Methods for the design and analysis of stepped wedge trials under model misspecification

Emily Voldal

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:
James Hughes, Chair
Patrick Heagerty
Susanne May

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Methods for the design and analysis of
stepped wedge trials under model misspecification

Emily Voldal

Chair of the Supervisory Committee:
James Hughes
Department of Biostatistics

This dissertation consists of three projects that attempt to address issues related to the design and analysis of stepped wedge trials (SWTs). Stepped wedge trials (SWTs) are a type of cluster-randomized trial that are commonly used to evaluate health care interventions. SWTs are sometimes preferred over parallel cluster-randomized designs due to practical concerns, however some elements of the SWT design (e.g. confounding between time and treatment) make analysis more challenging. Although the popularity of the SWT has been increasing, resources and information for designing and analyzing SWTs are limited compared to more traditional cluster-randomized trial designs. The three projects presented here provide (1) resources for understanding and correctly applying existing methods; (2) an analysis of how model misspecification affects existing methods; and (3) novel methods for analyzing SWTs that are robust to common misspecification concerns.

Project 1: Most SWT-related software packages have restrictive assumptions about the study design and correlation structure of the data. The objective of this project is to present a package and corresponding web-based graphical user interface (GUI) that provide researchers with another, more flexible option for SWT design and analysis. We developed an `R` package `swCRTdesign` ('stepped wedge Cluster Randomized Trial design'), which uses a random effects model to account for correlation in the data induced by a SWT design. Possible

sources of correlation include clusters, time within clusters, and treatment within clusters. `swCRTdesign` allows a user to calculate power, simulate SWT data to streamline simulation studies (e.g. to estimate power), and create descriptive summaries and plots. Additionally, a GUI, developed using `shiny`, is available to calculate power and create power curves and design plots. The `swCRTdesign` package accommodates a wide variety of SWT designs, and makes it easy to account for some sources of correlation which are not found in other packages. The user-friendly web-based GUI makes some `swCRTdesign` features accessible to researchers not familiar with `R`. These two resources will make appropriate and flexible SWT calculations more accessible to scientists from a wide variety of backgrounds.

Project 2: Mixed models are commonly used to analyze SWTs to account for clustering and repeated measures on clusters. One critical issue researchers face is whether to include a random time effect or a random treatment effect. When the wrong model is chosen, inference on the treatment effect may be invalid. We explore asymptotic and finite-sample convergence of variance component estimates when the model is misspecified and how misspecification affects the estimated variance of the treatment effect. For asymptotic results, we rely on analytical solutions rather than simulation studies, which allows us to succinctly describe the convergence of misspecified estimates, even though there are multiple roots for each misspecified model. We found that both direction and magnitude of the bias associated with model-based standard errors depends on the study design and magnitude of the true variance components. We identify some scenarios in which choosing the wrong random effect has a large impact on model-based inference. However, many trends depend on trial design and assumptions about the true correlation structure, so we provide tools for researchers to investigate specific scenarios of interest. We use data from a SWT on disinvesting from weekend services in hospital wards to demonstrate how these results can be applied as a sensitivity analysis, which quantifies the impact of misspecification under a variety of settings and directly compares the potential consequences of different modeling choices. Our results

will provide guidance for pre-specified model choices and supplement sensitivity analyses to inform confidence in the validity of results.

Project 3: Although mixed models are commonly used to analyze SWTs, they are susceptible to misspecification. This is in part because they use 'horizontal' or within-cluster information in addition to 'vertical' or between-cluster information. To use horizontal information in a mixed model, both the mean model and correlation structure must be correctly specified or accounted for, since time is confounded with treatment and time periods are likely correlated within clusters. Alternative non-parametric methods have been proposed that use only vertical information; these are more robust because between-cluster comparisons in a SWT preserve randomization, but these non-parametric methods are not very efficient. We propose a semi-parametric composite likelihood method that focuses on vertical information, but has the flexibility to recover some efficiency by using some horizontal information. We compare the properties and performance of various methods, using simulations based on COVID-19 data. We found that a vertical composite likelihood model that leverages baseline data is more robust than traditional methods, but more efficient than methods that use only vertical information. We hope that these results demonstrate the potential value of semi-parametric vertical methods, and that these new tools are useful to researchers analyzing SWTs who are concerned about misspecification of traditional models.

Finally, we discuss the implications of this work and plans for future work.

TABLE OF CONTENTS

Page

List of Figures iii

Glossary viii

Chapter 1: swCRTdesign: An R Package for Stepped Wedge Trial Design and Analysis 1
1.1 Introduction 1
1.2 The model 5
1.3 swCRTdesign overview 8
1.4 Examples 9
1.5 Web-based power calculator 14
1.6 Discussion 14

Chapter 2: Model misspecification in stepped wedge trials: Random effects for time or treatment 18
2.1 Introduction 18
2.2 Notation and models 21
2.3 Convergence of misspecified parameters 24
2.4 Impact on treatment effect variance 29
2.5 Example 35
2.6 Discussion 38

Chapter 3: Robust analysis of stepped wedge trials using composite likelihood models 43
3.1 Introduction 43
3.2 Methods: background and existing methods 46
3.3 Methods: Vertical composite likelihood estimators 53
3.4 Simulations 58
3.5 Application to LIRE trial 71

3.6 Discussion 72

Chapter 4: Conclusion 83

Bibliography 84

Appendix A: Appendix for Project 2 90
A.1 Relative frequencies of roots 90
A.2 Non-classic designs 94

Appendix B: Appendix for Project 3 96
B.1 Simulation setting details 96
B.2 Simulations: varying sample sizes 97
B.3 Possible extensions 99

LIST OF FIGURES

Figure Number Page

1.1 The design of the EPT trial with 4 sequences in which all sequences start in the control group. In each sequence, there are 6 independent clusters, separated by dotted lines. 3

1.2 Plot created using `swPlot` for the Gaussian example. 16

1.3 Power calculations for the Gaussian example, done in the web-based GUI. Inputs for all the arguments were entered using the sliders in the left-hand column. The range and appearance of the plot were changed in the "Customize plots" tab. 17

2.1 Validity ($var_m(\hat{\theta})/var_s(\hat{\theta})$) of Root 1 for both cases, for a variety of classic designs. Ratios above 1 (indicated by blue outlines) correspond to conservative $var_m(\hat{\theta})$ estimates. Ratios below 1 (indicated by red outlines) correspond to anti-conservative $var_m(\hat{\theta})$ estimates. For each case, $\sigma_t^2 = 5$ and $\tau_t^2 = 0.1$. To keep the ACC consistent, we chose $\eta_t^2 = 0.1$ and $\gamma_t^2 = 0.05$ for the time-fitted random treatment and treatment-fitted random time cases, respectively. . . 32

2.2 Validity ($var_m(\hat{\theta})/var_s(\hat{\theta})$) of Root 1 for both cases, for three designs and a variety of true ACCs. The three classic designs ($M = 2$, $M = 3$, and $M = 6$ sequences) each have $K = 20$ observations per cluster per time period. Throughout, $\sigma_t^2 = 1$. Each ACC is achieved in three different ways by adjusting the balance of τ_t^2 and γ_t^2 or $\eta_t^2/2$. If γ_t^2 or $\eta_t^2/2 = \tau_t^2$, 50% of the average between-cluster variance (the numerator of the ACC) comes from random intercepts. Similarly, if γ_t^2 or $\eta_t^2/2 = \tau_t^2/10$, then 90% of the variance comes from random intercepts and if γ_t^2 or $\eta_t^2/2 = \tau_t^2 * 10$, then 10% of the variance comes from random intercepts. 33

2.3 Efficiency ($var_m(\hat{\theta}_t)/var_s(\hat{\theta})$) of Root 1 for both cases, for three designs and a variety of true ACCs. The three classic designs ($M = 2$, $M = 3$, and $M = 6$ sequences) each have $K = 20$ observations per cluster per time period. Throughout, $\sigma_t^2 = 1$. Each ACC is achieved in three different ways by adjusting the balance of τ_t^2 and γ_t^2 or $\eta_t^2/2$. If γ_t^2 or $\eta_t^2/2 = \tau_t^2$, 50% of the average between-cluster variance (the numerator of the ACC) comes from random intercepts. Similarly, if γ_t^2 or $\eta_t^2/2 = \tau_t^2/10$, then 90% of the variance comes from random intercepts and if γ_t^2 or $\eta_t^2/2 = \tau_t^2 * 10$, then 10% of the variance comes from random intercepts. 34

2.4 Validity of the two proposed models in the disinvestment example if they were both misspecified. Throughout, $\tau_t = 0.28$ and $\sigma_t = 1.03$ were fixed, based on their average fitted values from the two models. We assumed there were $K = 177$ observations per cluster per time period, which was the average K across the trial. 37

3.1 Examples of simulated data under the four scenarios with no treatment effect ($\theta = 0$), with 11 sequences and two clusters per sequence. Each line represents the outcome for a particular cluster over time. 60

3.2 Point estimates of the treatment effect from a SWT with three sequences and 22 clusters in each sequence. Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), NPWP with synthetic control (SC), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA). 62

3.3 Point estimates of the treatment effect from a SWT with 11 sequences and six clusters in each sequence. Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA). 63

3.4 Type I error from a SWT with three sequences and 22 clusters in each sequence, with no treatment effect ($\theta = 0$). Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), NPWP with synthetic control (SC), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA). 64

3.5 Type I error from a SWT with 3 sequences and a varying number of clusters per sequence. The simulated data had no treatment effect, and either random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D). Data was analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA). 65

3.6 Type I error from a SWT with 11 sequences and six clusters in each sequence, with no treatment effect ($\theta = 0$). Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA). 66

3.7 Type I error from a SWT with 11 sequences and a varying number of clusters per sequence. The simulated data had no treatment effect, and either random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D). Data was analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA). 67

3.8 Power from a SWT with three sequences and 22 clusters in each sequence, with a substantial treatment effect ($\theta = -0.28$). Note that some of these are not accurate representations of power, since the corresponding Type I error was elevated. Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), NPWP with synthetic control (SC), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA). . 68

3.9 Power from a SWT with 11 sequences and six clusters in each sequence, with a substantial treatment effect ($\theta = -0.15$). Note that some of these are not accurate representations of power, since the corresponding Type I error was elevated. Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA). Due to their very poor Type I error, MM in scenarios C and D and MMT in scenario D are not shown here. 69

A.1 Frequency at which each root is observed in simulations using `lme4`'s default settings. Throughout, $\sigma_t^2 = 1$. For the 'large' values of τ_t^2, γ_t^2, and η_t^2, we used 0.125. For the 'small' values, we used 0.001. Note that these correspond to ACCs ranging between 0.001 and 0.16 for the time-fitted random treatment case, and ranging between 0.002 and 0.20 for the treatment-fitted random time case. 91

A.2 Efficiency of Root 1 for both cases, for a variety of classic designs. For each case, $\sigma_t^2 = 5$ and $\tau_t^2 = 0.1$. To keep the ACC consistent, we chose $\eta_t^2 = 0.1$ and $\gamma_t^2 = 0.05$ for the time-fitted random treatment and treatment-fitted random time cases, respectively. 92

A.3 Validity and efficiency of Root 1 for the time-fitted random treatment case, for four different SWT designs and a variety of true ACC's. The designs each have two sequences and $K = 5$ observations per cluster per time period. Throughout, $\sigma_t^2 = 1$ and the balance of τ_t^2 and η_t^2 is fixed at $\eta_t^2 = \tau_t^2/2$. . . . 93

B.1 Type I error from a SWT with 3 sequences and either 22 or 33 clusters per sequence. The simulated data had random cluster intercepts (scenario A), and no treatment effect ($\theta = 0$). Data was analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA). 98

GLOSSARY

ACC: Average Cluster Correlation

CAC: Cluster Auto-Correlation

CI: Confidence Interval

CL: Composite Likelihood

CLWP: Composite Likelihood Within-Period

CLWPA: Composite Likelihood Within-Period, Adjusted

CRT: Cluster Randomized Trial

EPT: Expedited Partner Therapy

GEE: Generalized Estimating Equation

GUI: Graphical User Interface

ICC: Intra-Cluster Correlation

LHJ: Local Health Jurisdiction

LIRE: Lumbar Imaging with Reporting of Epidemiology

MCLE: Maximum Composite Likelihood Estimator

MLE: Maximum Likelihood Estimator

MVN: Multivariate Normal

N: Normal/Gaussian

NPWP: Non-Parametric Within-Period

SD: Standard Deviation

SE: Standard Error

SWD: Stepped Wedge Design

SWT: Stepped Wedge Trial

WLS: Weighted Least Squares

ACKNOWLEDGMENTS

I would like to thank the members of my committee, Jim Hughes, Patrick Heagerty, Susanne May, and Kenny Sherr, for their time and patience. Their guidance and insights on these projects and many others have been invaluable. I am especially grateful to my advisor Jim Hughes for his tremendous patience and support. I would also like to extend a special thanks to the stepped wedge working group who have been heavily involved in all these projects, Jim Hughes, Patrick Heagerty, Fan Xia, and Avi Kenny. I feel incredibly lucky to have had the opportunity to learn from such a great group of people.

Although it does not have a chapter in this dissertation, I would like to acknowledge the incredibly valuable and formative experiences I've had working on applied projects during my time in this program. In particular, I would like to give special thanks to Sarah Monsell, Patrick Heagerty, and Bryan Comstock for their support and insights.

I would like to thank all the classmates and study buddies who have helped me get through tough problems. Finally, I would like to thank my family and friends for all their support and encouragement, for which I am eternally grateful. And to Stellaluna: good dog.

DEDICATION

To everyone who has supported me in this endeavor

Chapter 1

SWCRTDESIGN: AN R PACKAGE FOR STEPPED WEDGE TRIAL DESIGN AND ANALYSIS

This research has been published in the journal Computer Methods and Programs in Biomedicine (Voldal EC, Hakhu NR, Xia F, Heagerty PJ, and Hughes JP. swCRTdesign: An R Package for Stepped Wedge Trial Design and Analysis. *Computer Methods and Programs in Biomedicine*, 196:105514, 2020. doi: 10.1016/j.cmpb.2020.105514).

1.1 Introduction

1.1.1 Basics of stepped wedge trials

Scientists in medicine, public health, and many other fields rely on randomized controlled trials as the gold standard method to evaluate interventions. Typically, individuals are randomized to intervention or control. However, sometimes clusters of individuals (e.g., families, clinics, or communities) are randomized, in which case individual-level observations are correlated; this is called a cluster randomized trial (CRT). The motivation for choosing a CRT instead of a traditional randomized trial may be due to the nature of the intervention (e.g., policy changes at a hospital) or for practical or logistical reasons. To date, most CRTs have mirrored the design of individually randomized trials and used a parallel or matched parallel two group design. One exception is the stepped wedge trial (SWT) design. The SWT is a more recently developed type of CRT that has been growing in popularity [1]. SWTs are a type of crossover design, but crossover only occurs in one direction. The most common design for a SWT dictates that all clusters begin in the control state at baseline, and at every subsequent time point (also called periods) some of the clusters cross over until all clusters have received treatment. Cross-over times are pre-specified for each cluster

and usually evenly spaced after baseline. A group of clusters that cross over to treatment at the same time is called a sequence or wave. See Hughes, Granston, and Heagerty [2] for a more thorough introduction to stepped wedge designs. SWTs may have logistical or ethical advantages compared to other CRT designs, since the researcher only needs to implement the intervention in a few clusters at a time and every cluster eventually receives the intervention [1]. However, because of the complex design and correlated observations, specialized statistical methods and software are needed for designing SWTs.

The `R` [3] package `swCRTdesign` [4] is a tool to aid in the design and analysis of SWTs, including calculating power, simulating SWT data, and generating descriptive summaries. Compared to other resources (see below), this tool can be applied to a wider range of SWT designs and accommodate more complex data-generating models. Additionally, a web-based graphical user interface (GUI) for calculating power is available, with the hope of making some `swCRTdesign` features more accessible to researchers not familiar with `R`.

1.1.2 Motivating example

As a motivating example, we consider the stepped wedge trial on expedited partner therapy (EPT) in Washington state summarized by Golden et al [5]. With EPT, when an individual tests positive for a sexually transmitted infection, their sex partners are then offered treatment without medical evaluation. In this study, the clusters were local health jurisdictions (LHJs), which typically corresponded to counties. Although the authors analyzed several primary and secondary outcomes, here we focus on the effect of EPT on the percent of chlamydia tests that were positive among women between 14 and 25 years old at sentinel clinics within each LHJ. Researchers planned on randomizing 24 LHJs into four sequences of 6 LHJs each, stratified by region and LHJ size. Every sequence was observed in the control state at baseline, and at each subsequent time one sequence crossed over, resulting in a total of five time points. Figure 1.1 depicts the study design for this motivating example. Note that although the x-axis is labeled "Time 1", "Time 2", etc., the distance between observed times was between six and eight months, so these times could alternatively be labeled "Month

1", "Month 7", etc.

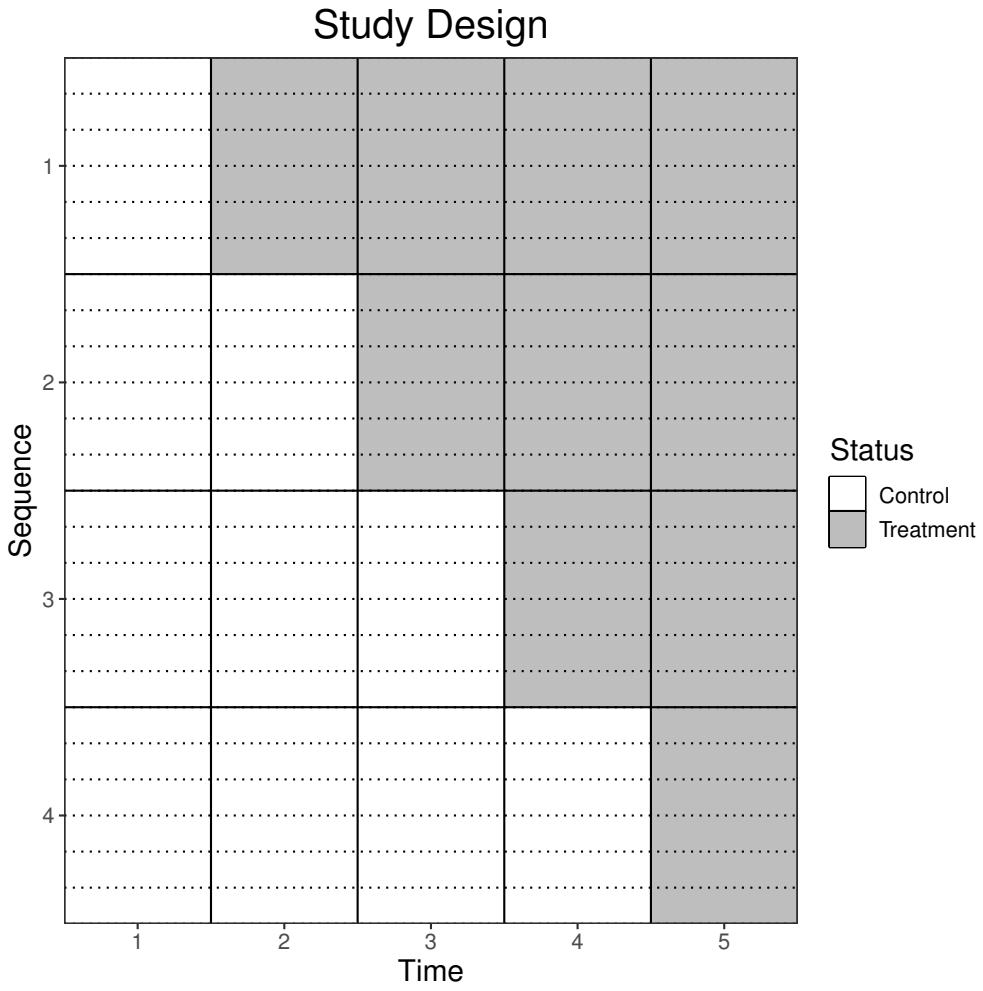

Figure 1.1: The design of the EPT trial with 4 sequences in which all sequences start in the control group. In each sequence, there are 6 independent clusters, separated by dotted lines.

1.1.3 Summary of methods and software

There are a wide variety of methods for designing and analyzing SWTs [1]. For calculating power, the two most common methods are based on (i) application of normal-theory models

via weighted least squares (WLS), with approximations for binary and count outcomes [6] and (ii) simulation [7]. A 'design effect' or 'sample size correction factor' is sometimes also used [8]. Baio et al [7] also provide an overview of these three techniques and their use for SWTs. `swCRTdesign` uses method (i). Very briefly, this method uses a mixed model structure to compute the variance of the WLS estimate of the treatment effect; power is based on a Wald statistic and relies on asymptotic Normality. For details, see [6] and [2] (especially Appendix A).

There are two other `R` packages that can be used for designing and analyzing SWTs. `SamplingDataCRT` [9] allows the user to simulate data and calculate power, but only for Gaussian outcomes and limited covariance structures. Power calculations are based on Hussey and Hughes [6] for cross-sectional data. `SWSamp` allows the user to simulate data and calculate power based on Hussey and Hughes, a design effect, or simulations, but most functionality is only supported for a limited number of covariance structures [10]. A web-based GUI developed by Hemming et al [11] allows the user to calculate power based on a design effect for Gaussian, binary, or count outcomes and several possible covariance structures. `swCRTdesign` provides support for more complex SWT designs (e.g., a sample size that varies over time and cluster) and alternative ways of specifying correlation structures.

1.1.4 Overview

The remainder of this chapter is organized as follows. Section 1.2 describes the underlying analytic model used to compute power for a SWT design and simulate SWT data. Section 1.3 contains an overview of the `swCRTdesign` package, followed by examples with `R` code in section 1.4. Section 1.5 describes the web-based GUI tool. We conclude with a discussion in section 1.6. We hope that making both `swCRTdesign` and the GUI freely available will help support the needs of investigators designing or analyzing a SWT.

1.2 The model

1.2.1 Primary parameterization

To account for the correlation within clusters, we use a random effects model [2]. This model is used for both the power calculations, and simulating SWT data. Let Y_{ijk} denote the response recorded for individual k from cluster i at time j. In the `swCRTdesign` package, Y_{ijk} may follow a Gaussian, Bernoulli, or (for simulations only) Poisson distribution. Also, let X_{ij} have a value of one if cluster i is assigned to treatment at time j, and zero otherwise. In this chapter and package, time is modeled as an unordered categorical factor. Then the mean response for cluster i at time j is:

$$\mu_{ij} = \mu + \beta_j + \theta * X_{ij} + u_i + w_{ij} + v_i * X_{ij}. \tag{1.1}$$

Here, μ is the baseline average (i.e., the average response among untreated individuals in the first time period), θ represents the treatment effect, and β_j represents the mean difference between time j and time 1, with $\beta_1 = 0$. The random effects are represented by u_i, w_{ij}, and v_i; each follows a Gaussian distribution with mean zero. The random intercepts, u_i, represent each cluster's unique baseline and have a standard deviation of τ. The random time effects, w_{ij}, represent cluster- and time- specific deviations and have a standard deviation of γ. The random treatment effects, v_i, represent the cluster-specific variations in the impact of treatment, and have a standard deviation of η. We assume that the correlation between u_i and v_i is ρ, and the correlation between w_{ij} and both u_i and v_i is zero. In addition, we assume the correlation between w_{ij} and $w_{ij'}$ with $j \neq j'$ is zero. So in cluster i with times $j = 1, \ldots, J$, the covariance of the vector of random effects is:

$$cov\begin{bmatrix} u_i \\ v_i \\ w_{i1} \\ \vdots \\ w_{iJ} \end{bmatrix} = \begin{bmatrix} \tau^2 & \rho\tau\eta & 0 & \ldots & 0 \\ \rho\tau\eta & \eta^2 & 0 & \ldots & 0 \\ 0 & 0 & \gamma^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \gamma^2 \end{bmatrix}. \tag{1.2}$$

When the response Y_{ijk} is assumed to be Gaussian, $Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$ where ϵ_{ijk} is Gaussian with mean zero and standard deviation σ; that is, the mean of Y_{ijk} is μ_{ij} and the standard deviation is $\text{sd}(Y_{ijk}|u_i, w_{ij}, v_i) = \sigma$. The ϵ_{ijk} are mutually independent, and independent of all random effects. When Y_{ijk} is assumed to be Bernoulli, the mean is μ_{ij} and the standard deviation is defined by the function $\text{sd}(Y_{ijk}|u_i, w_{ij}, v_i) = \sqrt{\mu_{ij}(1-\mu_{ij})}$. Similarly, when Y_{ijk} is assumed to be Poisson, the mean is μ_{ij} and the standard deviation is defined by the function $\text{sd}(Y_{ijk}|u_i, w_{ij}, v_i) = \sqrt{\mu_{ij}}$.

