink G, Robyn PJ and Small SS. Randomization inference for stepped-wedge cluster-randomised trials: An application to community-based health insurance. *Annals of Applied Statistics* 2017. 11. (1):1–20.
- [3] Wang R and DeGruttola V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. *Statistics in Medicine* 2017. 36. (18):2831–2843.

- [4] Lawrie J, Carlin JB and Forbes AB. Optimal stepped wedge designs. *Statistics and Probability Letters* 2015. 99:210–214.
- [5] Girling AJ and Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Statistics in Medicine* 2016. 35. (13):2149–2166.
- [6] Heagerty PJ and Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* 2001. 88. (4):973–985.
- [7] Bellan SE, Pulliam JR, Pearson CA, Champredon D, Fox SJ, Skrip L, Galvani AP, Gambhir M, Lopman BA, Porco TC, Meyers LA and Dushoff J. Statistical power and validity of Ebola vaccine trials in Sierra Leone: a simulation study of trial design and analysis. *The Lancet Infectious Disease* 2015. 15. (6):703–710.
- [8] Matthews JNS and Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. *Statistics in Medicine* 2017:1–18.
- [9] Hayes RJ and Moulton LH. Cluster Randomised Trials. 1st ed. USA: Chapman and Hall/CRC, 2009.
- [10] Wei LJ and Johnson WE. Combining dependent tests with incomplete repeated measurements. *Biometrika* 1985. 72. (2):359–364.
- [11] Moulton LH and Zeger SL. Analyzing Repeated Measures on Generalized Linear-Models Via the Bootstrap. *Biometrics* 1989. 45. (2):381–394.
- [12] Granston T. Addressing Lagged Effects and Interval Censoring in the Stepped Wedge Design of Cluster Randomized Clinical Trials. Thesis. University of Washington, USA, 2014.
- [13] Moulton LH, Golub JE, Durovni B, Cavalcante SC, Pacheco AG, Saraceni V, King B and Chaisson RE. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clinical Trials* 2007. 4. (2):190–199.
- [14] Hussey MA. Cluster randomized crossover trials: design and analysis of the stepped wedge design. Thesis. University of Washington, 2005.
- [15] Woertman W, Hoop E de, Moerbeek M, Zuidema SU, Gerritsen DL and Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology* 2013. 66. (7):752–758.

- [16] Morgan KE, Forbes AB, Keogh RH, Jairath V and Kahan BC. Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome. *Statistics in Medicine* 2017. 36. (2):318–333.
- [17] Turner RM, White IR, Croudace T and Group PIPS. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Statistics in Medicine* 2007. 26. (2):274–289.
- [18] Hooper R, Teerenstra S, Hoop E de and Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine* 2016. 35. (26):4718–4728.
- [19] Turner EL, Prague M, Gallis JA, Li F and Murray DM. Review of Recent Methodological Developments in Group-Randomized Trials: Part 2-Analysis. *American Journal of Public Health* 2017:e1–e9.
- [20] Morris TP, Kahan BC and White IR. Choosing sensitivity analyses for randomised trials: principles. *BMC Medical Research Methodology* 2014. 14:11.
- [21] Campbell M and Walters S. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Wiley, 2014.
- [22] Omar RZ and Thompson SG. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Statistics in Medicine* 2000. 19. (19):2675–2688.
- [23] Mdege ND, Man MS, Taylor Nee Brown CA and 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):936–948.
- [24] Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ and Fielding KL. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials* 2015. 16. (1):358.
- [25] Martin J, Taljaard M, Girling A and 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.
- [26] Grayling MJ, Wason JM and Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. *Trials* 2017. 18. (1):33.