Note that these calculations assume that each subject in the data set is only observed once, at a single time point. If the same subjects are observed at multiple time points, we refer to this as a 'cohort' design; this requires additional random effects [12]. Currently, `swCRTdesign` does not support power calculations or simulations for cohort data. The application developed by Hemming et al [11] allows the user to calculate power for some cohort designs. `SamplingDataCRT` also provides some support for cohort data, but only for Gaussian outcomes and cluster random effects [9].

1.2.2 Alternative parameterization

Instead of using the variance components above to quantify correlation (which we will refer to as the standard deviation or SD parameterization), some researchers prefer to use measures such as the within-period intra-cluster correlation (ICC) and cluster auto-correlation (CAC) [12]. When there are no random treatment effects (η=0), the two parametrizations are equivalent. However, the SD parameterization allows for random treatment effects (through η) and a correlation between the random treatment effects and the random intercepts (ρ), while a single ICC and CAC do not [13]. The ICC/CAC parameterization is less flexible than the SD parameterization since it requires the assumption that there are no random treatment effects (that is, $\eta = 0$ and $\rho = 0$). ICC and CAC are widely used in the literature both to summarize SWT data and perform power calculations [1]. We hope that including this parameterization makes power calculations more accessible for researchers familiar with ICC/CAC, but researchers interested in random treatment effects (e.g. [14]) will need to use

the SD parameterization.

The ICC is the correlation between individuals sampled from the same cluster at the same time, or the ratio of between-cluster variance to total variance [12]. In terms of the random effects parameters defined above,

$$\text{ICC} = \frac{\tau^2 + \gamma^2}{\tau^2 + \gamma^2 + \sigma^2}. \tag{1.3}$$

When the outcome is assumed to be Gaussian, σ is well defined. When the outcome is assumed to be Bernoulli, we substitute $\sigma = \sqrt{\bar{\mu}(1-\bar{\mu})}$, where $\bar{\mu}$ is the average of the mean outcome in the control group (μ_0) and treatment group (μ_1), $(\mu_0 + \mu_1)/2$. For example, if there is no time trend $\mu_0 = \mu$ and $\mu_1 = \mu + \theta$. The CAC is the ratio of the between-period ICC to the within-period ICC, or the ratio of the correlation between individuals from the same cluster at different times to the correlation between individuals from the same cluster within the same period [12]. In this scenario,

$$\text{CAC} = \frac{\tau^2}{\tau^2 + \gamma^2}. \tag{1.4}$$

In this package, ICC and CAC input is translated to random effect standard deviations. If the ICC is zero, both τ and γ are zero and the CAC is not well-defined. If the ICC is one, either σ is zero or either τ or γ is extremely large. Since solving for the random effects parameters is not possible in this case, it is best to use the SD parameterization directly when the ICC is close to one.

The formulas below can be used to manually convert from ICC, CAC, and σ to the SD parameterization:

For CAC= 1, $\gamma = 0$ and $\tau = \sigma * \sqrt{\text{ICC}/(1-\text{ICC})}$.

For CAC< 1, $\gamma = \sigma * \sqrt{\text{ICC} * (1-\text{CAC})/(1-\text{ICC})}$ and $\tau = \gamma * \sqrt{\text{CAC}/(1-\text{CAC})}$.

1.2.3 Fractional treatment effect

In the case described above, we assume that the clusters receive the full treatment effect θ as soon as they cross over to treatment (i.e., X_{ij} changes from 0 to 1). However, the

effect of an intervention is not always immediate, and usually depends on the scientific setting, mechanism of the intervention, and the planned timing of assessments in a trial. For example, perhaps when a cluster first crosses over to treatment the effect size is $\frac{\theta}{2}$, then θ at every subsequent time point. Fractional treatment effects may extend over as many time points as desired by specifying a series of X_{ij} between 0 and 1. Depending on the planned trial duration, some clusters may not be followed long enough for researchers to observe the full treatment effect. For more detail, see Hughes, Granston, and Heagerty [2].

1.3 `swCRTdesign` *overview*

The `swCRTdesign` package contains five main functions: `swDsn` for defining a SWT design, `swPwr` for calculating power, `swSim` for simulating data, and `swPlot` and `swSummary` for plotting and summarizing a SWT data set. The full documentation can be found at `https://cran.r-project.org/web/packages/swCRTdesign/swCRTdesign.pdf`.

The functions `swPwr` and `swSim` both depend on defining a SWT design through `swDsn`. To define a design, a vector (`clusters`) specifies the number of sequences (the length of the vector) and number of clusters per sequence. Designs can either begin with all sequences in the control setting (default), or begin with the first sequence on treatment. Extra time periods can be added while all sequences are on control, or after all sequences have switched to treatment. Designs may also have extra time points throughout the study during which no sequences cross over, or have transition periods.

`swPwr` calculates power for either a Bernoulli or Gaussian outcome. The number of observations per cluster may vary by cluster and over time. When simulating data with `swSim`, the outcome may be Gaussian, log-Gaussian, Poisson (identity or log link), or Bernoulli (identity, log, or logit link). Both `swPwr` and `swSim` allow either ICC/CAC or the SD parameterization in some cases.

`swSummary` summarizes stepped wedge data over time with respect to sequences, or with respect to individual clusters. The summaries can be by mean, sum, or number of observations. `swPlot` can plot trends by sequence or by cluster, either in a single plot or with one

plot for each sequence. Default colors, symbols, and line types differentiate between different clusters or sequences, and indicate when each sequence or cluster was under treatment or control. If desired, these can be customized or eliminated.

More details can be found in the documentation [4] on the syntax and flexibility of the `swCRTdesign` package. The examples below are meant to illustrate the capabilities of the package, but not provide a thorough tutorial of the package's use.

1.4 Examples

1.4.1 Gaussian example

To illustrate the use of the `swCRTdesign` package, suppose we are interested in a traditional SWT with 5 sequences and 6 clusters per sequence. We start by creating a design, and printing the design matrix (`swDsn.unique.clusters`). Rows and columns correspond to clusters and time points, respectively. Control and treatment status are represented by 0's and 1's.

```
> library("swCRTdesign")
> design <- swDsn(clusters = c(6, 6, 6, 6, 6))
> design$swDsn.unique.clusters

     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    1    1    1    1    1
[2,]    0    0    1    1    1    1
[3,]    0    0    0    1    1    1
[4,]    0    0    0    0    1    1
[5,]    0    0    0    0    0    1
```

Next, using our planned design and external information about the assumed variability in the data, we can calculate power. Here, we have Gaussian data with 50 individuals in every

cluster at every time point. Argument names mirror the symbols in the model described above (Equation 1.1).

```
> swPwr(design = design, distn = "gaussian", n = 50,
+    mu0 = 0, mu1 = 0.003, sigma = .03,
+    tau = .01, eta = 0, rho = 0, gamma = 0.001, silent = TRUE)
[1] 0.7399873
```

Alternatively, transforming from the SD parameterization to the ICC/CAC parameterization, we can obtain the same power estimate using ICC and CAC. Note that this is possible only because we assumed there are no random treatment effects ($\eta = 0$) in this scenario.

```
> icc.ex <- (.01^2 + .001^2) / (.01^2 + .001^2 + .03^2)
> icc.ex
[1] 0.1008991
> cac.ex <- .01^2 / (.01^2 + .001^2)
> cac.ex
[1] 0.990099
> swPwr(design = design, distn = "gaussian", n = 50,
+     mu0 = 0, mu1 = 0.003, sigma = .03,
+    icc = icc.ex, cac = cac.ex, silent = TRUE)
[1] 0.7399873
```

Additionally, we can simulate SWT data based on the assumed design and parameters. Note that we added a positive time trend (with `time.effect`); this corresponds to the vector of $\beta_1 \dots \beta_6$ in Equation 1.1. The default output contains the outcome (`response.var`), treatment indicator (`tx.var`), and the time and cluster IDs (`time.var`,`cluster.var`).

```
> set.seed(4)
> example.data <- swSim(design = design, family = "gaussian", n = 50,
+    mu0 = 0, mu1 = 0.003, time.effect = (0:5) * .002, sigma = .03,
```

```
+    tau = .01, eta = 0, rho = 0, gamma = 0.001)
> head(example.data)
  response.var tx.var time.var cluster.var
1 -0.004442856      0        1           1
2 -0.004531357      0        1           1
3 -0.005274664      0        1           1
4  0.064369213      0        1           1
5  0.049816709      0        1           1
6 -0.028956710      0        1           1
```

Using `swSummary`, we can examine the mean response for each sequence over time, and plot that using `swPlot` (Figure 1.2). The additional arguments in `swPlot` customize the location and width of the legends and provide a title; see package documentation for details.

```
> example.summary <- swSummary(response.var = response.var,
+    tx.var = tx.var, time.var = time.var, cluster.var = cluster.var,
+    data = example.data, type = "mean")
> #To view the matrix of means by sequence and time, we could run:
> #example.summary$response.wave


> swPlot(response.var = response.var, tx.var = tx.var,
+    time.var = time.var, cluster.var = cluster.var,
+    data = example.data, choose.ncol = 5,
+    choose.tx.pos = "bottom", choose.legend.pos = "bottomright",
+    choose.main = "SWT Example Plot")
```

1.4.2 Bernoulli example (EPT trial)

Recall the motivating example from the introduction, in which Golden et al [5] examined the effect of EPT on the percent of chlamydia tests that were positive. Each cluster is a LHJ, and

each individual observation is a chlamydia test result (positive or negative), which follows a Bernoulli distribution. When planning the study, researchers identified 24 eligible LHJs, and planned on randomizing them into four sequences. Below, we construct this design (note that this design is identical to the one in Figure 1.1).

```
> design.bernoulli <- swDsn(clusters = c(6, 6, 6, 6))
```

To estimate power for this SWT design, researchers assumed that there would be 162 tests per LHJ at each observed time, and that 5% of the tests would be positive in the control setting. They also estimated that $\tau = 0.0165$, and assumed that there would be no other random effects. The trial was powered to detect a multiplicative change of 0.7; that is, 3.5% positive in the treatment group. With a significance level of 0.05, the calculation below aligns with the researchers' desired power of at least 80%.

```
> swPwr(design = design.bernoulli, distn = "binomial", n = 162,
+    mu0 = 0.05, mu1 = 0.035,
+    tau = 0.0165, eta = 0, rho = 0, gamma = 0, silent = TRUE)
[1] 0.8468701
```

The researchers stated that they wished to use a log link in their model for the relationship between EPT and chlamydia positivity. We can simulate data that accounts for this. Because we are using a log link, `mu0` and `mu1` are on the log scale, so we enter the log of the prevalences. Similarly, `tau`, `eta`, `rho`, and `gamma` correspond to random effects that are linear on the log scale, so we use the predicted coefficient of variation between LHJs for `tau`. Note that this data does not exactly correspond with the mixed effects model used by Golden et al [5], since they also included random effects for individual clinics within the LHJs and a fixed time trend. The WLS technique used by `swPwr` for estimating power cannot easily account for a nonlinear link, but the data can be simulated; this is one scenario where simulation-based power calculations may be useful.

```
> example.data.bernoulli <- swSim(design = design.bernoulli,
```

```
+    family = binomial(link = "log"), n = 162,
+    mu0 = log(0.05), mu1 = log(0.035), time.effect = 0,
+    tau = .33, eta = 0, rho = 0, gamma = 0)
> #To view the simulated data, we could run:
> #head(example.data.bernoulli)
```

Unfortunately, while the researchers planned on 24 LHJs and 162 tests per LHJ at each time point, when the study was conducted they only had 22 LHJs and about 107 tests per LHJ at each time point. Researchers assigned six LHJs to each of the first three waves, and only four LHJs to the last wave, resulting in the design below.

```
> design.bernoulli.post <- swDsn(clusters = c(6, 6, 6, 4))
```

For the purpose of this example, suppose we know the exact number of observations in each LHJ at each time point. A plausible matrix of sample sizes is simulated below. Each row represents a LHJ, and each column represents a time point. LHJs are ordered by sequence, from first to cross over to last to cross over, so the sample sizes printed below correspond to the six LHJs in the first sequence to cross over. Within a sequence, the order of clusters is arbitrary.

```
> n.bernoulli.post <- matrix(data = rpois(n = 110, lambda = 107),
+    nrow = 22, ncol = 5)
> head(n.bernoulli.post)
     [,1] [,2] [,3] [,4] [,5]
[1,]   99   95  118  108  110
[2,]  105  139  104   93  121
[3,]  100   95  118  105  111
[4,]   95  113   85  124   96
[5,]  112  101  107  102  120
[6,]  116  117  113   97  128
```

Now we can see how these changes impact the power of the study.

```
> swPwr(design = design.bernoulli.post, distn = "binomial",
+    n = n.bernoulli.post, mu0 = 0.05, mu1 = 0.035,
+    tau = 0.0165, eta = 0, rho = 0, gamma = 0, silent = TRUE)
[1] 0.6432843
```

1.5 Web-based power calculator

The web-based GUI power calculator provides investigators with an alternative tool (that does not require knowledge of `R`) to compute power for SWTs and explore the impact on power of candidate SWT designs. This was developed using the `shiny` package [15]. The app can be accessed online (`https://swcrtdesign.shinyapps.io/stepped_wedge_power_calculation/`) or downloaded and run in `R` (`https://github.com/swCRTdesign/Stepped-wedge-power-calculation`). The language and inputs are very similar to the `swCRTdesign` package, described above, and a manual is provided in the 'Help and References' tab.

In addition to calculating the power for a specific design, the app can also plot a power curve, visualize and summarize the study design, and help translate between the ICC/CAC and SD parameterization. Some plots use the package `ggplot2` [16]. Figure 1.3 shows the power calculations from the Gaussian example above. Not pictured is the plot of the study design which the GUI also generates, in the same style as Figure 1.1.

1.6 Discussion

The `R` package `swCRTdesign` provides useful tools for the design and analysis of SWTs. These functions make it easy to calculate power, simulate data, and plot and summarize SWT data. The online power calculator app makes power calculations accessible without knowledge of `R`, and facilitates exploration of different designs. We believe these will be useful tools for applied researchers from a variety of backgrounds, as the use of SWTs continues to grow.

A limitation of `swCRTdesign` is that, currently, power cannot be calculated for cohort designs. Calculating power for SWTs with more complex correlation structures is an important extension that may be added in a future version of the package. Other potential additions include allowing more flexible specification of the design matrix and random effects covariance matrix, and adding power calculations for logit and log link models.

Computational details

The results in this paper were obtained using `R` 3.6.1 with the `swCRTdesign` 3.1 package. In addition, the `ggplot2` 3.2.0 package was used to create Figure 1.1.

Acknowledgments

Funding: Research reported in this work was partially funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1507-31750). The statements in this work are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors, or Methodology Committee. PCORI had no direct involvement in this work.

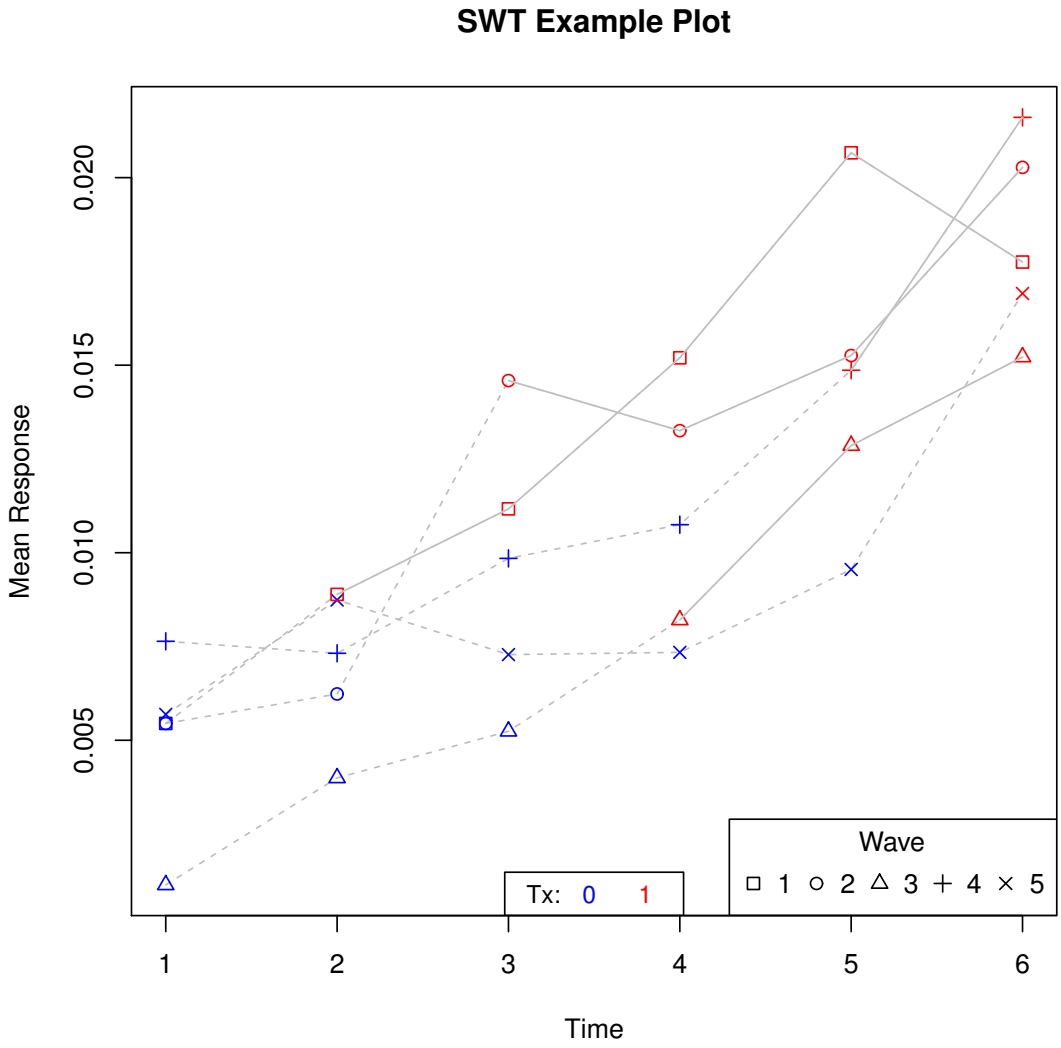

Figure 1.2: Plot created using `swPlot` for the Gaussian example.

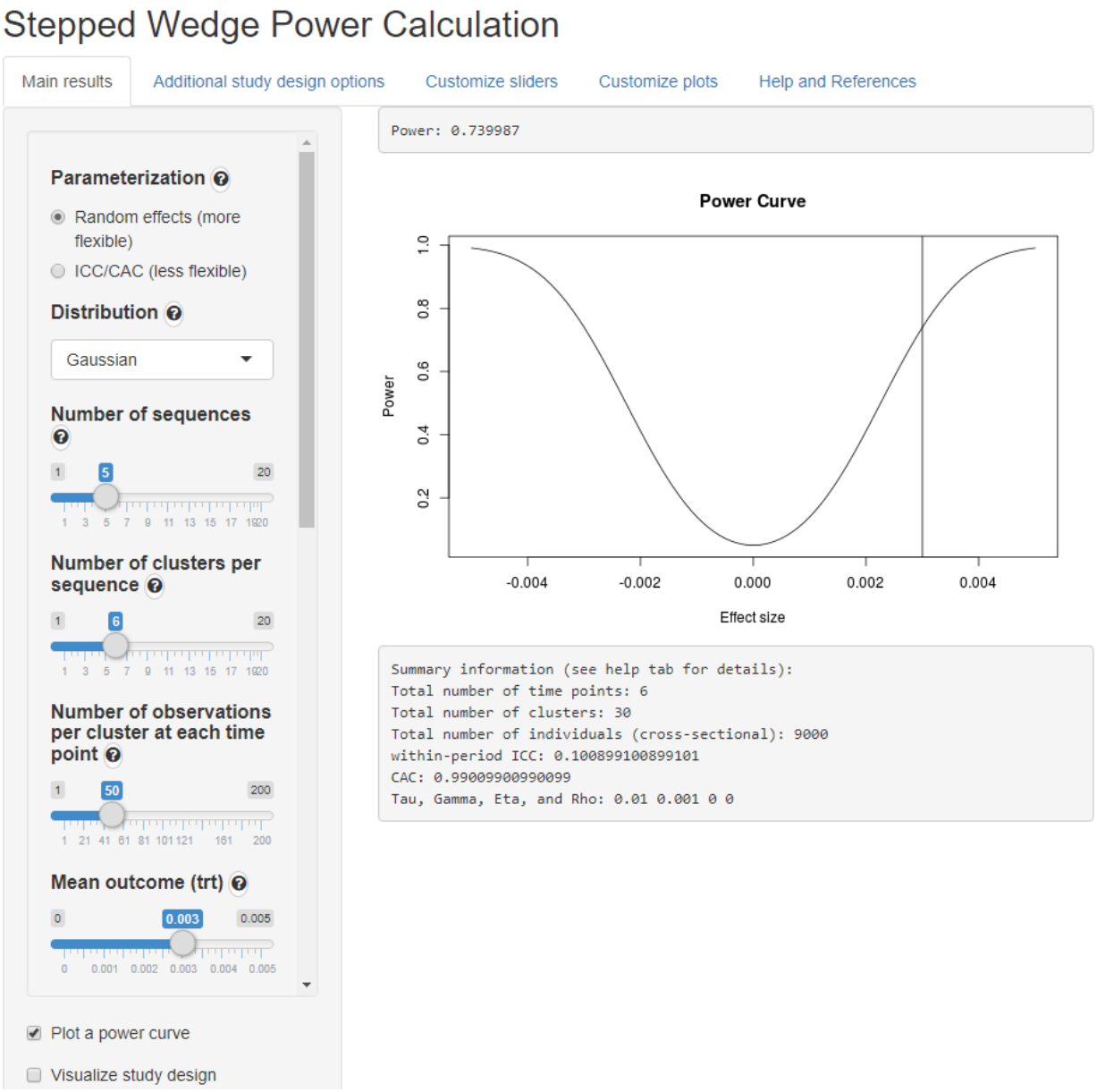

Figure 1.3: Power calculations for the Gaussian example, done in the web-based GUI. Inputs for all the arguments were entered using the sliders in the left-hand column. The range and appearance of the plot were changed in the "Customize plots" tab.

Chapter 2

MODEL MISSPECIFICATION IN STEPPED WEDGE TRIALS: RANDOM EFFECTS FOR TIME OR TREATMENT

This research has been published in the journal Statistics in Medicine (Voldal EC, Xia F, Kenny A, Heagerty PJ, Hughes JP. Model misspecification in stepped wedge trials: Random effects for time or treatment. *Statistics in Medicine*, 41(10):1751-1766, 2022. doi: 10.1002/sim.9326).

2.1 Introduction

Stepped wedge trials (SWTs) are a type of cluster randomized trial that have been growing in application[1]. In a typical SWT, all clusters begin in the control state at baseline, and at every subsequent time point some of the clusters cross over until all clusters have received treatment. A group of clusters that adhere to the same unique pattern of crossing over is called a sequence or wave. Cross-over times are pre-specified for each sequence and clusters are typically randomized into sequences. See Hughes, Granston, and Heagerty [2] for a more thorough introduction to SWTs.

One of the most common tools used for the analysis of SWTs is a mixed model [1]. The overall treatment impact is represented by a fixed effect, and because time and treatment are confounded in SWTs it is critical to also include a fixed effect for time [6]. The unique SWT structure gives rise to three natural choices for random effects: random intercept effects (sometimes called random cluster effects since each cluster is allowed a unique intercept), random time effects, and random treatment effects. Random intercept effects have been recommended as the most fundamental random effect by Hussey and Hughes [6]. Other authors have also recommended adding random time effects [17] and/or random treatment

effects [14] in addition to the random intercept effects. Most researchers using a mixed model for a SWT choose a model somewhere between the "full model" (random intercept, time, and treatment effects) and the random intercepts only model. SWTs that follow cohorts of individuals or that have more than two levels (e.g. subjects are grouped into schools, schools are grouped into school districts) may have additional random effects, but we do not address these designs here. Misspecification of the mean model via fixed effects is also an important topic, but in this manuscript we focus on potential misspecification of the covariance structure via the choice of random effects.

There has been ample research on the impacts of over-fitting and under-fitting random effects in mixed models. Choosing the full model may be inefficient but this strategy ensures that inference on the treatment effect is valid [18]. Researchers who choose a random intercepts only model when there are actually additional random effects risk invalid inference on the treatment effect [18]. As a result, some scientists advise erring in the direction of including potentially unnecessary random effects [18].

Unfortunately, it is not always possible or practical to fit the full model. If a SWT has a limited number of clusters, a researcher might be hesitant to specify such a complex model given limited replication across clusters. Even when a SWT has a large number of clusters, sometimes mixed model software will fit variance components as exactly zero, effectively simplifying the model's covariance structure; in this case, it may be unclear whether fitted values are actually the maximum likelihood estimators (MLEs) or are the result of algorithm convergence issues that may be associated with multiple roots. Alternatively, software might fit non-zero variance components but encounter other convergence issues, and the researcher might decide to remove a random effect to improve convergence. This decision to reduce the model is equivalent to a partially data-driven model selection procedure, which may not be desirable or appropriate. Some of these issues with model fit can be mitigated without removing components from the model, e.g. using strategies presented by Cheng et al [19]. However, if the full model is an over-parameterization, this can cause convergence issues which cannot be mitigated without removing terms from the model. Therefore, choice of

a primary pre-specified model is often challenging and driven by both desire for insurance of broad potential validity (favoring more complex models) and desire for stable estimation properties (favoring more simple models).

A researcher selecting a reduced random effects model must decide which component is the least important. In a SWT, the random intercept effects are the highest level of correlation, so we do not want to remove those from the model; instead, we may choose to remove either time or treatment random effects. Unfortunately, practical guidance on model choice and robustness is not readily available in the current literature. Other papers have examined the impact of excluding nested random effects [20], but time and treatment random effects are not nested in SWTs, so it is not immediately apparent which one might be less detrimental to exclude. Thompson et al. [17] used simulation studies to examine misspecification of random effects in a specific SWT design. They found that including a random time effect can account for some variation coming from a random treatment effect (and vice versa), although a model with random time effects was more robust than a model with random treatment effects. However, these simulations only covered binary outcomes in a cohort SWT with an unusual design: three sequences observed over only two time points, where one sequence remains exclusively on control, one crosses over, and the last receives treatment at both time points. It is unknown whether the observations from these simulations hold for more general SWT designs.

Although researchers who are interested in a very specific and simple case may find it feasible to explore model misspecification via simulations, there are some issues which make simulation studies unattractive. First, they are very time-consuming, particularly if a study has many clusters or complex random effects. Second, issues with model fit make it difficult to reflect real-life research decisions in automated simulations. For example, Thompson et al. [17] excluded simulations where the model failed to converge. Depending on the settings, this resulted in up to 33% of simulations being excluded from the results, potentially biasing conclusions. Last, reliance on simulations may conceal crucial details. For example, if there are multiple asymptotic solutions it may be very difficult to detect that with simulations.

For these reasons, we avoid simulations and rely instead on closed-form solutions whenever possible. This enables one to quickly and accurately examine a vast range of settings and reveals some complexities which would be difficult to discern through simulations.

As a motivating example, we consider a study reported by Haines et al. [21] that investigated the impact of removing weekend health services (or 'disinvesting') from 12 hospital wards in Australia. The original investigators did not report the use of any random time or treatment effects, yet we know that using an over-simplified covariance structure could impact the validity of their conclusions. A complex random effects model would be an ambitious choice given the limited number of clusters. Therefore, we use this concrete example to illustrate the practical issues around selection of alternative covariance models and in this case study we evaluate how that choice may affect inference.

The layout of the paper is as follows: in Section 2.2, we describe the models and notation. In Section 2.3, we examine the convergence of the misspecified parameters. In Section 2.4, we calculate and visualize the asymptotic impact of misspecification on the treatment effect variance. To demonstrate how our results might be used in practice, we apply this research to the disinvestment example in Section 2.5. Finally, a brief discussion is given in Section 2.6.

2.2 Notation and models

2.2.1 The SWT design

In this chapter, we consider SWT designs with M unique treatment sequences, each observed over J time points. We assume that each sequence contains N clusters, and each cluster contains K individuals at each point in time. We assume that each individual is observed only once; that is, a cross-sectional design as opposed to a cohort design. For each sequence, we denote T_m as the number of time points during which sequence m is receiving treatment. Throughout, $m = 1$ corresponds to the sequence that crosses over first, and $m = M$ corresponds to the last sequence to cross over. Figure 1.1 shows an example of a

'classic' SWT design; that is, every cluster starts on control and ends on treatment, and one sequence crosses over at each time point. In this example, there are $M = 4$ sequences, $J = 5$ time points, $N = 6$ clusters per sequence, and T_1=4, T_2=3, T_3=2, and T_4=1. In non-classic designs, modifications might include: extra time on treatment or control; some sequences never receiving treatment or control; or extra time points between transitions. Another common modification to the classic design is to vary the number of clusters per sequence or individuals per cluster; however, these generalizations of N and K are not addressed in this chapter. Note that classic designs can be fully described by just M, N, and K, since $J = M + 1$ and $T_1, \ldots, T_M = M, \ldots, 1$.

2.2.2 The model

Let Y_{ijk} denote the response recorded for individual k from cluster i at time j, where $k = 1, \ldots, K$, $j = 1, \ldots, J$, and $i = 1, \ldots, M * N$. Also, let X_{ij} have a value of one if cluster i is assigned to treatment at time j, and zero otherwise. Define the JK-by-1 vector of outcomes from cluster i as $Y_i = (Y_{i11}, Y_{i12}, \ldots, Y_{iJK})^T$. Then, for a Normal outcome and identity link, the mixed model can be written as

$$Y_i = X_i\Phi + Z_i a_i + \epsilon_i. \tag{2.1}$$

where X_i is the design matrix for the fixed effects, Φ is the vector of fixed effect coefficients, Z_i is the design matrix for the random effects, a_i is the vector of random effects, and ϵ_i is the residual variance. We assume the JK-by-1 vector ϵ_i has a $MVN(0, \sigma^2 I_{JK})$ distribution, where I_{JK} represents an identity matrix of dimension JK. Throughout, we assume that ϵ_i and a_i are independent. We also assume that $a_i \sim MVN(0, G)$, where Z_i, a_i, and G all depend on which random effects model we select. See below for details.

Although some results hold regardless of the mean model specification, we will primarily be considering fixed effects for just treatment and time, and modeling time as linear. Thus the fixed effect component for Y_{ijk} is $\mu+(j-1)\beta+\theta X_{ij}$, where μ is an intercept, β is the time slope, and θ is the treatment effect. Then we can write $\Phi = (\mu, \beta, \theta)^T$, and X_i is JK-by-3

with rows $(1, j-1, X_{ij})$.

Let α be the vector of all parameters in this model, including Φ, σ^2, and any parameters contained in G.

The random time effect model

In the random time effect model, $a_i = (u_i, w_{i1}, \ldots, w_{iJ})^T$, where u_i is the random intercept effect and w_{ij} is the random time effect, treating time as categorical for maximum flexibility. Note that frequently both the fixed and random time effects are categorical, but using the simplified linear fixed time effects allows us to compare designs with different numbers of time points while holding fixed effects constant. The covariance matrix of the random effects is $G = diag(\tau^2, \gamma^2, \ldots, \gamma^2)$, a diagonal matrix with dimension $J+1$. Note that this forces the strong assumption that the random intercept and random time effects are all independent. This is sometimes called a nested exchangeable correlation structure. Allowing random time effects to be correlated may be more realistic in some scenarios, but we leave that extension for future work. Z_i is a JK-by-$(J+1)$ matrix arranged so that the random effects component of Y_{ijk} is $u_i + w_{ij}$.

In cluster trials, the intra-cluster correlation (ICC) is frequently used to summarize the dependence between observations within the same cluster [12]. Note that, in this model, the ICC is $\frac{\tau^2+\gamma^2}{\sigma^2+\tau^2+\gamma^2}$, which represents a ratio of the between-cluster and total variance.

The random treatment effect model

In the random treatment effect model, $a_i = (u_i, v_i)^T$, where u_i is the random intercept effect and v_i is the random treatment effect. The covariance matrix of the random effects is $G = diag(\tau^2, \eta^2)$, a diagonal matrix with dimension two. Note that this forces the strong assumption that the random intercept and random treatment effects are independent. Sometimes random intercept and treatment effects are allowed to be correlated [14], which involves an additional parameter. Although a correlation between random effects may be important in some studies, in this chapter we make the simplifying assumption of independence for

tractability and interpretability of results. Z_i is a JK-by-2 matrix arranged so that the random effects component of Y_{ijk} is $u_i + X_{ij}v_i$.

In this model, ICC is not well-defined because clusters have different correlations while on control and treatment. For convenience, we will define the average cluster correlation (ACC) as $\frac{\tau^2+\frac{1}{JM}(\sum_{m=1}^{M} T_m)\eta^2}{\sigma^2+\tau^2+\frac{1}{JM}(\sum_{m=1}^{M} T_m)\eta^2}$, which represents a ratio of the average of the between-cluster and total variances when $X_{ij} = 0$ and $X_{ij} = 1$, weighted by the proportion of cluster-time spent on treatment. For a classic design where there are an equal number of cluster-periods on and off treatment, the expression simplifies to ACC=$\frac{\tau^2+\eta^2/2}{\sigma^2+\tau^2+\eta^2/2}$. For convenience, we will use ACC to refer collectively to ACC in the random treatment effect model and ICC in the random time effect model. Note that in a random time effect model, since correlation is constant over all clusters and time points, calculating an ACC in an analogous way produces the ICC.

2.2.3 The misspecified models

We are considering two primary cases. In the first case, the researcher chooses the random time effect model, but the true model is actually the random treatment effect model. We call this the 'time-fitted random treatment' case. In the second case, the researcher chooses the random treatment effect model, but the true model is actually the random time effect model. We call this the 'treatment-fitted random time' case. Table 2.1 shows what parameters are fit in each model, and the true parameter values. Throughout, the 't' subscript denotes a true parameter value.

2.3 Convergence of misspecified parameters

The MLE of the parameters in the misspecified model (vector $\hat{\alpha}$ where components depend on the case in question; see Table 2.1) converges to the value α^* that satisfies

$$\lim_{N\to\infty} E_{\alpha_t}[\sum_{i=1}^{MN} \frac{\partial}{\partial\alpha} \log p_\alpha(Y_i|X_i)\Big|_{\alpha^*}] = \vec{0} \tag{2.2}$$

Model	Fixed effects	Random intercept variance	Random time variance	Random treatment variance	Residual variance
Time-fitted random treatment					
Fitted model	$\hat{\Phi}$	$\hat{\tau^2}$	$\hat{\gamma^2}$	0	$\hat{\sigma^2}$
True model	Φ_t	τ_t^2	0	η_t^2	σ_t^2
Treatment-fitted random time					
Fitted model	$\hat{\Phi}$	$\hat{\tau^2}$	0	$\hat{\eta^2}$	$\hat{\sigma^2}$
True model	Φ_t	τ_t^2	γ_t^2	0	σ_t^2

Table 2.1: Parameters used in the fitted ($\hat{\alpha}$) and true (α_t) models for the two cases, where models have been fitted with erroneous random time effects (time-fitted random treatment case) and erroneous random treatment effects (treatment-fitted random time case).

where α_t represents the true values of the (correctly specified) parameters and $p_\alpha(Y_i|X_i)$ is the marginal (misspecified) likelihood for cluster i [22]. Note that under the model described above, $p_\alpha(Y_i|X_i)$ is the same for every cluster within the same sequence, and we assumed that each sequence has the same number of clusters N, so using Y_m to represent an arbitrary Y_i from sequence m we can reduce this equation to:

$$E_{\alpha_t}[\sum_{m=1}^{M} \frac{\partial}{\partial \alpha} \log p_\alpha(Y_m|X_m)\Big|_{\alpha^*}] = \vec{0} \tag{2.3}$$

Thus, the values the misspecified parameters converge to do not depend on the number of clusters per sequence. However, these values can be written as a function of the true parameters (α_t) and elements of the SWT design (e.g. M, J, K). Unfortunately, this system is not restricted to a single, unique root. Finding the roots of this system of equations allows us to determine what the misspecified parameters may converge to, which enables us to assess the variance of the treatment effect estimate under a variety of scenarios when the number of clusters N is sufficiently large (see Section 2.4).

The roots for both misspecification cases are presented below; see supplemental materials (`https://github.com/voldal/SWT_model_misspecification`) for details. Because we are using a linear link and only the random effects are misspecified, the fixed effects are unbiased and consistent [18]. Note that these roots are valid for any reasonable set of fixed effects, including using categorical time adjustment in the mean model instead of linear. For solving the system of estimating equations, some steps were done using Mathematica, version 12 [23].

For the time-fitted random treatment case, closed-form solutions could be found for all the roots of Equation 2.3. We found four different roots which were real and non-negative (see Table 2.2). Note that these roots all have the same total variance (i.e. the denominator of the ACC), and each root is a linear combination of the true variance components. The total variance $\sigma_t^2 + \tau_t^2 + \frac{1}{JM}(\sum_{m=1}^{M} T_m)\eta_t^2$ is an average between the total variance under control and the total variance under treatment, weighted by how many cluster-periods were on treatment vs control.

For the treatment-fitted random time case, two of the roots (Roots 3 and 4) are analogous to the roots from the time-fitted random treatment case and can be written in closed-form (see Table 2.2). However, we were not able to obtain closed-form solutions to the other roots (see supplemental files). From numerical solutions, it appears that there are two other roots, one with$\tau^{2*} = 0$ (Root 2) and one with all components non-zero (Root 1). The components of these two roots are more complicated than the ones from the time-fitted random treatment case. Although Roots 3 and 4 have the same total variance of $\sigma_t^2 + \tau_t^2 + \gamma_t^2$, it is difficult to detect a pattern in the total variance or average total variance of the numerical solutions to Root 1.

In both cases, Root 1 is the most appealing because it does not reduce the desired fitted model and does not have boundary issues like the roots where some components are exactly zero. Roots 2 and 4 are associated with models most scientists would find unsatisfying for an SWT, since they exclude random intercept effects. Root 3 corresponds to a reduced model with only random intercept effects. Although this is a common model used for SWTs, we are considering a scenario in which the scientist was originally interested in fitting the richer model so presumably Root 1 would be preferable over Root 3. Based on the abundance of sophisticated model fitting options, we hope that a researcher would rarely be forced to settle for Root 3. In fact, simulations (see appendix) show that Root 3 is relatively uncommon even when default fitting procedures are used, especially in scenarios where the study design is large enough to appeal to asymptotics; in those cases, Root 1 is by far the most common.

Because of the appeal and prevalence of Root 1, we will focus the rest of the discussion on this root and refer to it as the unreduced root since all components are nonzero. Researchers interested in other roots may modify the code provided (see supplemental files) to obtain results specific to their design and settings.

Time-fitted random treatment			
Root	σ^{2*}	τ^{2*}	γ^{2*}
1	σ_t^2	$\frac{\sum_{m=1}^{M} T_m^2 - \sum_{m=1}^{M} T_m}{(J^2-J)M}\eta_t^2 + \tau_t^2$	$\frac{J\sum_{m=1}^{M} T_m - \sum_{m=1}^{M} T_m^2}{(J^2-J)M}\eta_t^2$
2	σ_t^2	0	$\tau_t^2 + \frac{1}{JM}(\sum_{m=1}^{M} T_m)\eta_t^2$
3	$\sigma_t^2 + \frac{K}{J(JK-1)M}(J\sum_{m=1}^{M} T_m - \sum_{m=1}^{M} T_m^2)\eta_t^2$	$\frac{1}{J(JK-1)M}(-\sum_{m=1}^{M} T_m + K\sum_{m=1}^{M} T_m^2)\eta_t^2 + \tau_t^2$	0
4	$\sigma_t^2 + \tau_t^2 + \frac{1}{JM}(\sum_{m=1}^{M} T_m)\eta_t^2$	0	0

Treatment-fitted random time			
Root	σ^{2*}	τ^{2*}	η^{2*}
1	No closed form	No closed form	No closed form
2	No closed form	0	No closed form
3	$\sigma_t^2 + \frac{K(J-1)}{JK-1}\gamma_t^2$	$\frac{K-1}{JK-1}\gamma_t^2 + \tau_t^2$	0
4	$\sigma_t^2 + \tau_t^2 + \gamma_t^2$	0	0

Table 2.2: Roots of the system of equations in Equation 2.3. In the treatment-fitted random time case, cells marked 'No closed form' indicate values that do not have a simple closed form, but can be found via numerical methods (see supplemental files).

2.4 Impact on treatment effect variance

2.4.1 Calculating variance

For inference on the estimated treatment effect, we are interested in the variance of the estimated fixed effects $\hat{\Phi}$. A simple and wide-spread practice is to use the model-based variance [18]. However, the true variance of $\hat{\Phi}$ is given by the sandwich-based standard error form, which reduces to the model-based variance when the model is correct [24]. Some researchers prefer to use sandwich-based standard errors when there is a sufficiently large number of clusters, since they are consistent regardless of whether the random effects are misspecified [25]. In scenarios where use of a sandwich estimator is appropriate, large differences between the model-based and sandwich estimators can indicate model misspecification [26]. Throughout, we assume that N is large so that we can rely on asymptotic results on parameter convergence.

For a model with fitted covariance matrices of $\Sigma(\alpha)_i = cov(Y_i)$ for clusters $i = 1, \ldots, MN$, the model-based variance estimate is $cov(\hat{\Phi}) \approx [\sum_{i=1}^{MN} X_i^T \Sigma(\alpha)_i^{-1} X_i]^{-1}$. Since we assumed every sequence has N clusters and clusters within a sequence all have the same X_i and Z_i, we can rewrite this as a sum over sequences: $cov(\hat{\Phi}) \approx [N \sum_{m=1}^{M} X_m^T \Sigma(\alpha)_m^{-1} X_m]^{-1}$. Note that this is only possible because the fixed effects in this model do not include any additional cluster-level covariates. That is, all clusters within a sequence have a common mean design matrix.

We will use $var_m(\hat{\theta}_t)$ to denote the model-based variance of the estimated treatment effect from the correctly specified model. That is, the covariance matrices are correctly specified and contain the true parameter values.

We will use $var_m(\hat{\theta})$ for the model-based variance of the estimated treatment effect from the misspecified model. This involves the limit of the covariance matrices from the misspecified model, which are functions of the true parameter values. For example, in the treatment-fitted random time case, G will be a 2-by-2 diagonal matrix but instead of plugging in $\hat{\tau^2}$ and $\hat{\eta^2}$, we will plug in the roots τ^{2*} and η^{2*}, which are functions of the true

parameters σ_t^2, τ_t^2, and γ_t^2 (see Table 2.2).

We will use $var_s(\hat{\theta})$ to denote the true variance of the estimated treatment effect from the misspecified model ('s' for sandwich-form, as opposed to 'm' for model-based). In the misspecified model, the true variance has a sandwich form $cov(\hat{\Phi}) = A^{-1}BA^{-1}$, where $A = N\sum_{m=1}^{M} X_m^T\Sigma(\alpha)_m^{-1}X_m$ (see above), and $B = N\sum_{m=1}^{M} X_m^T\Sigma(\alpha)_m^{-1}\Sigma(\alpha_t)_m\Sigma(\alpha)_m^{-1}X_m$. Since the model is misspecified, we use the limit of the covariance matrices again for $\Sigma(\alpha)_m$, and the true covariance matrices for $\Sigma(\alpha_t)_m$. Note that if the model is correctly specified, this expression simplifies and $var_s(\hat{\theta}) = var_m(\hat{\theta}_t)$.

2.4.2 Ratios of variances

Note that in both the model-based and sandwich-form variance estimators, the only place N appears is as a multiplier of $\frac{1}{N}$ of the whole term, so the ratio of any two of these variances does not depend on N. That is, the multiplicative error of the misspecified model-based variance estimate is constant regardless of the number of clusters per sequence. Because it simplifies discussion of design choice, we will focus on multiplicative differences between variance estimates.

To examine the validity of misspecified model-based variances, we will use $var_m(\hat{\theta})/var_s(\hat{\theta})$. If this ratio is greater than one, it means that using $var_m(\hat{\theta})$ would result in conservative inference. If this ratio is less than one, using $var_m(\hat{\theta})$ is anti-conservative.

To examine how much efficiency is lost by choosing the incorrect model, we will use $var_m(\hat{\theta}_t)/var_s(\hat{\theta})$. Values much smaller than one indicate large losses of efficiency. In situations where it is appropriate to apply the sandwich-form estimator, we would expect this ratio to never be above one.

2.4.3 Results

Plots and calculations were done with `R`, version 3.6.1 [3].

We focus on classic designs for simplicity, since they can be completely described by only M and K. Recall that the validity and efficiency ratios do not depend on N, so these results

hold for any number of clusters per sequence.

In the time-fitted random treatment case, we found that the misspecified model-based variance $var_m(\hat{\theta})$ was conservative for some designs, and anti-conservative for others (Figures 2.1 and 2.2). The validity moved further from one as the number of observations per cluster-period (K) increased. For the time-fitted random treatment case, the trends in validity are dramatically different for a design with two sequences versus a design with more than two sequences, so we address these two scenarios separately.

For a design with two sequences in the time-fitted random treatment case, $var_m(\hat{\theta})$ was conservative for all the scenarios examined here. The validity was worse with high ACC, but trends related to the relative size of random intercept effects vs. treatment random effects were unintuitive and dependent on the ACC (Figure 2.2).

For designs with more than two sequences in the time-fitted random treatment case, $var_m(\hat{\theta})$ was anti-conservative for all the scenarios examined here, and would therefore lead to inflated Type I error. In the scenarios we examined, $var_m(\hat{\theta})$ was the worst for: larger number of sequences; higher ACC; and larger contribution of η_t^2 to the average between-cluster variance (Figures 2.1 and 2.2).