- [27] Hemming K, Girling A, Haines T and Lilford RJ. Protocol: Consort extension to stepped wedge cluster randomised controlled trial. Report. 2014.
- [28] Thompson JA, Fielding K, Hargreaves J and Copas A. The optimal design of stepped wedge trials with equal allocation to sequences and a comparison to other trial designs. *Clinical Trials* 2017.
- [29] Thompson JA, Fielding K, C D, Aiken AM, Hargreaves J and Hayes RJ. Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. *Statistics in Medicine* 2017. 36. (23):3670–3682.
- [30] Thompson J, Fielding K, Hargreaves J and Copas A. Considering the design effect for the stepped wedge trial: what can it tell us? *Trials* 2015. 16. (S2):O45.
- [31] Davey C and Thompson JA. Assessing sensitivity to assumptions in mixed effects analyses of stepped-wedge trials. *Trials* 2015. 16. (S2):O46.
- [32] Kanaan M. Proceedings of the First International Conference on Stepped Wedge Trial Design : York, UK, 10 March 2016. *Trials* 2016. 17. (1):311.
- [33] Burton A, Altman DG, Royston P and Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006. 25. (24):4279–4292.
- [34] Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S and Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology* 2004. 57. (8):785–794.
- [35] Baio G, Copas A, Ambler G, Hargreaves J, Beard E and Omar RZ. Sample size calculation for a stepped wedge trial. *Trials* 2015. 16. (1):354.
- [36] Ukoumunne OC, Forbes AB, Carlin JB and Gulliford MC. Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. *Statistics in Medicine* 2008. 27. (25):5143–5155.
- [37] Gail MH, Mark SD, Carroll RJ, Green SB and Pee D. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* 1996. 15. (11):1069–1092.
- [38] Zhan Z, Bock GH de, Wiggers T and Heuvel E van den. The analysis of terminal endpoint events in stepped wedge designs. *Statistics in Medicine* 2016. 35:4413–4426.

- [39] Turner RM, Omar RZ and Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine* 2001. 20. (3):453–472.
- [40] Kahan BC, Forbes G, Ali Y, Jairath V, Bremner S, Harhay MO, Hooper R, Wright N, Eldridge SM and Leyrat C. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials* 2016. 17. (1):438.
- [41] Liang KY and Zeger SL. Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* 1986. 73. (1):13–22.
- [42] Scott JM. Vaccine Efficacy Trials Using Stepped Wedge Design. Thesis. University of Washington, 2008.
- [43] Pepe MS and Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation* 1994. 23. (4):939–951.
- [44] Stram DO, Wei L and Ware JH. Analysis of Repeated Ordered Categorical Outcomes with Possibly Missing Observations and Time-Dependent Covariates. *Journal of the American Statistical Association* 1988. 83. (403):631–637.
- [45] Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clinical Trials* 2004. 1. (3):297–305.

Appendix

Appendix A: License Agreement for Paper A

Jennifer Thompson

From: PermissionsUK <PermissionsUK@sagepub.com>
Sent: 21 July 2017 16:51
To: John Devereux
Cc: Jennifer Thompson
Subject: RE: Information regarding your article

Dear Jennifer,

Thank you for your email. Under our general policy, journal article authors are permitted to include their article within their own thesis.

Please accept this email as permission for your request as detailed below. Permission is granted for the life of the edition on a non-exclusive basis, in the English language, throughout the world in all formats provided full citation is made to the original SAGE publication. If your thesis will publish prior to article, please note “published ahead of print” along with the journal citation. Additionally, as your article will also be publishing under a CC-BY license, please be certain to include the proper attribution to the license as well.

Please note approval excludes any graphs, photos, excerpts, etc. which required permission from a separate copyright holder at the time of publication. If your material includes anything which was not your original work, please contact the rights holder for permission to reuse those items.

If you have any questions, please let me know.

Best Wishes,

Craig Myles
on behalf of **SAGE Ltd. Permissions Team**

SAGE Publications Ltd
1 Oliver's Yard, 55 City Road
London, EC1Y 1SP
UK

www.sagepub.co.uk

SAGE Publications Ltd, Registered in England No.1017514

Los Angeles | London | New Delhi

Singapore | Washington DC

The natural home for authors, editors & societies

Thank you for considering the environment before printing this email.

From: John Devereux
Sent: Friday, July 21, 2017 5:10 AM
To: PermissionsUK <PermissionsUK@sagepub.com>
Cc: Jennifer Thompson <Jennifer.Thompson@lshtm.ac.uk>
Subject: RE: Information regarding your article

Hi Permissions dept,

Please can you confirm that the author will be able to add her paper (CC-BY) to her thesis. It will not be used in a commercial sense. The paper is still in production.

Appendix B: Ethics approval for Paper B

London School of Hygiene & Tropical Medicine

Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



Observational / Interventions Research Ethics Committee

Miss Jennifer Thompson
LSHTM

22 March 2016

Dear Jennifer

Study Title: Stepped wedge trial misspecification simulation study

LSHTM Ethics Ref: 11019

Thank you for your application for the above research project which has now been considered by the Observational Committee via Chair's Action.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved is as follows:

Document Type	File Name	Date	Version
Protocol / Proposal	Proposal	26/02/2016	1
Investigator CV	CV 2016	26/02/2016	1

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

At the end of the study, the CI or delegate must notify the committee using the End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: <http://leo.lshtm.ac.uk>.

Further information is available at: www.lshtm.ac.uk/ethics.

Yours sincerely,



Professor John DH Porter
Chair

ethics@lshtm.ac.uk
<http://www.lshtm.ac.uk/ethics/>

Improving health worldwide

Appendix C: License Agreement for Paper B

Statistics in Medicine
Published by Wiley (the "Owner")

LICENSE AGREEMENT FOR PUBLISHING CC-BY

Date: May 15, 2017

Contributor name: Jennifer Thompson

Contributor address:

Manuscript number: SIM-16-0376.R2

Re: Manuscript entitled Bias and inference from misspecified mixed effect models in stepped wedge trial analysis (the "Contribution")

for publication in Statistics in Medicine (the "Journal")

published by John Wiley & Sons Ltd ("Wiley")

Dear Contributor(s):

Thank you for submitting your Contribution for publication. In order to expedite the editing and publishing process and enable Wiley to disseminate your Contribution to the fullest extent, we need to have this Agreement executed. If the Contribution is not accepted for publication, or if the Contribution is subsequently rejected, this Agreement will be null and void.

Publication cannot proceed without a signed copy of this Agreement and payment of the appropriate article publication charge.

A. TERMS OF USE

1. The Contribution will be made Open Access under the terms of the [Creative Commons Attribution License](#) which permits use, distribution and reproduction in any medium, provided that the Contribution is properly cited.

2. For an understanding of what is meant by the terms of the Creative Commons License, please refer to [Wiley's Open Access Terms and Conditions](http://olabout.wiley.com/WileyCDA/Section/id-828079.html) (<http://olabout.wiley.com/WileyCDA/Section/id-828079.html>).
3. The Contributor may make use of the submitted and peer reviewed versions of the Contribution prior to publication, provided that the final Contribution is cited appropriately as set forth in paragraph F below. Nothing herein shall permit dual publication in violation of journal ethical practices.
4. The Owner (and Wiley, where Wiley is not the Owner) reserves the right to require changes to the Contribution, including changes to the length of the Contribution, as a condition of acceptance. The Owner (and Wiley, where Wiley is not the Owner) reserves the right, notwithstanding acceptance, not to publish the Contribution if for any reason such publication would in the reasonable judgment of the Owner (and Wiley, where Wiley is not the Owner), result in legal liability or violation of journal ethical practices.

B. RETAINED RIGHTS

The Contributor or, if applicable, the Contributor's Employer, retains all proprietary rights in addition to copyright, such as patent rights in any process, procedure or article of manufacture described in the Contribution.

C. LICENSE

In order to facilitate dissemination of the Contribution in accordance with paragraph A above, the Contributor grants to the Owner, during the full term of the Contributor's copyright and any extensions or renewals, a non-exclusive license of all rights of copyright in and to the Contribution, and all rights therein, including but not limited to the right to publish, republish, transmit, sell, distribute and otherwise use the Contribution in whole or in part in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, and to license or permit others to do so.

D. CONTRIBUTIONS OWNED BY EMPLOYER

If the Contribution was written by the Contributor in the course of the Contributor's employment (as a "work-made-for-hire"), and the employer owns the copyright in the Contribution, the employer company/institution must execute this Agreement (in addition to the Contributor) in the space provided below. In such case, the company/institution hereby grants to the Owner, during the full term of copyright, a non-exclusive license of all rights of copyright in and to the Contribution for the full term of copyright throughout the world as specified in paragraph C above. For company/institution owned work, signatures cannot be collected electronically and so instead please print off this Agreement, ask the appropriate person in your company/institution to sign the Agreement as well as yourself in the space provided below, and upload the signed Agreement to the Wiley Author Services Dashboard. For production editor contact details please visit the Journal's online author guidelines.