For the treatment-fitted random time case, we found that the misspecified model underestimated the variance of the treatment effect in all the scenarios we examined (Figures 2.1 and 2.2). The validity moved further from one as the number of observations per cluster-period (K) increased. For the treatment-fitted random time case, validity of $var_m(\hat{\theta})$ was worst for: smaller number of sequences; higher ACC; and larger contribution of γ_t^2 to the average between-cluster variance.

In both cases, for the scenarios examined in Figure 2.1 (classic designs with $2 \leq M \leq 7$, $K \leq 200$, and ACC=0.03), the loss of efficiency from using the misspecified model with a robust variance was not more than 5% (see appendix). Also, in both cases efficiency worsens as K increases. For an ACC of 0.03, loss of efficiency is not very significant; however, in some scenarios efficiency can be impacted dramatically (Figure 2.3). For the time-fitted random treatment case, efficiency is worse for higher ACC and larger η_t^2 when there were two or

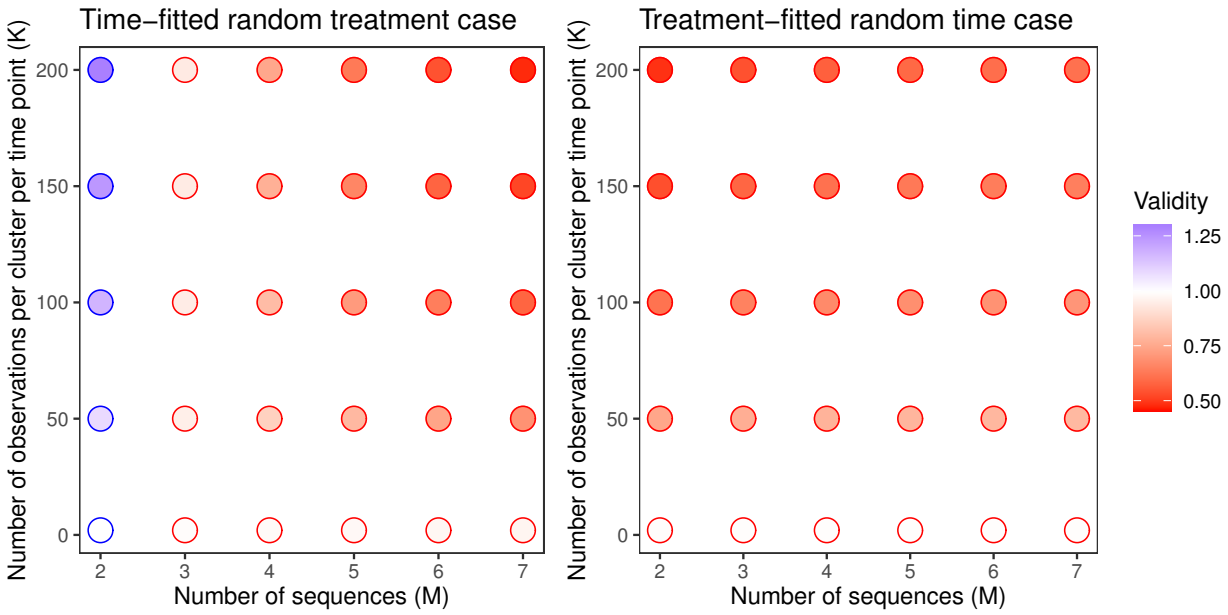

Figure 2.1: Validity $(var_m(\hat{\theta})/var_s(\hat{\theta}))$ of Root 1 for both cases, for a variety of classic designs. Ratios above 1 (indicated by blue outlines) correspond to conservative $var_m(\hat{\theta})$ estimates. Ratios below 1 (indicated by red outlines) correspond to anti-conservative $var_m(\hat{\theta})$ estimates. For each case, $\sigma_t^2 = 5$ and $\tau_t^2 = 0.1$. To keep the ACC consistent, we chose $\eta_t^2 = 0.1$ and $\gamma_t^2 = 0.05$ for the time-fitted random treatment and treatment-fitted random time cases, respectively.

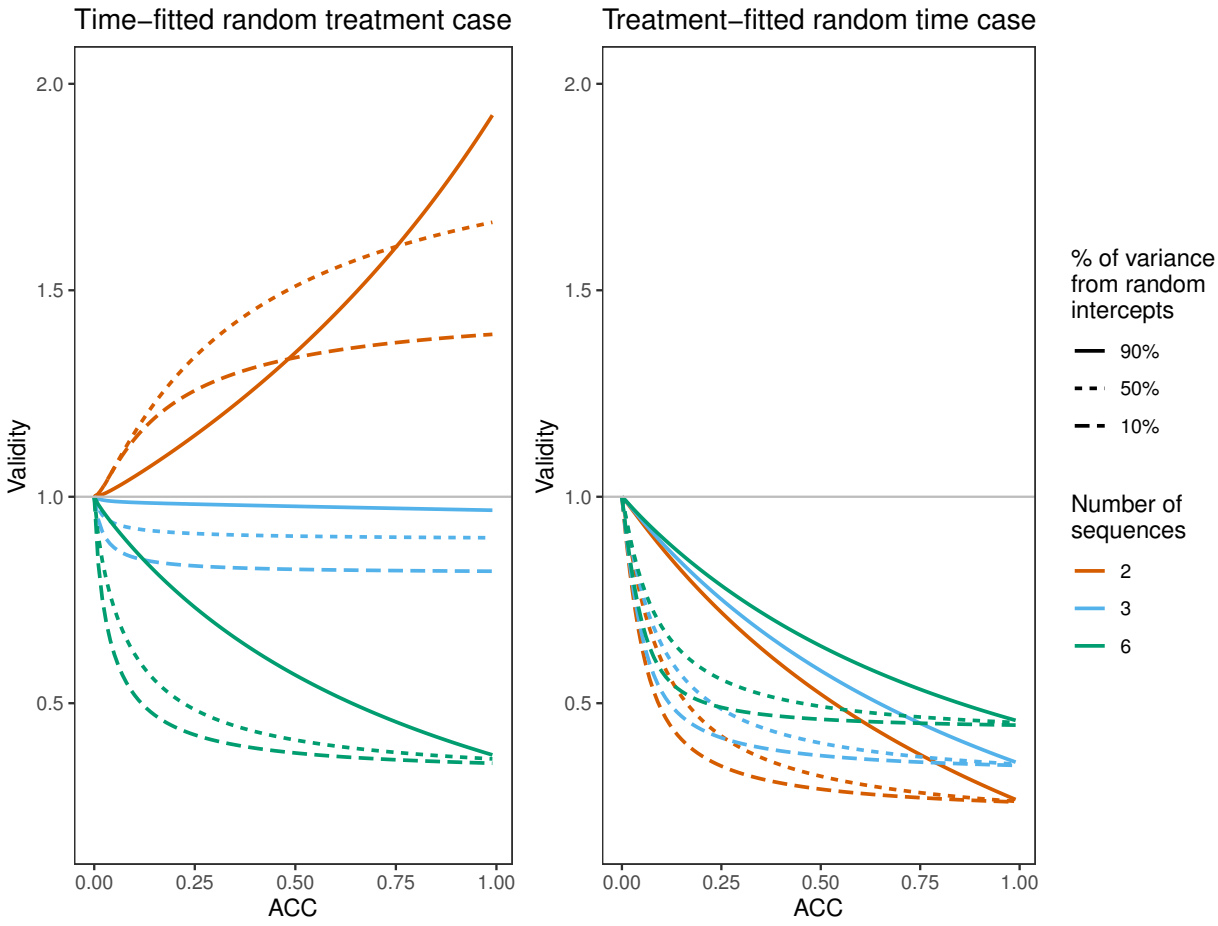

Figure 2.2: Validity ($var_m(\hat{\theta})/var_s(\hat{\theta})$) of Root 1 for both cases, for three designs and a variety of true ACCs. The three classic designs ($M = 2$, $M = 3$, and $M = 6$ sequences) each have $K = 20$ observations per cluster per time period. Throughout, $\sigma_t^2 = 1$. Each ACC is achieved in three different ways by adjusting the balance of τ_t^2 and γ_t^2 or $\eta_t^2/2$. If γ_t^2 or $\eta_t^2/2 = \tau_t^2$, 50% of the average between-cluster variance (the numerator of the ACC) comes from random intercepts. Similarly, if γ_t^2 or $\eta_t^2/2 = \tau_t^2/10$, then 90% of the variance comes from random intercepts and if γ_t^2 or $\eta_t^2/2 = \tau_t^2 * 10$, then 10% of the variance comes from random intercepts.

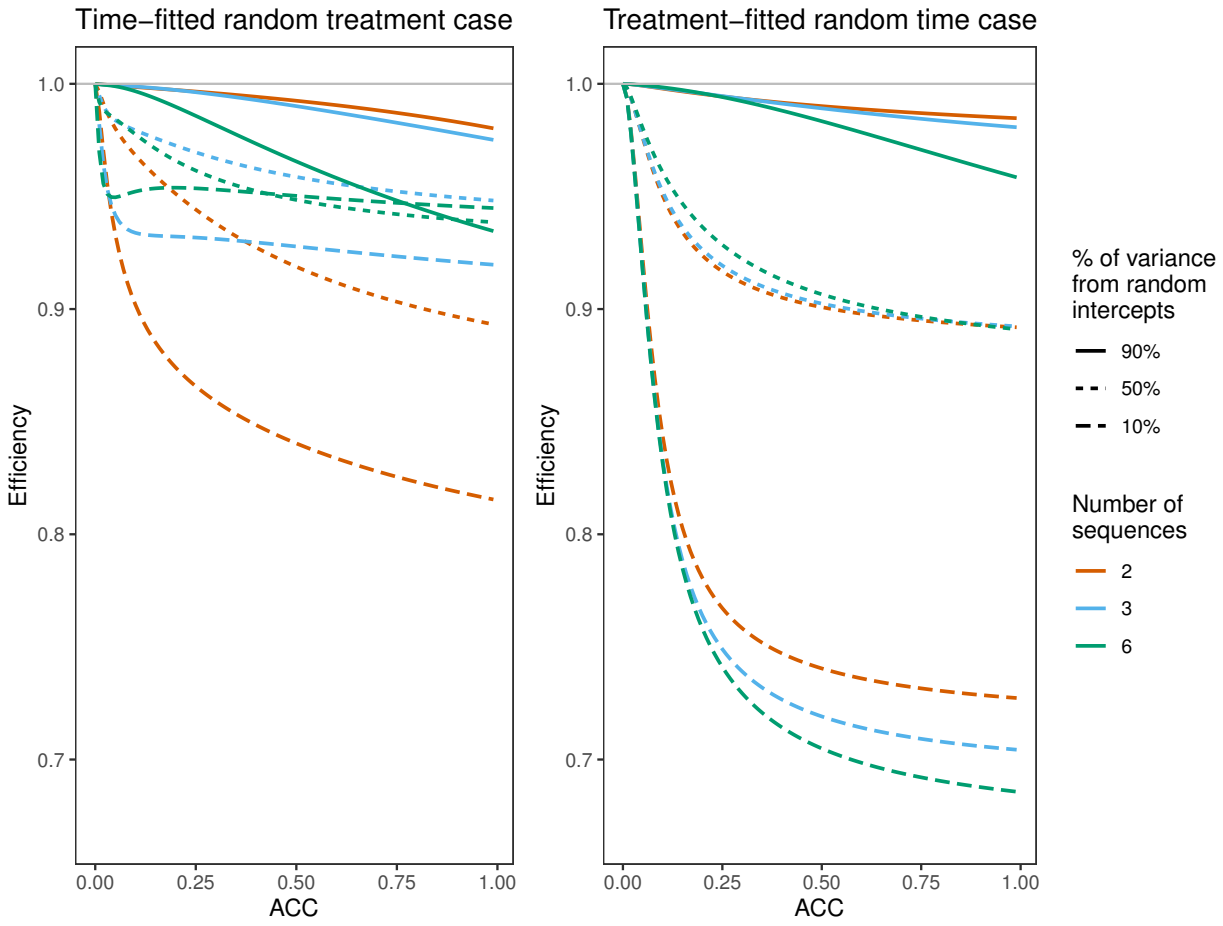

Figure 2.3: Efficiency $(var_m(\hat{\theta}_t)/var_s(\hat{\theta}))$ of Root 1 for both cases, for three designs and a variety of true ACCs. The three classic designs ($M = 2$, $M = 3$, and $M = 6$ sequences) each have $K = 20$ observations per cluster per time period. Throughout, $\sigma_t^2 = 1$. Each ACC is achieved in three different ways by adjusting the balance of τ_t^2 and γ_t^2 or $\eta_t^2/2$. If γ_t^2 or $\eta_t^2/2 = \tau_t^2$, 50% of the average between-cluster variance (the numerator of the ACC) comes from random intercepts. Similarly, if γ_t^2 or $\eta_t^2/2 = \tau_t^2/10$, then 90% of the variance comes from random intercepts and if γ_t^2 or $\eta_t^2/2 = \tau_t^2 * 10$, then 10% of the variance comes from random intercepts.

three sequences. For the six-sequence design, these trends only hold for some ranges of ACC and/or η_t^2. For the treatment-fitted random time case, efficiency is worse for higher ACC and larger γ_t^2.

The relative loss of efficiency for the two cases depends on the study design and values of the true variance components. However, in the scenarios considered here the treatment-fitted random time case had the largest potential losses in efficiency (Figure 2.3).

For non-classic designs, it is very difficult to predict trends in validity, and even (for the time-fitted random treatment case) whether the misspecified model is conservative or anti-conservative (see appendix). For some non-classic designs, whether the misspecified model is conservative or anti-conservative even depends on the ACC (see appendix). Researchers wondering about a specific design can use the roots in Table 2.2 (or system of equations in the supplemental files, for the treatment-fitted random time case Roots 1 and 2) to calculate validity and efficiency for their exact design. These roots hold for many non-classic designs, but do rely on the assumption that the number of clusters per sequence and number of observations per cluster-period are constant.

These results have demonstrated some broad trends in validity and efficiency and have shown that the impact of misspecification can be very severe. Although we examined only classic designs here, it is clear that the precise design of a SWT plays a key role in how misspecification affects inference. For this reason, it is important for researchers to examine these effects for their specific SWT design. See supplemental files for `R` code which can be used to perform these calculations, which can be easily adapted for most SWT designs, alternative roots, and other ways of modeling time. Other fixed effects can also be included, although consideration must be given to the asymptotic nature of these results.

2.5 Example

We will use a study reported by Haines et al. [21] as an example of how these results might be used. Researchers conducted a SWT in two metropolitan teaching hospitals in Australia. The study involved 14,834 patients clustered into 12 hospital wards. Researchers

were interested in the effect of removing weekend allied health services from wards, which included physical therapy, social work, and other patient services. Researchers listed several outcomes of interest, but for the purposes of this example we will focus on log-transformed length of stay in days. Start times at the two hospitals were slightly different, but for simplicity we will disregard this so that we are considering a classic six-sequence SWT with two clusters in each sequence. We will also assume that there was no hospital-level clustering, although it would be straightforward to add hospital as a fixed effect.

Haines et al. [21] used mixed-effects models, but focused on traditional nested mixed effects (i.e. hospital and ward effects). Suppose that we are conducting a post-hoc exploratory analysis which also considers random time and treatment effects. Using the data published by Haines et al. [21], we attempt to fit a full model with random intercept, time, and treatment effects in `R`. Unfortunately, we find that the fitted random treatment variance is zero, and additional warning messages suggest that the fitted values may not be very accurate. For the purposes of this example, suppose we have followed the suggestions of Cheng et al. [19] without success (see supplemental files for details), and that there is no previously reported study which can be used to inform the choice of random effects in this particular setting.

To improve model fit, we abandon the full model and fit two reduced models: a random time effect model, and a random treatment effect model. Both models produce similar estimates of the random intercept standard deviation (SD) τ (0.28) and residual SD σ (1.02 vs. 1.03 for the random time and random treatment models, respectively). In the random time model, the estimated random time SD γ is 0.12. In the random treatment model, the estimated random treatment SD η is 0.10. Unfortunately, these two models also produce different estimates of the treatment effect (0.13 vs. 0.10) and model-based treatment effect standard error (0.05 vs. 0.04). We might be inclined to use the random time model, since that was favored when fitting the full model. However, information on which model is more robust to misspecification might influence our decision. Quantifying the impact of potential misspecification may also affect our confidence in our results.

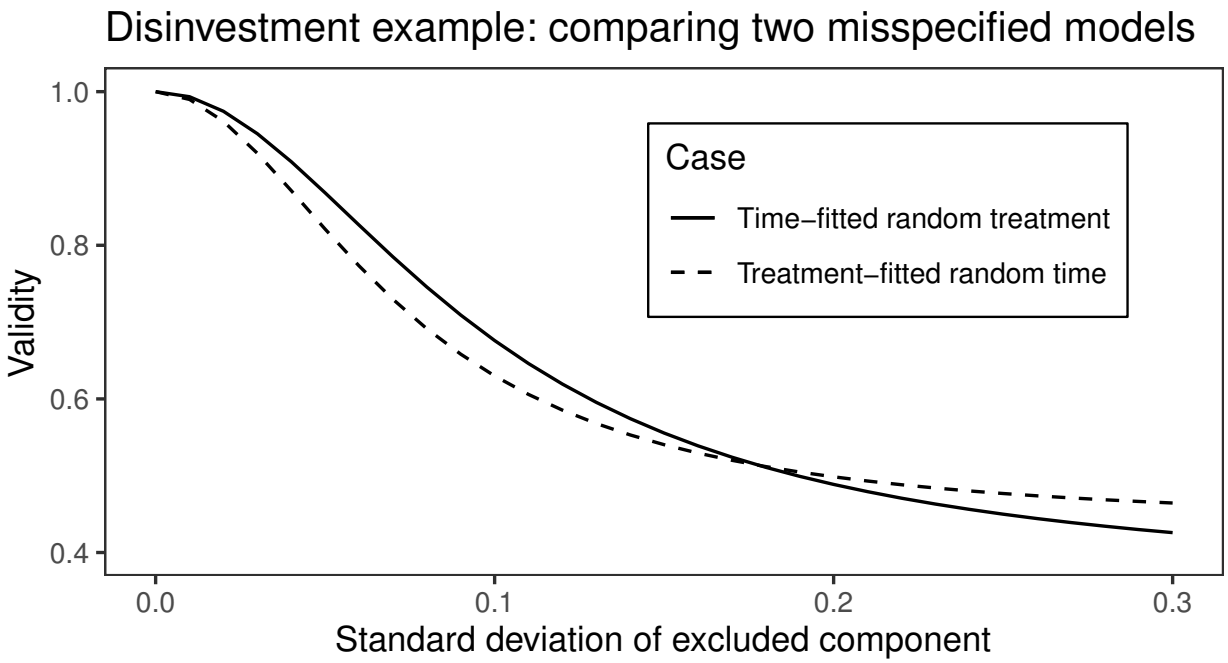

Figure 2.4: Validity of the two proposed models in the disinvestment example if they were both misspecified. Throughout, $\tau_t = 0.28$ and $\sigma_t = 1.03$ were fixed, based on their average fitted values from the two models. We assumed there were $K = 177$ observations per cluster per time period, which was the average K across the trial.

Figure 2.4 shows the impact on validity of the two models considered in this example. Using the results of the reduced models in which both random time and treatment SDs were around 0.1, it is clear that choosing the wrong model in this scenario could have a dramatic effect on the validity of our conclusions, underestimating the variance of the treatment effect by around 35% asymptotically. The impact of incorrectly excluding a random time effect is slightly more extreme than the impact of incorrectly excluding a random treatment effect. These results support the choice of the random time effect model. However, because the effect of misspecification may be large in this scenario, these results also suggest that scientific conclusions from the random time effect model should be made with caution. We did not examine trends in efficiency for this example, since calculating a sandwich variance based on only 12 clusters may not be advisable [25].

2.6 Discussion

In this chapter, we have explored how choosing the wrong random effect in a SWT mixed model analysis can affect the estimated variance of the treatment effect. Since our method relies on analytical solutions instead of simulations, we were able to precisely study a vast landscape of study designs and settings. We focus here on the results for scenarios most relevant to a SWT: small ACC (< 0.25) and the variance of the excluded random effect makes up a small portion of the average between-cluster variance. Some unsurprising trends hold across both cases. Validity and efficiency tended to be closer to 1.0 for smaller ACC and smaller variance of the excluded random effect. Other trends were more unexpected. Validity and efficiency both worsen as the number of observations per cluster-time period K increases. When the true model includes a random treatment effect but the researcher incorrectly fits a random time effect (time-fitted random treatment case), the number of sequences in a classic design has a dramatic impact on trends in both validity and efficiency. Our exploration of efficiency showed that in most cases, the cost of using the misspecified model with sandwich-based variance is relatively small. Although it was not the focus of this chapter, researchers might consider using robust variance estimates when there are a

sufficient number of clusters.

We used analytical solutions instead of simulations, which avoided some key issues. In addition to the obvious computational burden and variability of results, simulation studies in these types of settings can struggle significantly with the convergence of fitted models. This reduces efficiency and could bias results. Using analytical solutions also allowed us to gain more insight into the details of the model misspecification. For example, if we had used simulations we would not have detected the issue with multiple roots.

Through our motivating example, we have demonstrated how our methods may be used as a sensitivity analysis to assess robustness to misspecification without the use of simulations. Ideally, model selection including the choice of random effects would be informed primarily by scientific beliefs and analyses of similar data. However, when choosing random effects these may not be readily available, forcing researchers to rely more heavily on other considerations such as robustness to misspecification. For a large trial, if researchers have already committed to using robust standard errors then it may be helpful to focus on efficiency as a metric for model choice instead of validity.

In our motivating example, we used the data to inform assumptions about values of true variance parameters. To use this method for creating a pre-specified analysis, assumptions about true variance parameters would be based on pre-existing data [27] and scientific beliefs instead of being estimated from the data of interest. This may sometimes be difficult, but we hope researchers can draw on their experience in estimating power for SWTs, since the methods of making assumptions about variance components before data collection should be similar. If some parameters are particularly difficult to estimate, it is easy to consider a range of potential values as we did in our example. In settings with abundant pre-existing data from similar studies, it may be possible to use that data to directly test for the existence of specific random effects.

The practical impact of the existence of multiple asymptotic solutions is unclear. Since these are asymptotic results and simulations suggest that the unreduced root (Root 1) becomes more common as the number of clusters increases, we believe that our focus on the

unreduced root is appropriate for studies that are large enough to appeal to asymptotics. We informally examined some simulated datasets to determine whether multiple roots might exist for a single dataset. Although this was not an exhaustive analysis, we found no evidence of multiple roots within a dataset, so we hypothesize that this is a between-datasets issue rather than within-datasets. To see whether the issue of multiple roots is unique to misspecified models, we examined the behavior of a correctly specified model using the same methods described in this chapter. In the few cases we checked, the correctly specified model equations also had four potential roots following the same pattern as the four roots identified in misspecified models, with only one root having all nonzero components. Although solutions on the boundary corresponding to roots 2, 3, and 4 occur in correctly specified models, consistency of the MLE suggests that only Root 1 is asymptotically relevant when the model is correctly specified.

These results cannot be applied directly to simulations done by Thompson et al. [17] since those simulations were for a cohort SWT with binary outcomes. However, our observations suggest that the specific SWT design can have a dramatic impact on results, and even the direction (conservative vs. anti-conservative) of validity in the time-fitted random treatment case. Thus, researchers concerned about misspecification in cohort SWTs with binary outcomes may want to conduct their own simulations similar to those done by Thompson et al. [17] instead of relying on results that may not generalize to other SWT designs.