E. GOVERNMENT CONTRACTS

In the case of a Contribution prepared under U.S. Government contract or grant, the U.S. Government may reproduce, without charge, all or portions of the Contribution and may authorize others to do so, for official U.S. Government purposes only, if the U.S. Government contract or grant so requires. (U.S. Government, U.K. Government, and other government employees: see notes at end.)

F. COPYRIGHT NOTICE

The Contributor and the company/institution agree that any and all copies of the Contribution or any part thereof distributed or posted by them in print or electronic format as permitted herein will include the notice of copyright as stipulated in the Journal and a full citation to the final published version of the Contribution in the Journal as published by Wiley.

G. CONTRIBUTOR'S REPRESENTATIONS

The Contributor represents that the Contribution is the Contributor's original work, all individuals identified as Contributors actually contributed to the Contribution, and all individuals who contributed are included. If the Contribution was prepared jointly, the Contributor has informed the co-Contributors of the terms of this Agreement and has obtained their written permission to execute this Agreement on their behalf. The Contribution is submitted only to this Journal and has not been published before, has not been included in another manuscript, and is not currently under consideration or accepted for publication elsewhere. If excerpts from copyrighted works owned by third parties are included, the Contributor shall obtain written permission from the copyright owners for all uses as set forth in the standard permissions form or the Journal's Author Guidelines, and show credit to the sources in the Contribution. The Contributor also warrants that the Contribution and any submitted Supporting Information contains no libelous or unlawful statements, does not infringe upon the rights (including without limitation the copyright, patent or trademark rights) or the privacy of others, or contain material or instructions that might cause harm or injury and only utilize data that has been obtained in accordance with applicable legal requirements and Journal policies. The Contributor further warrants that there are no conflicts of interest relating to the Contribution, except as disclosed. Accordingly, the Contributor represents that the following information shall be clearly identified on the title page of the Contribution: (1) all financial and material support for the research and work; (2) any financial interests the Contributor or any co-Contributors may have in companies or other entities that have an interest in the information in the Contribution or any submitted Supporting Information (e.g., grants, advisory boards, employment, consultancies, contracts, honoraria, royalties, expert testimony, partnerships, or stock ownership); and (3) indication of no such financial interests if appropriate.

H. USE OF INFORMATION

The Contributor acknowledges that, during the term of this Agreement and thereafter, the Owner (and Wiley, where Wiley is not the owner) may process the Contributor's personal data, including storing or transferring data outside of the country of the Contributor's residence, in order to process transactions related to this Agreement and to communicate with the Contributor. By entering into this Agreement, the Contributor agrees to the processing of the Contributor's personal data (and, where applicable, confirms that the Contributor has obtained the permission from all other contributors to process their personal data). Wiley shall comply with all applicable laws, statutes and regulations relating to data protection and privacy and shall process such personal data in accordance with Wiley's Privacy Policy located at: <http://www.wiley.com/WileyCDA/Section/id-301465.html>.

I agree to the OPEN ACCESS AGREEMENT as shown above, consent to execution and delivery of the Open Access Agreement electronically and agree that an electronic signature shall be given the same legal force as a handwritten signature, and have obtained written permission from all other contributors to execute this Agreement on their behalf.

Contributor's signature (type name here): Jennifer Thompson

Date: May 15, 2017

SELECT FROM OPTIONS BELOW:

Contributor-owned work

U.S. Government work

Note to U.S. Government Employees

A contribution prepared by a U.S. federal government employee as part of the employee's official duties, or which is an official U.S. Government publication, is called a "U.S. Government work", and is in the public domain in the United States. Contributor acknowledges that the Contribution will be published in the United States and other countries. If the Contribution was not prepared as part of the employee's duties or is not an official U.S. Government publication, it is not a U.S. Government work.