Some of the assumptions about the mixed models we considered here may be particularly restrictive. Because we focused on a Normal outcome with a linear link, our marginal and conditional fixed effects are equivalent and unbiased. With other link functions, marginal and conditional effects may differ, and the estimated treatment effect may be biased when random effects are misspecified. The assumption that all random effects are independent also has important implications. In particular, assuming that the random treatment effect is independent of the random intercept effect implies that variability of the outcome is higher in the treated time periods compared to control. Since these results about efficiency and validity are driven by nuanced differences in random effect structures, the exact correlation

structure may be important. In addition to adding correlation between the random intercept and other effects, researchers might consider the many ways that random time effects could be related within a cluster. The model we considered here is probably unrealistic in most scenarios, since it implies that the correlation between two individuals from the same cluster but different time periods does not depend on how far apart those time periods are. Other more complex random time effect models (e.g. exponential decay) might be more realistic. Researchers could even consider a scenario where the true model is the full model, with correlations between random treatment and random time effects as well. Because of the complexity and diversity of these extensions, we have not addressed them in this chapter. However, we have provided a Mathematica file that demonstrates how to obtain solutions for some of these extended cases. Alternatively, researchers could turn to simulations like Thompson et al. [17], who did allow for correlation between some random effects. Last, the fixed effects we considered were very minimal; in particular, we used a linear model for the fixed time effect, whereas it may often be important to use a more flexible model for time. If a model for time where the number of parameters depends on the study design (e.g. modeling time as categorical with $J-1$ indicator variables) is used, the trends relating validity or efficiency to the number of sequences may be different. Researchers familiar with `R` can modify the functions provided in the supplemental files in order to account for additional fixed effects or a more flexible time model.

We hope that scientists struggling to choose between random time effects and random treatment effects can use these results to understand how model choice might impact the validity and efficiency of inference on the treatment effect. Although there is no universal 'correct' choice, for some scenarios validity and efficiency are dramatically different between the time-fitted random treatment case and the treatment-fitted random time case. Particularly when supplementary data and pre-existing scientific beliefs are weak, the methods presented in this chapter can be an important tool for developing statistical analysis plans and performing sensitivity analyses.

Acknowledgments

Research reported in this paper was supported in part by the National Institutes of Health under award number AI29168. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data accessibility

The code that supports the findings of this paper are available at `https://github.com/voldal/SWT_model_misspecification`. The data from the disinvestment example are openly available at `https://doi.org/10.1371/journal.pmed.1002412` [21].

Chapter 3

ROBUST ANALYSIS OF STEPPED WEDGE TRIALS USING COMPOSITE LIKELIHOOD MODELS

3.1 Introduction

Stepped wedge trials (SWTs) are a type of cluster randomized trial in which every cluster typically begins in the control state, and crosses over into the treatment condition at some point during the trial [2]. A group of clusters following the same cross-over schedule is called a sequence or wave. Although observational SWTs exist, typically clusters are randomized into sequences, and each sequence's cross-over time is pre-specified.

To account for correlation within clusters, researchers frequently use mixed models to estimate treatment effects in SWTs [1]. However, mixed models are susceptible to misspecification in both the mean model via fixed effects and the correlation structure via random effects. The specification of the underlying time trend is a particular concern; since time and treatment are confounded in a SWT, misspecification of the time trend can bias treatment effect estimates. Many researchers recommend using a flexible model for time [6], but this assumption may still be violated if there is not a common time trend (e.g. if time trend varies between geographic regions). Validity of inference from a mixed model also depends on correctly specifying the random effects structure; these random effects may be difficult to anticipate when pre-specifying an analysis, and selecting the wrong random effects can cause inference to be either conservative or anti-conservative (see e.g. the previous chapter).

In response to these concerns about misspecification, many researchers have proposed variations on the usual models for analyzing SWTs. Random effects in mixed models range from very simple [6] to quite complex, e.g. allowing the time trend and treatment effect to vary by cluster [14, 17] with more realistic correlation between time periods [28]. Although

more complex random effects may be more effective at representing realistic data-generating models, they are still susceptible to misspecification. Even in the rare case that a researcher has confidence in the form of the data-generating random effects, it may be computationally difficult to fit a very complex model if there are not many clusters. In response to concerns about misspecification in mixed models, some researchers have proposed extensions that are more robust. For example, some methods allow for a more flexible semi-parametric random effects model [29]. Generalized estimating equation (GEE) models are sometimes an attractive alternative because valid inference is not contingent on correctly specifying the entire correlation structure, although assumptions about the time trend are still important to ensure an unbiased treatment effect estimate. However, GEEs generally require a large number of clusters, while SWTs are often designed with very few clusters [30]. Small-sample adjustments for GEEs have been proposed, but can be complicated and performance depends on attributes of the SWT design [25, 31].

Mixed models are susceptible to misspecification in part because they use both within-period and within-cluster information about the treatment effect. Within each time period, contrasting the clusters randomized to treatment vs. control during that particular time period mimics a traditional parallel cluster randomized trial. As a result, concerns about confounding of treatment and time do not apply to within-period contrasts. These within-period treatment effect estimates are often called 'vertical' estimators. In contrast, within-cluster treatment effect estimates rely on comparing a cluster's outcomes before and after cross-over; this is sometimes called a 'horizontal' estimator. To obtain a valid treatment effect estimate from horizontal contrasts, researchers need to adequately adjust for underlying time trends, and account for correlation between observations coming from the same cluster at different time periods. Treatment effect estimates from mixed models can be thought of as a weighted average of the vertical and horizontal treatment effect estimates, with more weight generally given to horizontal contrasts [32]. An estimator that uses only vertical contrasts would generally be more robust to misspecification of time trends and correlation structure, but less efficient than a correctly specified mixed model that incorporates horizontal contrasts.

A number of robust methods have been proposed that use only vertical information in stepped wedge trials. Hughes et al. [33] developed closed-form estimators of the treatment effect, based on vertical contrasts and permutation of clusters over sequences. Thompson et al. [34] developed a non-parametric within-period (NPWP) estimator, using permutation tests for inference. Kennedy-Shaffer et al. [35] sought to improve the efficiency of the NPWP by using synthetic controls based on each cluster's history of outcomes. Both the NPWP and NPWP with synthetic control estimators were more robust than a mixed model, but less efficient. Moreover, both the NPWP and cross-over methods use permutation for inference, which tests for equivalence of distributions; that is, if the treatment affects either the mean or the variance of the outcome, these methods may return a small p-value [36]. This is called a strong or sharp null, since under the null hypothesis the treatment has no impact on the outcome. In the context of SWTs, researchers are almost certainly interested in identifying whether treatments affect the mean of the outcome, i.e. a weak null. So, if there is a random treatment effect, inference using these non-parametric methods may be undesirable.

Although these methods have demonstrated the advantages of vertical methods in terms of robustness, for practical application it is important to explore methods for improving efficiency. We propose a semi-parametric method for obtaining a vertical treatment effect estimate that allows researchers to improve efficiency by adding some horizontal information in line with best practices for cluster-randomized trials, and has a closed-form asymptotic option for inference, which is robust to many common sources of covariance and tests the more familiar weak null. To account for the complicated correlation structure underlying the set of all vertical contrasts, we use composite likelihood (CL) models. CL models are more robust than their traditional full likelihood counterparts, as assumptions need to be made on small pieces of the likelihood instead of the entire joint distribution [37]. A CL is constructed by multiplying together many component likelihoods, which consist of conditional or marginal densities [37]. Each component likelihood typically models a specific piece of the data (e.g. an observation, or all the observations from a cluster), but the form varies widely and usually reflects the context or underlying data structure. Because the CL is the product of component

likelihoods which may not be independent, the CL can be thought of as a misspecified model. For inference, robust methods are used with a 'working' independence structure, so that deviations from the working correlation structure affect efficiency but not validity [37]. The CL framework is very flexible and robust to many types of deviations from typical data-generating models, but the familiar parametric mean model of the CL may offer an opportunity to improve on the efficiency of fully non-parametric methods such as the NPWP estimator. In this paper, we propose a set of estimators using CL methods and compare them to other robust methods and more standard mixed models and GEE.

3.2 Methods: background and existing methods

3.2.1 Stepped wedge design

Although many variations on the SWT exist, for simplicity we restrict our attention to 'classic' designs, in which every cluster starts on control and ends on treatment, and one sequence crosses over between every time point (e.g. [6]). These methods would extend to deviations from the classic design (e.g. extra follow-up time after all clusters have crossed over), but some of these factors may also affect the relative performance of different methods (e.g. by affecting the relative amount of information in vertical vs. horizontal contrasts). Similarly, although these methods would be applicable to a cohort SWT, we focus on cross-sectional designs, in which each individual is observed only once. Assume our SWT has M unique sequences, each with N clusters, for a total of $M * N$ clusters indexed by $i = 1, \ldots, M * N$. Because of the cross-over schedule, there will be $M + 1$ time points, which we will index with $j = 1, \ldots, J$ for convenience. Assume that the number of individuals in each cluster at each point in time is constant, and call this K; see appendix for extension to a varying cluster size.

We assume outcomes are Normally distributed, and like Thompson et al. [34] we will focus on cluster-period level summaries. Note that this is an important restriction, as some methods (e.g. mixed models, GEE, CL) can be used at either the cluster-period level or the

individual level [38], while others are more specific to the cluster-period level (e.g. NPWP). The importance of this restriction varies; for example, if there are important individual-level covariates, it may be more desirable to use individual-level data. Throughout this paper, we use a linear link, assume that clusters are equally sized, and either share a common fixed treatment effect, or have a random treatment effect that follows a Normal distribution. As a result, the estimand targeted by a cluster-level and individual-level analysis are the same. Although we do not consider it in this paper, in some situations the distinction between the treatment effect on clusters and the average treatment effect over the entire population of individuals is more impactful, and researchers may need to take this into consideration when selecting a method. Extension of CL methods to binary and other non-Normal outcomes is important for SWTs, but left to future work.

Let Y_{ij} denote the cluster-period mean (over the K individuals from this cluster-period) for cluster i at time j. Then we can represent our data via a mixed model-style notation:

$$Y_{ij} = \vec{X}_{ij}\Theta + \vec{Z}_{ij}a_{ij} + \epsilon_{ij} \tag{3.1}$$

Here, $\vec{X}_{ij}\Theta$ represents the mean model which includes the treatment effect θ and some model for time. Throughout, we will use an unordered categorical model for time when fitting models, to reduce the risk of residual confounding between time and treatment. $\vec{Z}_{ij}a_{ij}$ represents the random effects; for a SWT, this would typically include a random cluster effect, but may also include random effects for time, treatment, and correlation between those effects. We assume $\epsilon_{ij} \sim N(0, \sigma^2)$, so σ^2 is the residual variance for the cluster-period means, which is equal to the individual-level residual variance divided by K. Typically, the residual variance is assumed to be independent, so $\epsilon_{i1}, \ldots, \epsilon_{iJ} \sim MVN(\vec{0}, \sigma^2 I_J)$ for each cluster i, where I_J is an identity matrix of dimension J. However, researchers may sometimes be concerned about residual correlation between cluster-periods, and allow the covariance of $\epsilon_{i1}, \ldots, \epsilon_{iJ}$ to be e.g. autoregressive [39]. Note that we have assumed that the residual variance is constant both over time and between clusters. However, this does not imply that the total variance is constant over time. For example, if there is a random treatment effect

the total variance can be different in treated cluster-periods and untreated cluster-periods.

3.2.2 Traditional methods

We wish to compare our proposed methods to traditional methods used to analyze SWTs. We include some basic versions of mixed models and GEE in this paper as a benchmark of efficiency from a typical analysis, but acknowledge that model performance varies widely even within the mixed models or GEE family, with many more sophisticated options available to researchers analyzing SWTs. It would be important for future research to compare more sophisticated extensions of these traditional methods to the proposed methods. For example, methods for applying GEE to small samples or fitting mixed models that are more robust to assumptions about the correlation structure.

For consistency, we apply these traditional methods using cluster-period means instead of individual-level data, although this does not necessarily reflect common use. In the setting we consider here with constant cluster-period sizes and a Normal outcome with a linear link, Grantham et al. [28] suggest that mixed models using cluster-period summaries are as efficient as individual-level models as long as within-period observations are exchangeable. Similar properties hold for GEE's [38].

It is important to note that mixed models estimate a conditional treatment effect, whereas other methods presented in this paper estimate a marginal treatment effect (e.g. GEE, CL models). Because the outcome is Normal with an identity link, the marginal and conditional models have the same target of inference [40].

Simple mixed model

This method estimates the treatment effect θ with the following model:

$$E[Y_{ij}|u_i] = \mu + \beta_j + X_{ij}\theta + u_i \tag{3.2}$$

This simple mixed model, first proposed by Hussey and Hughes [6], includes only a ran-

dom cluster intercept effect (u_i), which is assumed to follow a Normal distribution $N(0, \tau^2)$. The fixed time trend is represented by $\beta_1, \ldots, \beta_J$ with $\beta_1 = 0$ for identifiability, and μ is the fixed intercept. Throughout, X_{ij} is an indicator of whether cluster i is receiving treatment at time j, with $X_{ij} = 0$ for control and $X_{ij} = 1$ for treatment.

Inference can be based on asymptotics following the usual procedure, or based on permutation [41, 42]. Although permutation tests are robust to some types of misspecification (e.g. misspecified distribution of random cluster effects), in this paper we use the standard asymptotic standard error (SE) estimates for inference on mixed models. We use the Satterthwaite approximation for obtaining p-values from mixed models [43]. Note that if the correlation structure is misspecified by choosing the wrong random effects, we would expect asymptotic-based inference to be invalid, but since we are using a linear link we would expect the point estimates to be unbiased. In a mixed model using a nonlinear link, misspecification of random effects can bias the estimated treatment effect.

Note that although this model does not explicitly contain any random effects for time, because we aggregated the data to cluster-period means, the mixed models in Equation 3.2 and Equation 3.3 are already consistent with a random cluster-period intercept (e.g. as implemented in the previous chapter). Although we do not consider them in this paper, other types of random time effects might be identifiable with cluster-period data, e.g. random effects that cause cluster-periods to be more strongly correlated if they are closer together in time [39, 28].

Mixed model with random treatment effect

This mixed model builds on the previous model (Equation 3.2) by adding a random effect for treatment, allowing for a more complicated correlation structure. It includes both a random cluster intercept effect (u_i) and a cluster-specific random treatment effect (v_i).

$$E[Y_{ij}|u_i, v_i] = \mu + \beta_j + X_{ij}\theta + u_i + X_{ij}v_i \tag{3.3}$$

where $[u_i, v_i] \sim N([0,0], diag(\tau^2, \eta^2))$. Note that this model assumes that the random cluster intercepts and random treatment effects are independent, but it may sometimes be desirable to allow for correlation between the random effects.

GEE with working independence

We use the same mean model as above, and a working independence model within clusters.

$$E[Y_{ij}] = \mu + \beta_j + X_{ij}\theta \tag{3.4}$$

Here, the working covariance matrix for a single cluster i is $Var(Y_{i1}, \ldots, Y_{iJ}) = \sigma^2_{GEE} I_J$, where I_J is an identity matrix of dimension J. For inference, we use robust standard errors. As a result, uncertainty about the within-cluster correlation structure does not affect the validity of the inference from this model.

In both GEE and the mixed models above, we have modeled time in the mean model using a very flexible categorical time model. It is important to fully adjust for time in a SWT, since time and treatment are confounded. If researchers are confident that the time trend is common to all clusters, then a simpler model for time (e.g. linear trend) could be considered to improve efficiency, at the risk of some bias if it is not correct. Because most SWTs have a reasonably small number of time periods, the cost of using a categorical time model may often be negligible. If the time trend varies randomly between clusters but is centered around some common mean time trend (e.g. a mixed model with a random cluster-period effect), this is easy to account for. It is more difficult to account for a time trend that differs between clusters and is not centered around a common mean time trend. We focus on a setting in which there are two strata of clusters, and each stratum has its own distinct time trend; it would be difficult to account for this with random effects, since the distribution of a random effect would be bimodal, not grouped around a common mean time trend. If researchers can identify strata a priori, modeling stratified time trends is a bias-variance tradeoff. For example, if hospitals are the unit of randomization and hospitals are being recruited from

two larger healthcare systems, researchers might suspect that the time trend could differ between healthcare systems. In this case, researchers can fit a model with an interaction between time and healthcare system to ensure that time trends have been accounted for, and even stratify randomization by healthcare system. Unfortunately, it may not always be possible for researchers to account for time trends that differ systematically between clusters. For example, based on pre-existing data researchers may incorrectly assume a common time trend, and not account for it in the design or model. If researchers are considering a time-varying treatment effect, there may be identifiability issues with complex time trend models [44]. If the time trend varies systematically by strata but the model assumes a common time trend, this could compromise the variance estimates for both mixed models and GEE (if the number of clusters is not large enough) using the methods described above. If the time trends happen to be balanced over sequences by chance, then we would still expect treatment effect estimates from mixed models or GEE to be unbiased. However, if time trends are correlated with sequence by chance, then there is residual confounding between time and treatment after adjusting for a common time trend. This may bias point estimates from mixed models or GEE in finite samples.

Note that although GEEs in general use both horizontal and vertical information to estimate a treatment effect, the GEE we consider in Equation 3.4 with working independence and a categorical time trend uses almost exclusively vertical information [45]. Unlike methods specially derived to use vertical information, Equation 3.4 explicitly models a time trend and includes outcomes from the first and Jth time periods, which do not contribute any information about a vertical treatment effect. If the model were adjusted for other factors such as precision variables, the first and Jth periods could indirectly affect inference by contributing information about those other factors. To avoid confusion, we will continue to refer to GEE models as traditional methods that may utilize both vertical and horizontal information, but acknowledge that Equation 3.4 is a special case that we would expect to resemble vertical methods.

3.2.3 Existing vertical estimators

Non-parametric within-period

Thompson et al. [34] proposed the non-parametric within-period (NPWP) estimator, in which time-specific estimated treatment effects are averaged using inverse-variance weighting. First, for $j = 2, \ldots, J-1$ the vertical contrasts $\hat{\theta}_j$ are estimated by taking the average Y_{ij} on treatment and subtracting the average Y_{ij} on control. Note that the first time period is not used since all clusters are on control, and similarly the last time period is not used since all clusters have crossed over. Then the average treatment effect is $\hat{\theta} = \sum_{j=2}^{J-1} \frac{w_j}{\sum w_j} \hat{\theta}_j$, where the weights are based on the empirical variance of the clusters on control ($s^2_{j,0}$) and treatment ($s^2_{j,1}$), where $c_{j,0}$ and $c_{j,1}$ represent the number of clusters on control and treatment at time j, respectively:

$w_j = \hat{var}(\hat{\theta}_j)^{-1} = [(\frac{(c_{j,0}-1)s^2_{j,0}+(c_{j,1}-1)s^2_{j,1}}{c_{j,0}+c_{j,1}-2})(\frac{1}{c_{j,0}} + \frac{1}{c_{j,1}})]^{-1}$

Permutation tests are used to obtain exact p-values and/or confidence intervals. In this context, the 'average' treatment effect that is being estimated is a weighted average over time and over cluster-periods, so the estimand of interest is a treatment effect at the cluster level.

Non-parametric within-period with synthetic control

This method proposed by Kennedy-Shaffer, De Gruttola, and Lipsitch [35] builds on the NPWP estimator, but instead of using a simple average of the cluster-periods on control and treatment to obtain vertical contrasts, a synthetic control estimator is constructed for each treated cluster-period to form vertical contrasts.

The synthetic control Z_{ij} for a particular treated cluster-period in time j from cluster i is a weighted average of the untreated cluster-periods from time j, $Z_{ij} = \sum_{i':X_{i'j}=0} w_{i'j} Y_{i'j}$. The weights $w_{i'j}$ are chosen to minimize the mean squared difference of analogous contrasts in time periods in which cluster i was untreated, $MSPE_{ij} = \sum_{j':X_{ij'}=0} (Y_{ij'} - \sum_{i':X_{i'j}=0} w_{i'j} Y_{i'j'})^2$. Weights must also be nonnegative and sum to 1 for each Z_{ij}. After obtaining the appropriate

synthetic controls, for each treated cluster-period the estimate of the treatment effect is $\hat{\theta}_{ij} = Y_{ij} - Z_{ij}$. To estimate an overall treatment effect, we average the cluster-period estimators $\hat{\theta} = \frac{1}{\sum_{j=2}^{J-1}\sum_{i=1}^{MN} X_{ij}} \sum_{j=2}^{J-1}\sum_{i:X_{ij}=1} \hat{\theta}_{ij}$. Inference proceeds in the same way as the NPWP method; see Kennedy-Shaffer, De Gruttola, and Lipsitch [35] for details.

The synthetic control estimator is an unbiased estimate of the expected outcome among the control cluster-periods if the cluster-period-level outcomes have a symmetric distribution around some global mean. In this paper, we examine this method using equal weights across cluster-periods, although weighting can be used to increase efficiency; a within-period synthetic control estimator with weights chosen to decrease the variance of the treatment effect estimator performed better than the equally weighted version, but differences between the two were smaller than the differences between different classes of methods [35].

3.3 Methods: Vertical composite likelihood estimators

Here we present some simple CL models that focus on within-period information; see Varin Reid and Firth [37] for a more general introduction to CL methods.

Simple vertical CL

The first vertical composite likelihood estimator we propose uses vertical contrasts between treated and untreated clusters, similar to the NPWP method. We will call this the CLWP (composite likelihood estimation of within-period treatment effect) method. The primary motivation for using vertical contrasts - that is, $Y_{ij} - Y_{i'j}$ where $X_{ij} = 1, X_{i'j} = 0$ - is the removal of nuisance parameters. To formalize this, if the data-generating structure follows Equation 3.1 with a mean model of $\mu + \beta_j + X_{ij}\theta$ then for any pair of treated and untreated observations from the same time period, a sufficient statistic for the nuisance parameters μ and β_j is $Y_{ij} + Y_{i'j}$. If the joint distribution of $(Y_{ij}, Y_{i'j}|X_{ij} = 1, X_{i'j} = 0)$ is conditioned on the sufficient statistic, the portion of the likelihood involving the parameter of interest θ depends on the data only through $Y_{ij} - Y_{i'j}$. Thus, using vertical contrasts $Y_{ij} - Y_{i'j}$ to study θ is both intuitive and supported from a likelihood perspective of removing nuisance

parameters.

The composite likelihood corresponding to a within-period estimate is presented below (Equation 3.5). Note that each vertical contrast is represented by a separate component likelihood, and we have used a working independence structure between the contrasts. Because these units of observation can no longer be separated into discrete clusters, most traditional methods of analyzing clustered data (including mixed models and GEE) cannot be applied easily to these vertical contrasts; instead we use this CL model with working independence to account for the complicated correlation structure:

$$\mathscr{L}_{CLWP}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} f(\Phi; Y_{ij} - Y_{i'j}) \tag{3.5}$$

where $f(\Phi; Y_{ij} - Y_{i'j})$ is a Normal density with mean θ and some variance σ^2_{CLWP}, and Φ is the vector of all parameters in the model. Note that this mean model relies on the assumption that all clusters share a common fixed time trend (at least within periods that contain vertical contrasts, e.g. $j = 2, \ldots, J-1$) so that period effects cancel out.

If each component likelihood is correctly specified, the estimate $\hat{\theta}$ that maximizes the CL (the maximum composite likelihood estimate, or MCLE) is an unbiased estimate of the treatment effect [37]. Equation 3.5 assumes constant variance over all vertical contrasts; although this is consistent with random cluster intercepts (Equation 3.2) or a random treatment effect (Equation 3.3), it may be violated by other concerns such as random time effects with variance that grows or shrinks over time. Since this CL model uses aggregated cluster-period means, varying cluster-period sample sizes would also violate the constant variance assumption.

In this simple model (Equation 3.5), we give each likelihood component equal weight. If all cluster-periods have the same sample size, then equal weighting may be sensible. Note that the number of clusters on treatment and control in each time period was a concern in selecting efficient weights for the non-parametric methods, but this has effectively already been accounted for in the CL because time periods with more balance between treatment and control produce more likelihood components.

Inference on this CL model could be done via resampling methods such as jackknife or bootstrap [37]. In this paper, however, we focus on inference based on asymptotic properties, as they are much faster to compute and more analogous to traditional methods.

Estimation of a standard error for the model in Equation 3.5 is more complicated than for CLs commonly fit on cluster-randomized data [37], because each unit of observation is a difference between clusters. It is not possible to partition the set of all $Y_{ij} - Y_{i'j}$ into independent groups, which would typically be necessary to apply a GEE or mixed model. In some cases it is possible to apply standard GEE software to nonnested clustering, but in this case the number of clusters would be so large that this would likely be computationally unfeasible [46]. In some ways, this is similar to CL models used for analyzing spatial data, in the sense that observations are connected in a network and cannot be separated into truly independent groups. In the case of data without independent replicates, researchers often attempt to assemble pseudo-independent, possibly overlapping groups to estimate standard errors; see e.g. the window subsampling procedure coined by Heagerty and Lele [47]. However, the setting of vertical contrasts is unique because the data generating process that we used to create these contrasts involves plausibly independent clusters. We propose one estimator of the asymptotic variance below, but further research is needed to explore other estimators and compare their performance.

Under some regularity conditions, MCLEs are asymptotically Normal with a closed-form variance [37]. Instead of the usual Fisher information, CLs use the Godambe information matrix, G. In a traditional cluster-randomized CL model, if clusters are independent and the component likelihoods are correctly specified, $\sqrt{MN}(\hat{\Phi}-\Phi) \xrightarrow{d} N(0, G^{-1}(\Phi))$. The Godambe information matrix is estimated using a sandwich form $G(\Phi) = H(\Phi)J^{-1}(\Phi)H(\Phi)$; see Varin Reid and Firth [37] for more information about the sensitivity matrix $H(\Phi)$ and variability matrix $J(\Phi)$. $G^{-1}(\Phi)$ is a robust variance estimate, in the sense that the sandwich form accounts for our belief that the covariance matrix induced by the model is misspecified [37]. We use these traditional CL methods as a framework, but acknowledge that traditional proofs of consistency and asymptotic Normality (e.g. [48]) depend on independent replicates

of the observations, which is violated in the setting we consider here. Although the vertical contrasts have a complicated correlation structure without clear independent groups, we have constructed these standard errors in an attempt to leverage the underlying independent clusters from the SWT, and demonstrate the performance of this estimator using simulations.

Let $\ell_{CLWP}(\Phi; y)$ denote the log composite likelihood for this model, and let the composite score function with respect to $\Phi = (\theta, \sigma^2_{CLWP})$ be $u_{CLWP}(\Phi; y) = \nabla_\Phi \ell_{CLWP}(\Phi; y)$. To estimate $H(\Phi)$, we use:

$\hat{H}(\Phi) = -\frac{1}{MN} \sum_{j=1}^{J} \sum_{i,i':X_{ij}=1,X_{i'j}=0} \nabla_\Phi u_{CLWP}(\hat{\Phi}; Y_{ij} - Y_{i'j})$

To estimate $J(\Phi)$, we allow each contrast $Y_{ij} - Y_{i'j}$ to 'belong' to both the treated cluster i and untreated cluster i'. In this context, let d_k denote the set of all vertical differences $Y_{ij} - Y_{i'j}$ for $j = 1, \ldots, J$ that involve a specific cluster k, so either $i = k$ or $i' = k$ (so each $Y_{ij} - Y_{i'j}$ is counted twice)

$\hat{J}(\Phi) = \frac{1}{MN} \sum_{i=1}^{MN} u_{CLWP}(\hat{\Phi}; d_i) u_{CLWP}(\hat{\Phi}; d_i)^T$

By continuing to group units of observation by the underlying independent clusters, we hope to capture the asymptotic mechanism of interest, that is increasing the total number of clusters MN. However, with some basic assumptions about the data-generating mechanism it is possible to quantify the network of correlations between each vertical contrast, based on whether they share any cluster-period means or have any clusters in common. It may be possible to use this information to create a better estimate of the standard error, if it can be used to divide the contrasts into pseudo-independent groups. Note also that although the calculation of the point estimate does not require independent clusters, we did assume independence between clusters to calculate the variance. Adaptations for CL models without strictly independent replicates may be applicable [47], but in this paper we restrict our attention to this simple sandwich estimator and study its performance under violations of the independent clusters assumption.

Note that this asymptotic inference requires a sufficiently large number of clusters. Understanding these limits is especially important for application to SWTs, since trials often have few clusters.

Vertical CL adjusted for baseline

In this model, we take advantage of the parametric portion of the model by adjusting the mean model for the baseline outcomes. Recall that for a classic design, the CLWP does not use any data from the baseline period, since no clusters have crossed over by this time. We call this method the CLWPA, since it is the CLWP with an adjustment.

$$\mathscr{L}_{CLWPA}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} f(\Phi; Y_{ij} - Y_{i'j} | Y_{i1} - Y_{i'1}) \tag{3.6}$$

where $f(\Phi; Y_{ij} - Y_{i'j} | Y_{i1} - Y_{i'1})$ is a Normal density with mean $\theta + \gamma(Y_{i1} - Y_{i'1})$ and some variance σ^2_{CLWPA}. The motivation for adjusting for the baseline difference is efficiency [49]. In SWTs, many possible adjusted models might be considered because vertical contrasts $Y_{ij} - Y_{i'j}$ have been observed in the previous $j-1$ time points. We focus on this particular adjustment because it is available for all classic SWTs, and it is a simple model because every vertical contrast $Y_{ij} - Y_{i'j}$ has $X_{i1} = X_{i'1} = 0$. Intuitive alternatives such as adjusting for the observations in the previous time period must account for the fact that some vertical contrasts $Y_{ij-1} - Y_{i'j-1}$ have $X_{i1} = X_{i'1} = 0$, but others have $X_{ij} = 1, X_{i'j} = 0$. It is easy to adjust Equation 3.6 for other variables when necessary (e.g. stratifying factors) or desirable (e.g. precision variables). This is an advantage of the parametric mean model in the CL, as it is easier to identify and follow established best practices for cluster randomized trials compared to the fully nonparametric methods such as NPWP.

Another way of thinking about this model is to think about looking at the difference in residuals after factoring out the influence of the baseline. That is, in terms of the mean model and conditional on the baseline observations Y_{i1} and $Y_{i'1}$, $E[Y_{ij} - Y_{i'j}] = E[\theta + \gamma(Y_{i1} - Y_{i'1})]$ is equivalent to $E[(E[Y_{ij}] - \gamma Y_{i1}) - (E[Y_{i'j}] - \gamma Y_{i'1})] = \theta$. Other characteristics can be factored out of the residual as well by adjusting for their difference. For example, suppose we have a precision variable P_i which is a characteristic of each cluster i that is either fixed or measured at baseline, which does not have any interaction with the treatment. We would add this to Equation 3.6 by adjusting for $P_i - P_{i'}$.

We note that the data-generating model implied by adjusting for baseline outcomes is not always very plausible. For example, it implies that correlation between baseline outcomes and outcomes at time j is constant for $j = 2, \ldots, J$; this is true for an exchangeable correlation structure of cluster-periods over time, but not e.g. an autoregressive correlation structure which might be more realistic. This model would also be violated if cluster-period sample sizes differed between clusters or over time. However, estimates of the treatment effect are generally consistent even when the model assumptions about adjustments are not correct [50, 51], and in some cases even the model-based variance estimate (i.e. the standard errors were computed assuming all model assumptions hold, not using robust methods) is valid [52]. Based on this previous research, we hypothesize that adjusting for baseline outcomes should improve efficiency (assuming use of robust SE) without biasing the treatment effect estimate, and we explore the magnitudes of these improvements in simulations. We would expect the baseline outcomes to be a better precision variable when the fitted model is correct (i.e. exchangeable cluster-period correlation structure), and when baseline observations are strongly correlated with later outcomes (e.g. large variance of random cluster intercepts).

3.4 Simulations

3.4.1 Simulation settings

We simulate data from a classic SWT design to compare the behavior of the methods listed above. Our simulation settings are based on data made available by Johns Hopkins on COVID-19 trends [53]. Although this data does not come from a SWT, the setting is broadly similar to one in which a SWT might be conducted. See appendix for details on this data and how settings were derived. In this hypothetical SWT, each cluster is a county, each time period is one month long, and the outcome of interest is log-transformed average daily number of confirmed COVID-19 cases.

We examine two designs with the same total number of clusters: three sequences with 22 clusters per sequence, and 11 sequences with six clusters per sequence. Note that these

two designs also have a different length (four vs. 12 time periods) and a different number of recruited individuals (three times more in the 11-sequence design). Although the data-generating settings are analogous for the two trial designs, the implications may be different. For example, when cluster-periods have an autoregressive correlation structure, the average correlation is larger in the 3-sequence design since cluster-periods can be no more than three time points apart by design. To understand the limitations of these CL models with a small number of clusters, we also perform a set of simulations using these settings but varying the number of clusters per sequence.

We consider four data-generating settings:

A. Data is generated with a common time trend and exchangeable correlation structure (Equation 3.2), with $\sigma = 0.43$ and $\tau = 1.27$.

B. Data is generated as in scenario A, but with an AR(1) correlation structure within clusters so that for two observations t time points apart, $corr(Y_{ij}, Y_{ij+t}|u_i) = 0.31^t$.

C. Data is generated as in scenario A, but with a random treatment effect (Equation 3.3) with $\eta = 0.90$.

D. Data is generated as in scenario A, but half the clusters are randomly selected to follow a different time trend.

See appendix for details, and Figure 3.1 for an example of simulated data from each of these scenarios. For each scenario, we examine bias of the treatment estimate, coverage, and power for each of the methods. For methods that use permutation for inference, we use 500 randomly sampled permutations of the sequences. Each setting is simulated 1,000 times, so if the true Type I error rate of a method is 5%, there is a 95% chance that the empirical Type I error rate will be between 3.65% and 6.35%. Code for the NPWP and NPWP with synthetic control was based on code made available by Kennedy-Shaffer, De Gruttola, and Lipsitch [35]. The methods we will compare are:

- MM: Mixed model with random cluster intercepts (Equation 3.2)

- MMT: Mixed model with random cluster intercepts and random treatment (Equation

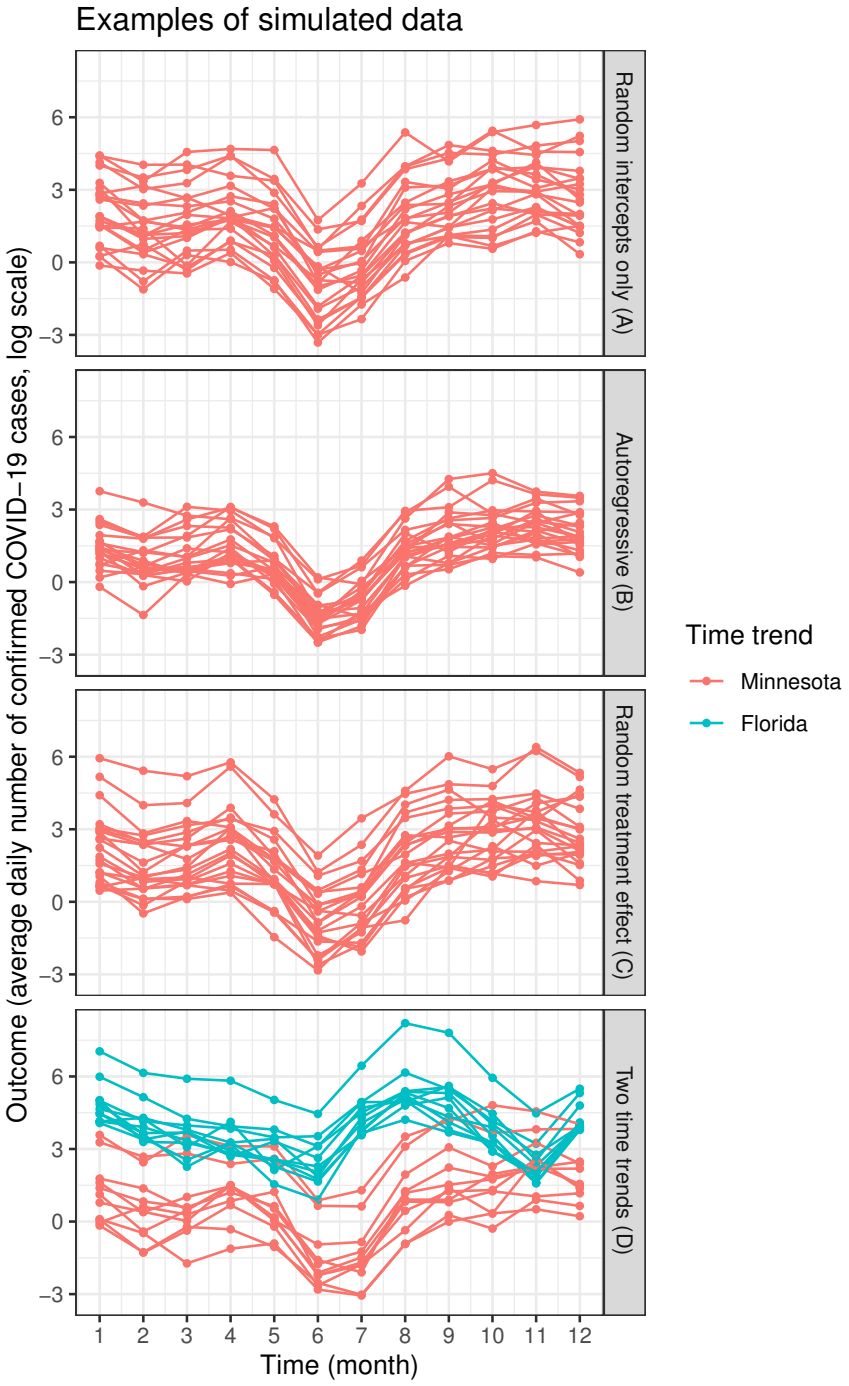

Figure 3.1: Examples of simulated data under the four scenarios with no treatment effect ($\theta = 0$), with 11 sequences and two clusters per sequence. Each line represents the outcome for a particular cluster over time.

3.3)

- GEE: GEE with working independence (Equation 3.4)
- NPWP
- SC: NPWP with synthetic control
- CLWP: Vertical CL (Equation 3.5)
- CLWPA: Vertical CL adjusted for baseline (Equation 3.6)

Throughout, we do not apply any small-sample corrections, continuing to use the asymptotic-based inference for mixed models, GEE, and CL, and permutation-based inference for NPWP. Due to large computational demands of the synthetic control method, we only fit this method for designs with a small number of time periods.

3.4.2 Simulation results

For the 3-sequence SWT under scenarios A, B, C, and D, all methods appear unbiased (Figure 3.2). Broadly, the point estimates from GEE, NPWP, and CLWP had similar empirical variability, which was substantially larger than the variability of the MM and MMT, although the magnitude of the difference depended on the scenario. The variability of the treatment estimate from CLWPA was mostly intermediate compared to other methods, but in the scenario with two time trends (D) was very similar to the mixed models (MM, MMT). The SC method was less variable than the NPWP, but slightly more variable than the CLWPA. Relative behavior of the point estimates was similar for the 11-sequence design (Figure 3.3). Behavior of point estimates under an alternative hypothesis was the same, as expected with a linear link, and not presented here.

For the 3-sequence SWT, most methods had close to a 5% Type I error rate, with the exception of the simulations with a random treatment effect (Figure 3.4). Of the methods

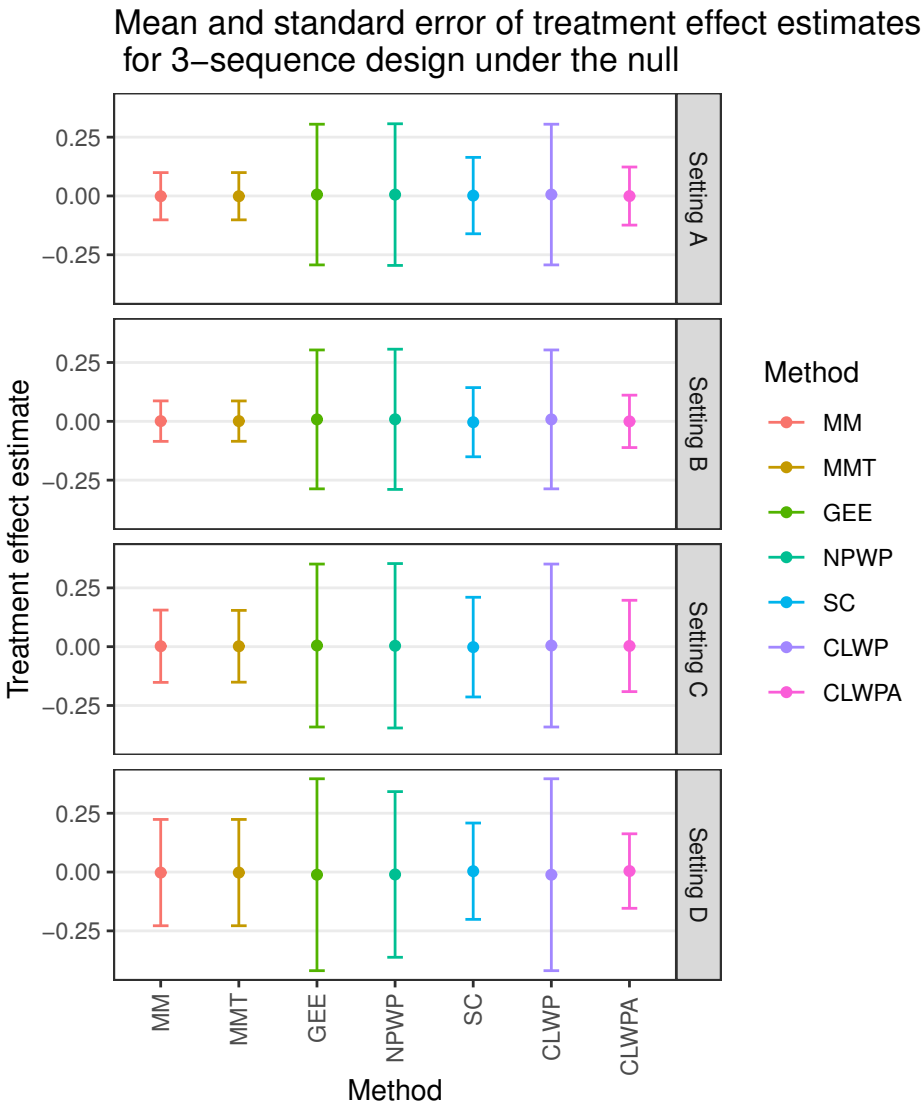

Figure 3.2: Point estimates of the treatment effect from a SWT with three sequences and 22 clusters in each sequence. Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), NPWP with synthetic control (SC), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA).

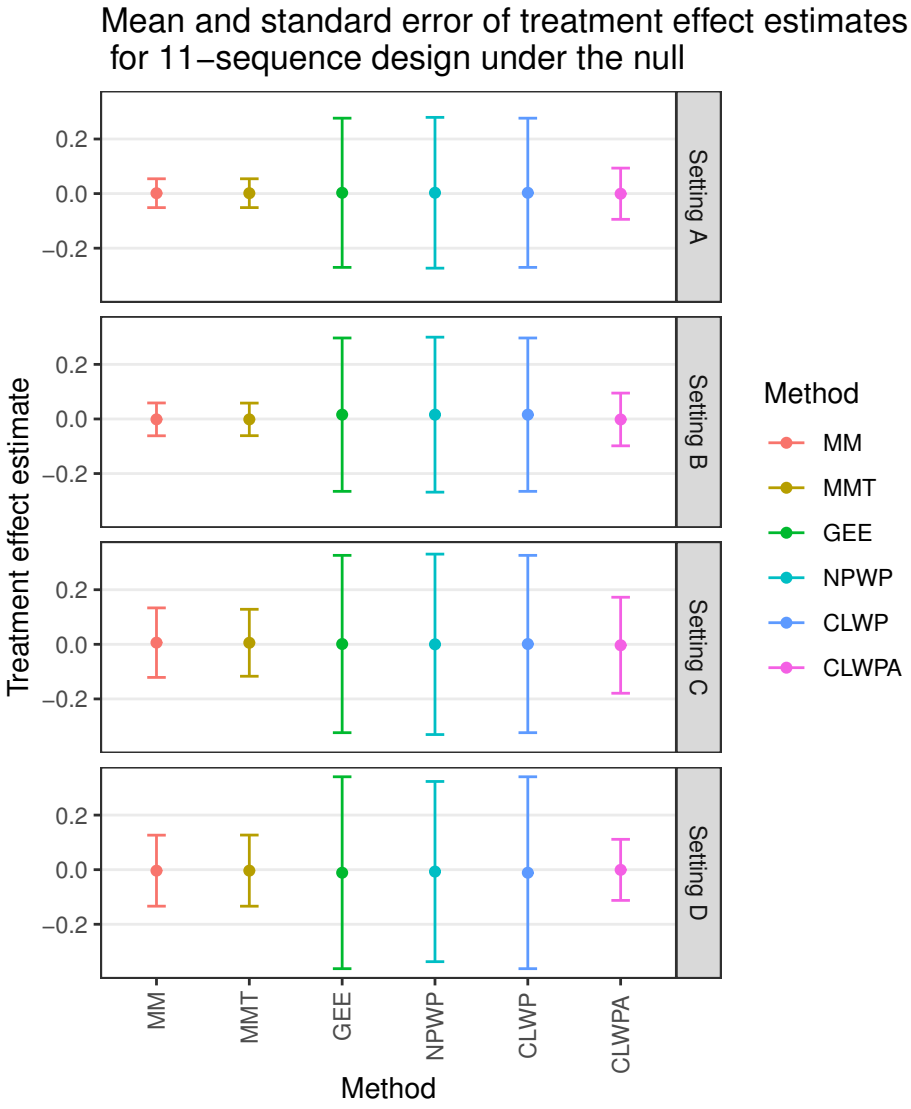

Figure 3.3: Point estimates of the treatment effect from a SWT with 11 sequences and six clusters in each sequence. Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA).