U.K. Government work (Crown Copyright)

Note to U.K. Government Employees

For Crown Copyright this form cannot be completed electronically and should be printed off, signed in the Contributor's signatures section above by the appropriately authorised individual and uploaded to

the Wiley Author Services Dashboard. For production editor contact details please visit the Journal's online author guidelines. *The rights in a contribution prepared by an employee of a UK government department, agency or other Crown body as part of his/her official duties, or which is an official government publication, belong to the Crown and must be made available under the terms of the Open Government License. Contributors must ensure they comply with departmental regulations and submit the appropriate authorisation to publish. If your status as a government employee legally prevents you from signing this Agreement, please contact the Journal production editor.*

[] Other

Including Other Government work or Non-Governmental Organisation work

Note to Non-U.S., Non-U.K. Government Employees or Non-Governmental Organisation Employees

For Other Government or Non-Governmental Organisation work this form cannot be completed electronically and should be printed off, signed in the Contributor's signatures section above by the appropriately authorised individual and uploaded to the Wiley Author Services Dashboard. For production editor contact details please visit the Journal's online author guidelines. *If you are employed by the World Health Organization, please download a copy of the license agreement from <http://olabout.wiley.com/WileyCDA/Section/id-828023.html> and return it to the Journal Production Editor. If your status as a government or non-governmental organization employee legally prevents you from signing this Agreement, please contact the Journal production editor. If your status as a government or non-governmental organisation employee legally prevents you from signing this Agreement, please contact the Journal production editor.*

Name of Government/Non-Governmental Organisation:

[] Company/institution owned work (made for hire in the course of employment)

For "work made for hire" this form cannot be completed electronically and should be printed off, signed and uploaded to the Wiley Author Services Dashboard. For production editor contact details please visit the Journal's online author guidelines.

Name of Company/Institution:

Authorized Signature of Employer:

Date:

Signature of Employee:

Date:

Appendix D: Ethics approval for Paper C

London School of Hygiene & Tropical Medicine

Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



Observational / Interventions Research Ethics Committee

Miss Jennifer Thompson
LSHTM

18 July 2016

Dear Jennifer

Study Title: Reanalysis of XpertMTB/RIF trial in Brazil

LSHTM Ethics Ref: 11741

Thank you for your application for the above research project which has now been considered by the Observational Committee via Chair's Action.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved is as follows:

Document Type	File Name	Date	Version
Investigator CV	CV 2016	24/06/2016	1
Protocol / Proposal	Protocol Final	24/06/2016	1

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

At the end of the study, the CI or delegate must notify the committee using the End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: <http://leo.lshtm.ac.uk>.

Further information is available at: www.lshtm.ac.uk/ethics.

Yours sincerely,



Professor John DH Porter
Chair

ethics@lshtm.ac.uk
<http://www.lshtm.ac.uk/ethics/>

Improving health worldwide

Appendix E: Ethics approval for Paper D

London School of Hygiene & Tropical Medicine

Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



Observational / Interventions Research Ethics Committee

Miss Jennifer Thompson
LSHTM

13 July 2017

Dear Jennifer

Study Title: Stata command enabling permutation tests for stepped wedge trials

LSHTM Ethics Ref: 14361

Thank you for your application for the above research project which has now been considered by the Observational Committee via Chair's Action.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved is as follows:

Document Type	File Name	Date	Version
Investigator CV	CV 2017-06	27/06/2017	1
Protocol / Proposal	Study proposal	28/06/2017	2
Local Approval	Aprovac, a~o CONEP Rollout 1	29/06/2017	1
Local Approval	Aprovac, a~o CONEP Rollout 2	29/06/2017	1
Local Approval	aprovac, a~o CEP Manaus	29/06/2017	1
Local Approval	Aprovac, a~o CEP MRJ	29/06/2017	1

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

At the end of the study, the CI or delegate must notify the committee using the End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: <http://leo.lshtm.ac.uk>.

Further information is available at: www.lshtm.ac.uk/ethics.

Yours sincerely,



Professor John DH Porter
Chair

ethics@lshtm.ac.uk
<http://www.lshtm.ac.uk/ethics/>