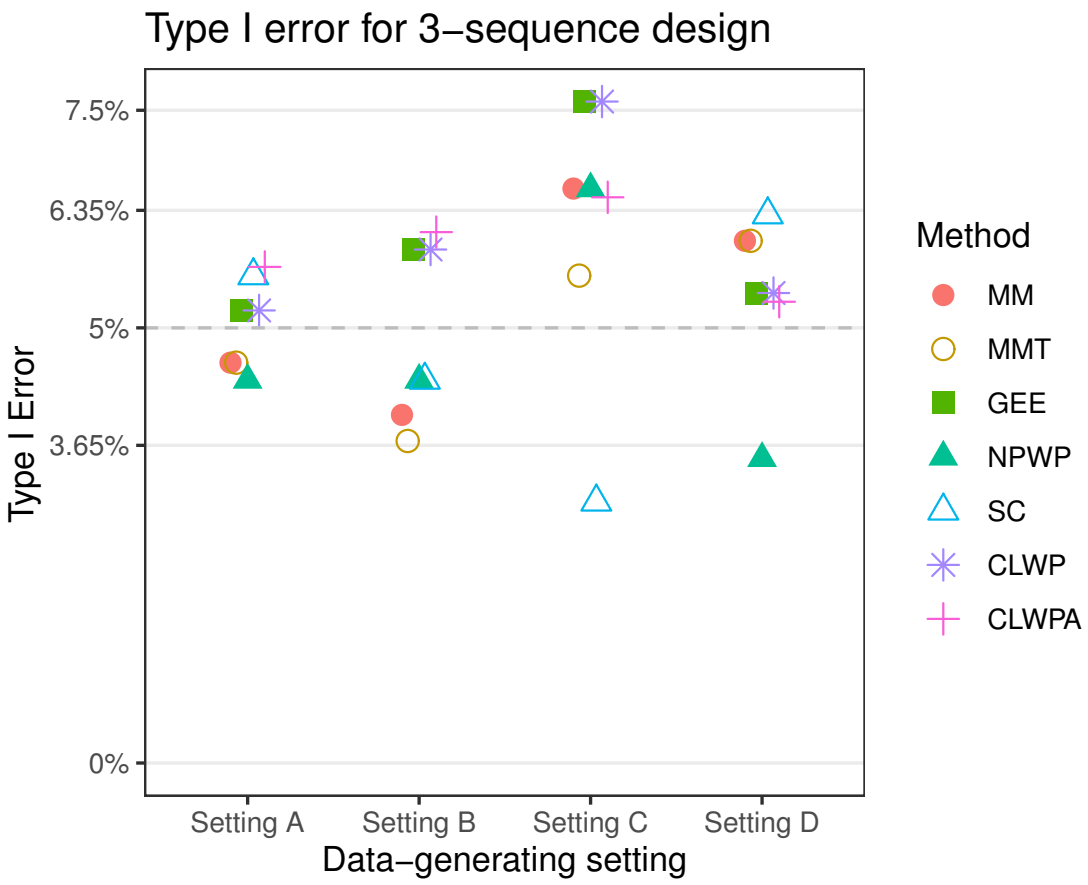

Figure 3.4: Type I error from a SWT with three sequences and 22 clusters in each sequence, with no treatment effect ($\theta = 0$). Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), NPWP with synthetic control (SC), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA).

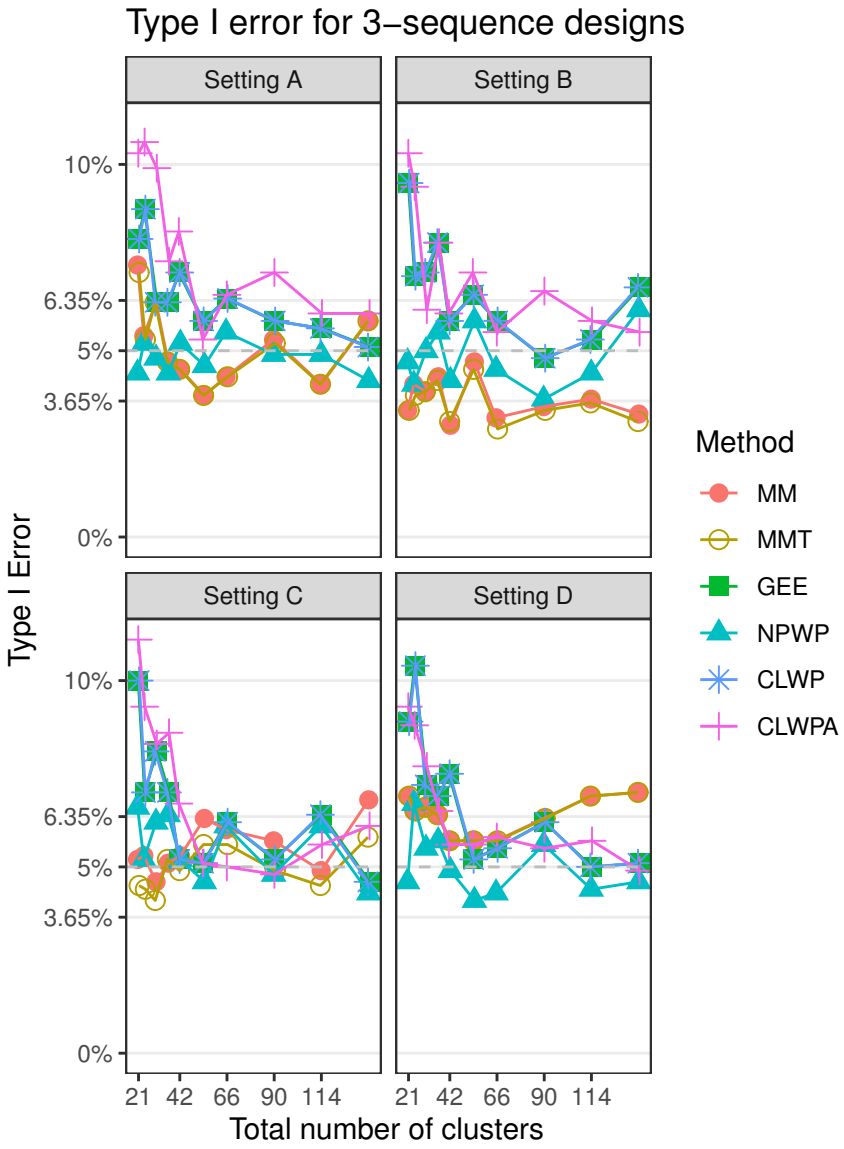

Figure 3.5: Type I error from a SWT with 3 sequences and a varying number of clusters per sequence. The simulated data had no treatment effect, and either random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D). Data was analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA).

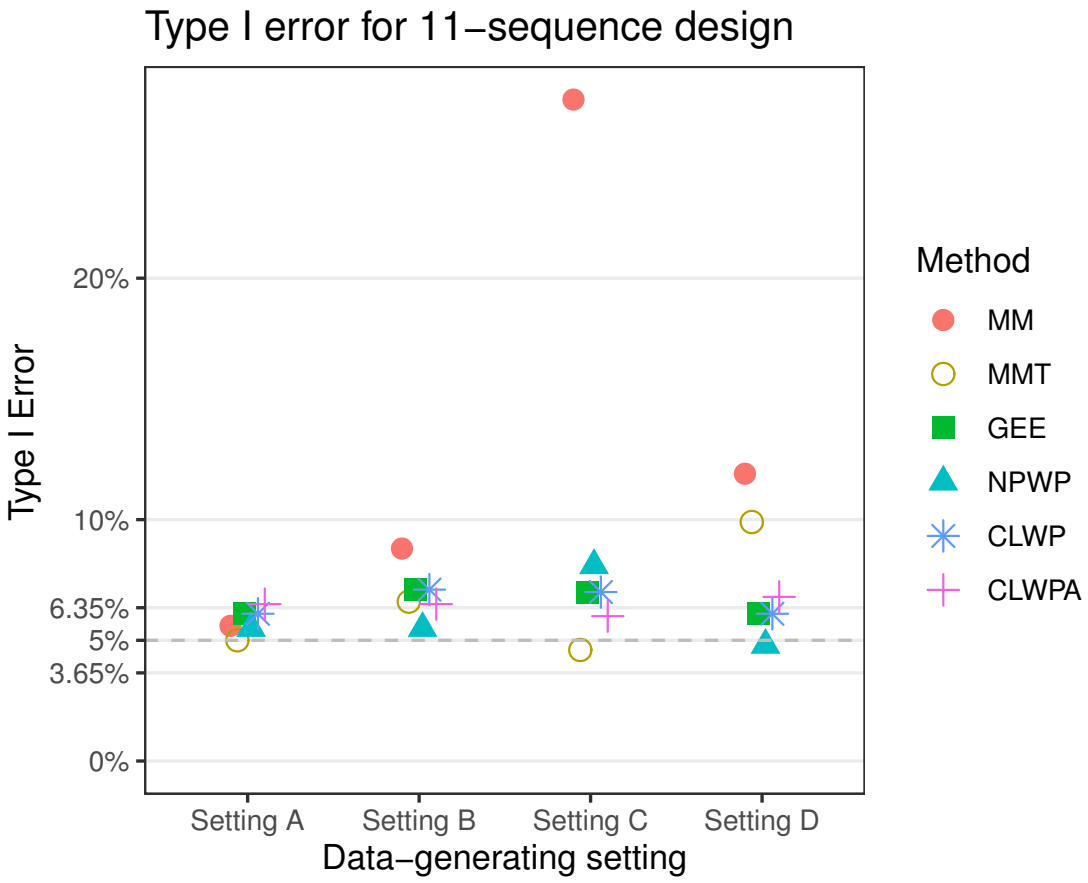

Figure 3.6: Type I error from a SWT with 11 sequences and six clusters in each sequence, with no treatment effect ($\theta = 0$). Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA).

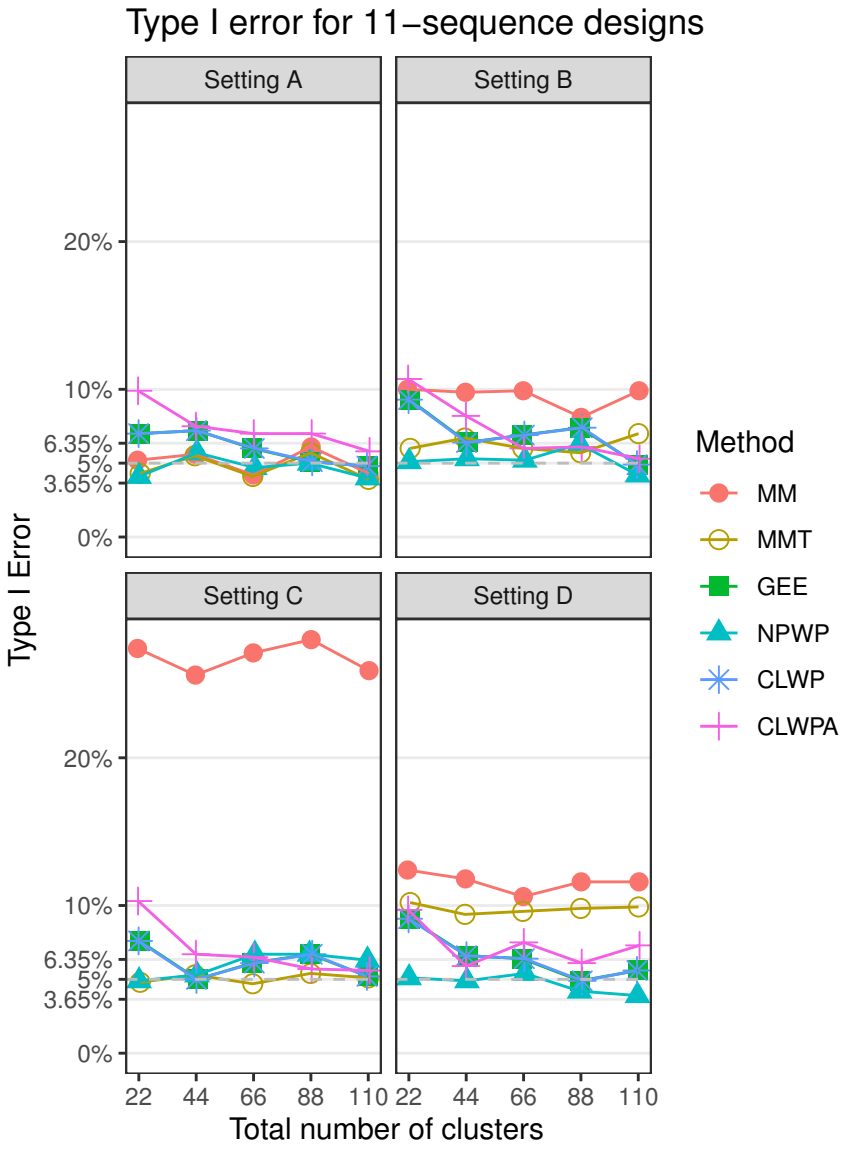

Figure 3.7: Type I error from a SWT with 11 sequences and a varying number of clusters per sequence. The simulated data had no treatment effect, and either random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D). Data was analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA).

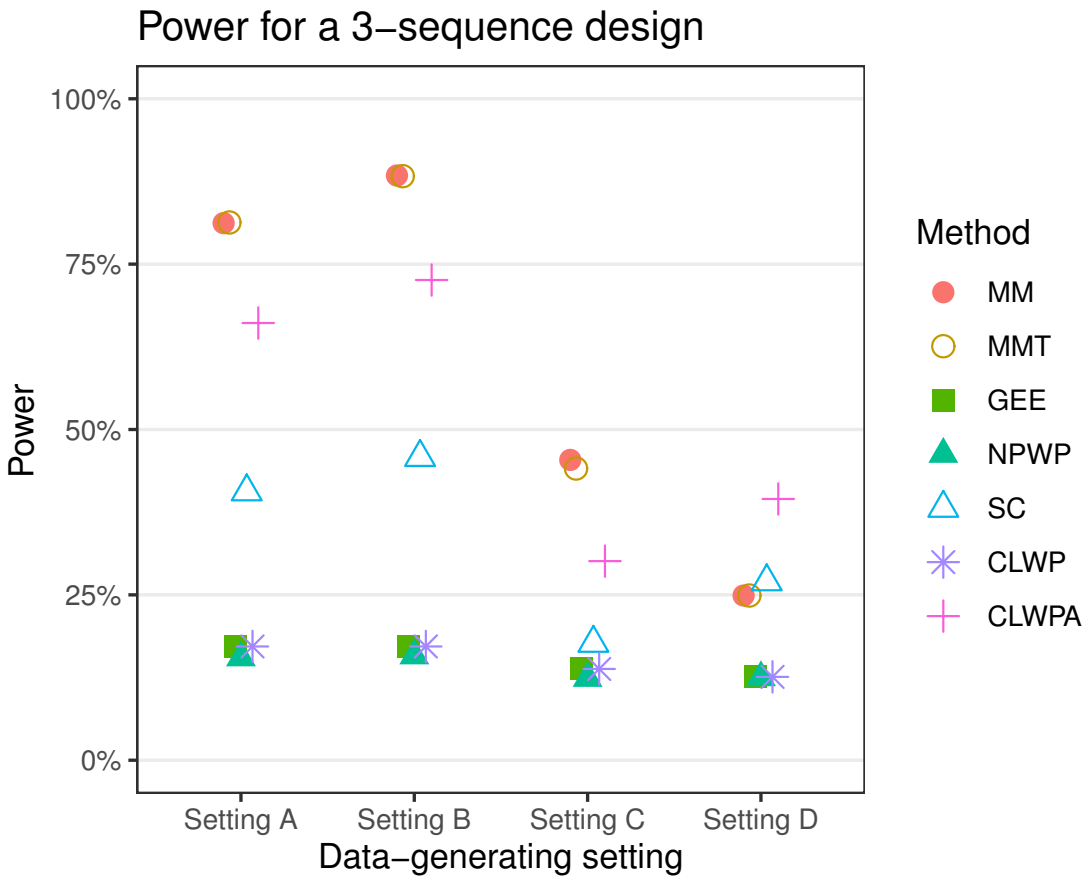

Figure 3.8: Power from a SWT with three sequences and 22 clusters in each sequence, with a substantial treatment effect ($\theta = -0.28$). Note that some of these are not accurate representations of power, since the corresponding Type I error was elevated. Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), NPWP with synthetic control (SC), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA).

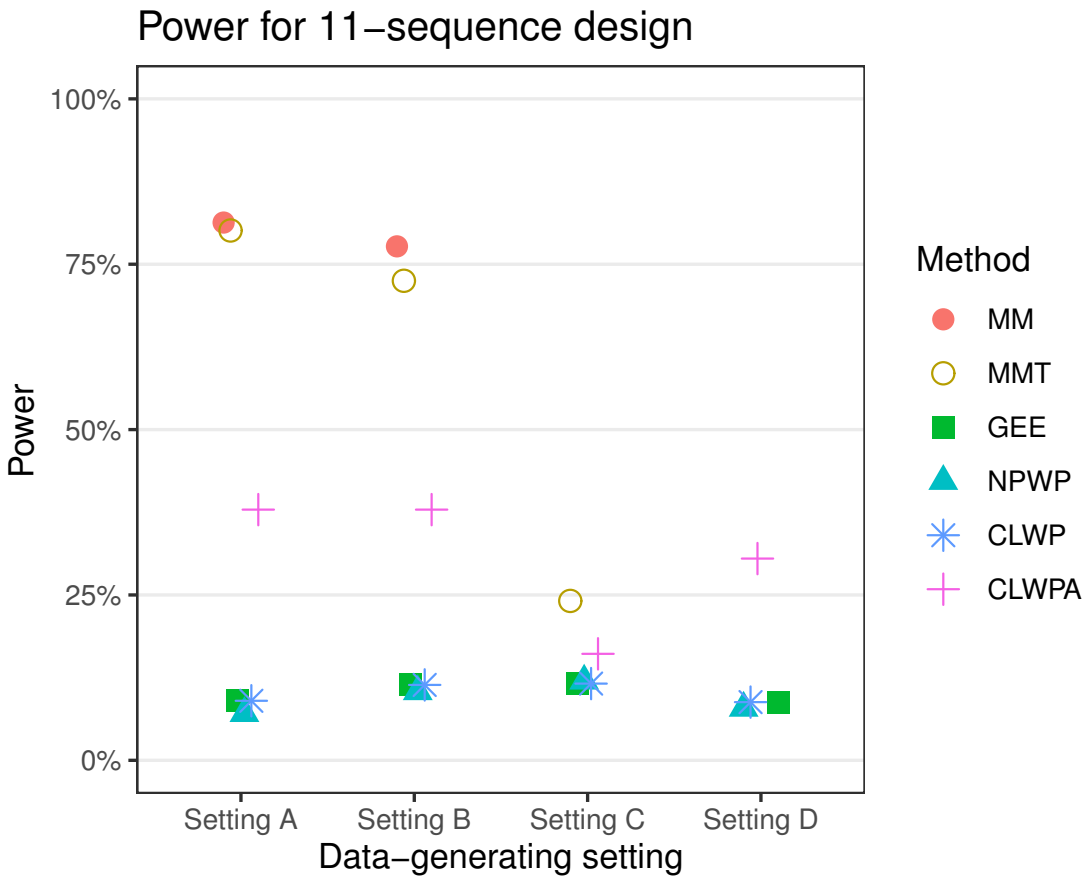

Figure 3.9: Power from a SWT with 11 sequences and six clusters in each sequence, with a substantial treatment effect ($\theta = -0.15$). Note that some of these are not accurate representations of power, since the corresponding Type I error was elevated. Data was generated with random cluster intercepts (A), AR(1) correlation (B), random treatment effects (C), and two cluster-specific time trends (D), and analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA). Due to their very poor Type I error, MM in scenarios C and D and MMT in scenario D are not shown here.

with an in appropriate Type I error rate in these simulations, the MM, NPWP, and SC methods are not compatible with a random treatment effect, while GEE, CLWP, and CLWPA are (but the presence of a large random treatment effect impacts the efficiency of a working independence covariance matrix). For the 3-sequence design, it appears that the GEE, CLWP, and CLWPA converge towards an appropriate Type I error rate as the number of clusters increases, while the MM and NPWP do not (Figure 3.5). In fact, we would expect the Type I error rate of the NPWP method to increase with the total number of clusters, since it has greater power to correctly reject the strong null hypothesis. The issues with Type I error rates were more dramatic in the 11-sequence SWT (Figure 3.6). The mixed models in particular had grossly inflated Type I error, above 10% for the simulations with two time trends and above 20% for the simple mixed model under a random treatment effect. Unlike the 3-sequence SWT, even with a very large number of clusters these Type I error rates were far from 5% (Figure 3.7).

Simulations with a meaningful treatment effect (Figure 3.8, Figure 3.9) demonstrate that when MM or MMT are correctly specified, they are the most powerful methods. Note that power has not been adjusted for Type I error rate. GEE, NPWP, and CLWP had relatively consistent but low power. Throughout, CLWPA is more powerful than GEE, NPWP, or CLWP. In the scenario with two time trends (D), CLWPA is the most powerful method that has close to reasonable Type I error control (although the power is still very low). SC is more powerful than NPWP, but less powerful than CLWPA.

In these simulations, the MMT method sometimes indicated potential issues with model fit. In scenario A with 11 sequences, the MMT had a message about model fit (e.g. fitting the random treatment variance as exactly zero) 51%-52% of the time. In other scenarios for the 11-sequence design, there were indications of issues with MMT fit no more than 1% of the time. For the 3-sequence design, scenario A was similar (49%-52%), but there were also messages 13%-14% of the time for scenario B and 100% of the time for scenario D.

Our simulations examining these methods in scenarios with a small number of clusters showed that the CL methods generally followed the same patterns as the GEE in Equation

3.4 (Figure 3.5, Figure 3.7). As expected, the NPWP and mixed model methods had better Type I error control than GEE and the CL methods when the number of clusters was low, when correctly specified.

3.5 Application to LIRE trial

To demonstrate the application of these methods to data from a SWT, we analyzed the Lumbar Imaging with Reporting of Epidemiology (LIRE) data [54]. The LIRE trial was a SWT designed to see whether including extra information in imaging reports (e.g. information about the prevalence of findings in routine imaging for asymptomatic individuals) might reduce unnecessary interventions. 98 primary care clinics from four large health systems were randomized to five sequences and observed over six time periods. Randomization was stratified on clinic size and health system. The primary analysis found that the intervention did not have a significant impact on health care utilization, but found a small but statistically significant decrease in the odds of a participant being prescribed opioids by a study clinician within 1 year of imaging (odds ratio 0.95, 95% CI 0.91-1.00). The analyses were done using mixed models with robust standard errors, adjusted for time (linear), participant age and imaging type, clinic size (categorical), and health care system. Given that this trial was conducted between October 2013 and September 2016 in clinics from different health systems across the US, it is plausible that researchers might be concerned about strong time trends. For the opioid outcome, if health systems implemented different policies about prescribing opioids, it may be possible that time trends differ between health systems.

To analyze this data using the methods described above, we aggregated outcomes to the cluster-period level. Two primary care clinics did not observe any participants during the baseline time period, and were excluded from these analyses. One important difference from the simulations above is that the number of participants per cluster-period is not constant in the LIRE trial. A total of 250,401 participants were included in the trial. The number of participants per primary care clinic in each time period ranges from one to 4,019, with a median of 102.5 participants per cluster-period. Although not exhaustive, cursory

simulations did not reveal major issues with applying these methods to varying cluster-period sizes (see appendix). For the purposes of comparing these methods, we will adjust models for only a categorical time model, and not any additional covariates used in the original analysis of the trial.

Because it was statistically significant in the original study, we focus on the secondary opioid prescription outcome. At the cluster-period level, this outcome is the percent of patients at each clinic who were prescribed opioids by study clinicians within 1 year of imaging, and appears roughly Normal. The treatment effect is a difference in percentages. Table 3.1 shows treatment effect estimates from each method, 95% confidence intervals, and the time required to execute each method. The estimated treatment effect is relatively similar across methods, ranging from 0.29% for the synthetic control method to -1.1% for the NPWP method. Although all p-values are greater than 0.05 in this example, it is clear that choice of method could be important. All the methods examined here were quick to compute, with the exception of the synthetic control method. Table 3.1 also shows 95% confidence intervals for methods based on model-based or sandwich standard errors. It is also possible to calculate confidence intervals for the non-parametric methods by inverting the permutation test [35], but this would be very computationally intensive for the synthetic control method in this example.

3.6 Discussion

3.6.1 Summary

In this paper, we have proposed a novel semi-parametric method for obtaining estimates of a treatment effect in SWTs using vertical information. In our simulations, we have demonstrated that, like other vertical methods, vertical CL estimators are more robust than traditional methods, but less efficient. For example, with the 11-sequence design the mixed model with random cluster intercepts had inflated Type I error under settings B, C, and D, while composite likelihood methods had Type I error around 5% for all simulations with

Method	Estimated treatment effect	95% CI	p-value	Run time
Mixed model (MM)	-.93%	-3.2%,1.3%	0.42	<1 second
Mixed model with random treatment (MMT)	-.93%	-3.2%,1.3%	0.42	<1 second
GEE with working independence (GEE)	-.92%	-4.0%,2.2%	0.56	<1 second
Non-parametric within-period (NPWP)	-1.1%		0.48	3 seconds
NPWP with synthetic controls (SC)	0.29%		0.89	12 hours
Vertical CL (CLWP)	-.92%	-4.0%,2.2%	0.56	1 second
Vertical CL adjusted for baseline (CLWPA)	-.98%	-3.4%,1.4%	0.42	6 seconds

Table 3.1: Results from the LIRE trial, estimating the difference the percent of patients at a clinic who were prescribed opioids. Note: the NPWP confidence interval and run time are for 500 permutations; for SC due to computational burden we used only 100 permutations. The MMT model fit the random treatment effect as exactly zero, reducing it to a mixed model with random cluster intercepts only.

a large number of clusters. The mixed model with a random treatment effect fared better than the simpler mixed model, but still had very high Type I error when there were two time trend strata. Although vertical methods are less efficient than correctly-specified mixed models, adjusting for the baseline outcome substantially improves the efficiency of composite likelihoods in most examined scenarios. Improvements in power from adjusting for baseline were especially dramatic in the 3-sequence design, with the CLWPA method having nearly three times the power of the pure vertical methods in settings A and B. Simulations suggest that sample size requirements for valid inference with CL methods in the SWT setting are similar to those for GEE's. Table 3.2 gives a broad summary of how the characteristics of vertical CL models compare to mixed models, GEE, and NPWP estimators. In short, CL models are similar to GEE models, but there is some additional flexibility in CL models that may be useful for SWTs which we discuss below.

Similar to previous work, we conclude that traditional methods which take advantage of both horizontal and vertical information are excellent efficient methods when model assumptions hold. However, if researchers are concerned about violations then vertical methods offer a more robust alternative. Although vertical methods are generally less efficient than traditional methods, statistical methods such as synthetic controls and adjusting for precision variables offer options for improving their efficiency.

In most of the simulations we considered here, the GEE model had similar Type I error control and power as the NPWP estimator. In SWTS with few clusters, we would expect NPWP to perform better, and in settings with a random treatment effect we would prefer the inference from a GEE. Although the GEE method departs from the typical two-step setup that characterizes many other vertical estimators [55], because it uses almost exclusively vertical information [45] and has comparable performance to the NPWP when there are many clusters and no random treatment effect, we would argue that the GEE with working independence and a categorical time trend is itself a valid vertical effect estimator. This may be a departure from previous literature. For example, when Thompson et al. [34] attempted to use GEE to obtain a vertical effect, they fit a model very similar to Equation

Issue/assumption	**Mixed models**	**GEE models**	**NPWP**	**CLWP models**
Inference method presented here	Asymptotic	Asymptotic, sandwich SE	Permutation	Asymptotic, sandwich SE
Uses horizontal information	Yes	Optional	No	Optional
Type of null hypothesis	Weak	Weak	Strong	Weak
Need to model time	Yes	Yes	No	No
Specify distribution of outcomes	Yes	Yes	No	Yes
Correlation structure	Defined by model	Estimated from data	No random treatment effect	Estimated from data
Simple to account for random treatment effects	Yes	Yes	No	Yes
Easy to adjust for covariates or time-varying treatment	Yes	Yes	No	Yes
Required number of clusters	Moderate	Large	Small	Large
Efficiency when correctly specified	High	Low to moderate	Low	Low to moderate

Table 3.2: Table summarizing properties of different method families. Many variations and extensions exist that we do not address here; this table is not exhaustive, but a high-level summary of strengths and weaknesses.

3.4, but instead of a single treatment effect θ they fit one treatment effect for each time period, then averaged the period-specific effects using inverse variance weighting. Even in their simulations with many clusters, this estimator had low confidence interval coverage and was not recommended for use in any scenario. The authors noted that fitting the GEE model with categorical time and time-by-treatment required many parameters relative to the number of clusters, whereas the GEE in Equation 3.4 combines the time-specific estimates as part of the GEE fitting process. GEE models in general are already a staple of robust SWT analyses, but special cases like Equation 3.4 have not had as much attention in the literature surrounding vertical estimators.

The vertical methods examined here are all valid ways of analyzing a SWT, but each method has its own strengths and weaknesses. Computationally, closed-form asymptotic standard error estimates for a CL model are preferable to permutation-based inference. Permutation-based inference has the additional disadvantages of testing a strong null, and relying on the intervention not affecting the variance of the outcome. The vertical CL models, on the other hand, are congruous with a random treatment effect and easily adapted to any other design element that might cause clusters to differ (e.g. SWTs with offset start times). The parametric mean model in a CL model also makes it very easy to adjust for stratifying variables and account for more complicated parametric forms of the treatment effect (e.g. time-varying). However, the CL model does make assumptions about the underlying distribution of the outcome, while the NPWP does not. A major weakness of the CL model is that like GEE, it requires a substantial number of clusters in order to rely on closed-form standard errors. Future work may mitigate this via small-sample corrections or exploring other methods of estimating standard errors. Researchers concerned about the performance of sandwich standard errors might prefer to proceed using permutation-based or resampling methods.

The estimates produced by the CLWP method were essentially identical those from the GEE in Equation 3.4. There are subtle differences between the models; for example, the working independence structure in the CLWP is consistent with a residual variance that

depends on treatment status, while the GEE is not. This is because for the models examined in this paper, GEE uses cluster-period means as the units of observation, while CLWP uses vertical contrasts; vertical contrasts still have constant variance if the residual variance depends on treatment status, while the cluster-period means do not. But because these differences are immaterial in realistic simulations, the simple CLWP is not an interesting novel estimator. However, there are some interesting extensions of the CL model (e.g. CLWPA), some of which would be difficult or impossible to replicate using a GEE model; see below.

3.6.2 Extensions

Although we presented some simple options here, CL models are very flexible and can be modified to reflect different goals and assumptions. Many variations on the mean model are possible. If there is a time-varying treatment effect, θ can be replaced with a more appropriate time-dependent model (see Appendix). For example, researchers can estimate the treatment effect specific to each time after treatment (e.g. a classic 2-sequence SWT would have a θ_1 and a θ_2 for the effect of treatment one and two time periods after cross-over, respectively). In that case, if there is no random treatment effect, contrasts could be made between two treated clusters as well, where the mean would be the difference of two treatment effects, and include any time periods after all clusters have crossed over. Or if there is a random treatment effect, the variance assumptions could be relaxed by fitting different variance terms depending on how many clusters in each vertical contrast had already crossed over. The mean model can also be easily modified to include other parameters, such as precision variables, any factors that were incorporated in randomization, and even interactions between treatment and cluster characteristics.

In these CL models, component likelihoods were equally weighted, so time periods with more vertical contrasts may contribute more information about the treatment effect. Researchers might use weights on the likelihood components to target different estimands. For example, a researcher could give equal weight to each time period. Weights could also be used to reflect beliefs about similarity between clusters. For example, if researchers have

reason to believe that some clusters followed different time trends (e.g. time trend differed by region), certain contrasts could be down-weighted (e.g. by distance between cluster locations) or removed (e.g. if regions were discrete and identifiable, include only contrasts within the same region).

Although the CL models with working independence performed well, it may be possible to construct CL models that are more efficient by reflecting the correlation between vertical contrasts in the likelihood components (see Appendix). Although in GEE's it may be common to choose between working independence and working exchangeable, in CL models it is easy to construct a working structure that reflects only the strongest beliefs with minimal assumptions, e.g. a working correlation which is exchangeable between adjacent time points and independent between all others by using joint likelihoods for observations that are close together in time [56]. At the cluster-period level, a model like this might be very sensible in that it could model correlation between adjacent time points for efficiency, with relatively mild assumptions (e.g. constant correlation between adjacent time periods across clusters and time). However, modeling correlation between vertical contrasts creates a much more complicated model because each unit of observation involves two clusters. Assembling pairwise component likelihoods that involve vertical contrasts from adjacent time periods creates a network of modeled correlations, which takes careful thought to simplify or model, and may be difficult to understand and fit (see Appendix). However, with further research this could be one mechanism for adding horizontal information in a more thoughtful way.

It might also be possible to leverage more of the pre-crossover data; our CLWPA model adjusted only for control observations from the first time period, but for some vertical contrasts there might be many prior control periods with additional information (see Appendix). In particular, if researchers suspect cluster-periods are not exchangeable, adjusting for more recent control periods may be more informative than adjusting for baseline (see Appendix). A feature of the SWT design in this context is that for later sequences, there may be many observed time periods before crossover which can be leveraged for efficiency. In a CL, it is possible to adjust for a variable number of prior control periods, reflecting the SWT design

and researchers' beliefs about what information might be relevant (see Appendix). This is one potential advantage of the CL over a GEE.

Although our limited simulations did not indicate any problems with applying CL methods to SWTs with varying cluster-period sample sizes, it is possible to adapt the model to explicitly allow for this (see Appendix).

These CL methods also extend naturally to individual-level outcomes instead of aggregated cluster-period level outcomes (see Appendix), at the expense of computing time. This may be especially attractive if there are important individual-level baseline precision variables or effect modifiers researchers wish to account for.

In this paper, we have focused on Normal outcomes with identity links. However, CL's can be applied to a wide variety of outcomes [37]. Of particular importance for SWTs is the binary outcome with a logit link. One important difference from the results presented here is that the CL produces a marginal treatment estimate, analogous to the GEE. On the logit scale, this is no longer asymptotically equal to the conditional treatment effect estimates from mixed models, so using the correctly specified mixed model as a 'gold standard' of efficiency may not be appropriate since the estimands are different. In this case, the GEE with a correctly specified working correlation structure may be a better comparison for understanding the efficiency of CLs.

The CL models we presented here used only vertical contrasts. One potential way to improve efficiency would be to incorporate some horizontal contrasts. Kennedy-Shaffer et al. [35] examined ensemble estimators that combined vertical and horizontal estimators, in an attempt to improve efficiency. This method performed well, although the authors note that this is dependent on the settings. In a similar manner, it would be possible to construct CL models that combine vertical and horizontal information by multiplying together component likelihoods that represent each piece of information. This would allow researchers to explicitly control the relative importance of the vertical and horizontal information, and it may be possible to construct a more robust CL estimator that performs well when either the vertical or horizontal estimates are correct [57]. This is a possible advantage over GEE and other

traditional methods. Adding horizontal information to the GEE in Equation 3.4 by e.g. using a working exchangeable correlation structure induces some weighted average between horizontal and vertical information, but those weights may not be apparent or desirable. However, inference on these CL methods may be very complicated, and simpler methods of using horizontal information (e.g. Equation 3.6) may be preferable.

3.6.3 Concerns for application

Although the results presented in this paper suggest that CL methods can be a useful tool for analyzing SWTs, some practical concerns need to be addressed and further clarified to ensure that researchers using CLs can continue to follow modern best practices and guidelines. One important issue is how to estimate sample size for a CL model. It is possible to calculate the expected asymptotic variance of the treatment effect under a simple data-generating correlation structure (e.g. exchangeable). Although this would be a valid way to estimate sample size under simple models, we expect that researchers pursuing a CL model are concerned about more complex correlation structures. In this case, simulation-based sample size calculations with more realistic correlation structures may be more appropriate.

Another important issue is effective reporting in published analyses. Because of the rigid structure and established terminology of GEE and mixed models, authors often accurately describe their analytic models without writing out the likelihood or estimating equations. Because of the rich variety of CL models and their relative novelty, we suggest that authors make the full composite likelihood available to readers, to ensure that the statistical methods and precise assumptions are clear.

In reporting results, one important difference from a GEE with working exchangeable correlation or mixed model with random cluster intercepts is that an intra-cluster correlation (ICC) is not necessarily estimated with a CL model. However, to inform future studies and adhere to CONSORT guidelines [13] it is helpful to report an ICC. In some cases, the assumptions behind an ICC may contradict the assumptions made in the CL. For example, if researchers were worried about a random time effect and constructed a likelihood with

different variances for each time period, this is incompatible with an ICC, which assumes variances are constant over time. In this case, researchers may have to resort to fitting a separate GEE or CL with an exchangeable correlation structure to estimate an ICC.

Last, a difficult prospect for using a CL model is the choice of model, especially in a pre-specified statistical analysis. We suggested many extensions to the basic CLWP model in the section above, but there is little guidance in this context for how to balance the desire for a simple model with the desire to construct a CL that closely reflects the assumed data-generating structure. If researchers are interested in using CL models, it would be helpful for future research to provide more guidance about model choice based on researchers' beliefs about the data-generating structure.

3.6.4 Final comments

These novel within-period composite likelihood estimators are more robust than traditional mixed models, and in some contexts may be preferable to other existing non-parametric vertical estimators. For Type I error control, semi-parametric methods like CLs or specific GEE models may be preferable to NPWP when there is a random treatment effect. For efficiency, the CL adjusted for baseline outcomes is preferable to the NPWP in all the scenarios we examined. We hope these composite likelihood estimators are useful options and avenues for future research for researchers concerned about misspecification in SWT.

Acknowledgments

The following individuals have contributed significantly to this chapter: Avi Kenny, Fan Xia, Patrick Heagerty, James Hughes

Research reported in this paper was supported in part by the National Institutes of Health under award number AI029168. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Research reported in this publication was completed by investigators at the University of Washington's Clinical Learning, Evidence And Research (CLEAR) Center for Muscu-

loskeletal Disorders and supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) of the National Institutes of Health under Award Number P30AR072572. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data motivating the simulations is freely available from the COVID-19 data repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [53].

Chapter 4
CONCLUSION

In this paper, we have attempted to address three important issues in the design and analysis of SWTs by: (1) providing tools and guidance to help researchers apply existing methods correctly; (2) quantifying the impact of misspecification in mixed models so that researchers can use that information in statistical analysis plans and sensitivity analyses; and (3) proposing alternative vertical methods for analyzing SWTs that are robust to common misspecification concerns.

Although we have provided some insights, it is still a major challenge for applied researchers to make informed pre-specified decisions about appropriate methods and models when information about the underlying correlation structure is limited. We hope that new research such as the CLOUD database continues to address this [27]. As research about the impact of misspecification and more robust methods continues to develop, one important issue is how statistical concerns and choices should impact the design of SWTs. For example, the 'classic' SWT has been the standard for mixed models, but would not be appropriate if the primary analysis used only vertical information. Research that has been done on optimal SWT designs for traditional methods can now be extended to these new methods. In both the second and third projects, we observed that relatively minor changes in the SWT design can have major implications. Although trial design is often driven by logistical constraints and practical concerns, it is also important to understand the impact of the design choices on the analysis. Improving this understanding and clarifying how it might interface with trial design is one potential area for future research.

We hope that the research presented here provides valuable tools and insight for researchers designing and analyzing SWTs.

BIBLIOGRAPHY

[1] Barker D, McElduff P, D'Este C, and Campbell MJ. Stepped wedge cluster randomised trials: A review of the statistical methodology used and available. *BioMed Central Medical Research Methodology*, 16(69), 2016.

[2] Hughes JP, Granston TS, and Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. *Contemporary Clinical Trials*, 45:55–60, 2015.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[4] Hughes J, Hakhu NR, and Voldal E. *swCRTdesign: Stepped Wedge Cluster Randomized Trial (SW CRT) Design*, 2019. R package version 3.1.

[5] Golden MR, Kerani RP, Stenger M, Hughes JP, Aubin M, Malinski C, and Holmes KK. Uptake and population-level impact of expedited partner therapy (ept) on *Chlamydia trachomatis* and *Neisseria gonorrhoeae*: The washington state community-level randomized trial of ept. *PLoS Medicine*, 12(1):e1001777, 2015.

[6] Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28(2):182–191, 2007.

[7] Baio G, Copas A, Ambler G, Hargreaves J, Beard E, and Omar RZ. Sample size calculation for a stepped wedge trial. *Trials*, 16(354), 2015.

[8] Woertman W, de Hoop E, 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*, 66(7):752–758, 2013.

[9] Trutschel D and Treutler H. *samplingDataCRT: Sampling Data Within Different Study Designs for Cluster Randomized Trials*, 2017. R package version 1.0.

[10] Baio G and Leech R. *SWSamp*, 2019. R package version 0.3.1.

[11] Hemming K, Kasza J, Hooper R, Forbes A, and Taljaard M. A tutorial on sample size calculation for multiple-period cluster randomised parallel, cross-over, and stepped wedge trials using the shiny crt calculator. *International Journal of Epidemiology*, 49(3):979–995, 2020.

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

[13] Hemming K, Taljaard M, McKenzie JE, Hooper R, Copas A, Thompson JA, Dixon-Woods M, Aldcroft A, Doussau A, Grayling M, Kristunas C, Goldstein CE, Campbell MK, Girling A, Eldridge S, Campbell MJ, Lilford RJ, Weijer C, Forbes AB, and Grimshaw JM. Reporting of stepped wedge cluster randomised trials: Extension of the consort 2010 statement with explanation and elaboration. *BMJ*, 363(k1614), 2018.

[14] Hemming K, Taljaard M, and Forbes A. Modeling clustering and treatment effect heterogeneity in parallel and stepped-wedge cluster randomized trials. *Statistics in Medicine*, 37(6):883–898, 2018.

[15] Chang W, Cheng J, Allaire J, Xie Y, and McPherson J. *shiny: Web Application Framework for R*, 2019. R package version 1.3.2.

[16] Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[17] Thompson JA, Fielding KL, Davey C, Aiken AM, Hargreaves JR, and Hayes RJ. Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. *Statistics in Medicine*, 36(23):3670–3682, 2017.

[18] Verbeke G and Molenberghs G. *Linear mixed models for longitudinal data*. Springer, 2000.

[19] Cheng J, Edwards LJ, Maldonado-Molina MM, Komro KA, and Muller KE. Real longitudinal data analysis for real people: building a good enough mixed model. *Statistics in Medicine*, 29(4):504–520, 2010.

[20] Moerbeek M. The consequences of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, 39(1):129–149, 2004.

[21] Haines TP, Bowles KA, Mitchell D, O'Brien L, Markham D, Plumb S, May K, Philip K, Haas R, Sarkies MN, Ghaly M, Shackell M, Chiu T, McPhail S, McDermott F, and Skinner EH. Impact of disinvestment from weekend allied health services across acute medical and surgical wards: 2 stepped-wedge cluster randomised controlled trials. *PLoS Medicine*, 14(10):e1002412, 2017.

[22] Heagerty PJ and Kurland BF. Misspecified maximum likelihood estimates and generalised linear models. *Biometrika*, 88(4):973–985, 2001.

[23] Inc Wolfram Research. *Mathematica*. Wolfram Research, Inc, Champaign, Illinois, 2020.

[24] Huber PJ. The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 5(1):221–233, 1967.

[25] Thompson JA, Hemming K, Forbes A, Fielding K, and Hayes R. Comparison of small-sample standard-error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: A simulation study. *Statistical Methods in Medical Research*, 30(2):425–439, 2021.

[26] Chavance M and Escolano S. Misspecification of the covariance structure in generalized linear mixed models. *Statistical Methods in Medical Research*, 25(2):630–643, 2016.

[27] Korevaar E, Kasza J, Taljaard M, Hemming K, Haines T, Turner EL, Thompson JA, Hughes JP, and Forbes AB. Intra-cluster correlations from the clustered outcome dataset bank to inform the design of longitudinal cluster trials. *Clinical Trials*, 18(5):529–540, 2021.

[28] Grantham KL, Kasza J, Heritier S, Hemming K, and Forbes AB. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. *Statistics in Medicine*, 38(11):1918–1934, 2019.

[29] Zhang D and Davidian M. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57(3):795–802, 2001.

[30] Leyrat C, Morgan KE, Leurent B, and Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? *International Journal of Epidemiology*, 47(1):321–331, 2018.

[31] 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*, 26(2):583–597, 2017.

[32] Matthews JNS and Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. *Statistics in Medicine*, 36(24):3772–3790, 2017.

[33] Hughes JP, Heagerty PJ, Xia F, and Ren Y. Robust inference for the stepped wedge design. *Biometrics*, 76(1):119–130, 2020.

[34] Thompson JA, Davey C, Fielding K, Hargreaves JR, and Hayes RJ. Robust analysis of stepped wedge trials using cluster-level summaries within periods. *Statistics in Medicine*, 37(16):2487–2500, 2018.

[35] Kennedy-Shaffer L, de Gruttola V, and Lipsitch M. Novel methods for the analysis of stepped wedge cluster randomized trials. *Statistics in Medicine*, 39(7):815–844, 2020.

[36] Thompson J, Davey C, Hayes R, Hargreaves J, and Fielding K. swpermute: Permutation tests for stepped-wedge cluster-randomized trials. *The Stata Journal*, 19(4):803–819, 2019.

[37] Varin C, Reid N, and Firth D. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

[38] Li F, Yu H, Rathouz PJ, Turner EL, and Preisser JS. Marginal modeling of cluster-period means and intraclass correlations in stepped wedge designs with binary outcomes. *Biostatistics*, 23(3):772–788, 2022.

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

[40] Lee Y and Nelder JA. Conditional and marginal models: another view. *Statistical Science*, 19(2):219–238, 2004.

[41] Wang R and De Gruttola V. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. *Statistics in Medicine*, 36(18):2831–2843, 2017.

[42] Ji X, Fink G, Robyn PJ, and Small DS. Randomization inference for stepped-wedge cluster-randomized trials: an application to community-based health insurance. *The Annals of Applied Statistics*, 11(1):1–20, 2017.

[43] Fai AH and Cornelius PL. Approximate f-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation*, 54(4):363–378, 1996.

[44] Kenny A, Voldal E, Xia F, Heagerty PJ, and Hughes JP. Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect. *Statistics in Medicine*, 2022.

[45] Tian Z, Preisser JS, Esserman D, Turner EL, Rathouz PJ, and Li F. Impact of unequal cluster sizes for gee analyses of stepped wedge cluster randomized trials with binary outcomes. *Biometrical Journal*, 64(3):419–439, 2022.

[46] Miglioretti DL and Heagerty PJ. Marginal modeling of nonnested multilevel data using standard software. *American Journal of Epidemiology*, 165(4):453–463, 2007.

[47] Heagerty PJ and Lele SR. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111, 1998.

[48] Molenberghs G and Verbeke G. *Models for discrete longitudinal data*. Springer, 2005.

[49] Hooper R, Forbes A, Hemming K, Takeda A, and Beresford L. Analysis of cluster randomised trials with an assessment of outcome at baseline. *BMJ*, 360:k1121, 2018.

[50] Yang L and Tsiatis AA. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

[51] Lin W. Agnostic notes on regression adjustments to experimental data: reexamining freedman's critique. *Annals of Applied Statistics*, 7(1):295–318, 2013.

[52] Wang B, Ogburn EL, and Rosenblum M. Analysis of covariance in randomized trials: more precision and valid confidence intervals, without model assumptions. *Biometrics*, 75(4):1391–1400, 2019.

[53] Dong E, Du H, and Gardner L. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, 2020.

[54] Jarvik JG, Meier EN, James KT, Gold LS, Tan KW, Kessler LG, Suri P, Kallmes DF, Cherkin DC, Deyo RA, Sherman KJ, Halabi SS, Comstock BA, Luetmer PH, Avins AL, Rundell SD, Griffith B, Friedly JL, Lavallee DC, Stephens KA, Turner JA, Bresnahan BW, and Heagerty PJ. The effect of including benchmark prevalence data of common imaging findings in spine image reports on health care utilization among adults undergoing spine imaging: a stepped-wedge randomized clinical trial. *JAMA Network Open*, 3(9):e2015713, 2020.

[55] 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*, 16:358, 2015.

[56] Varin C and Czado C. A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics*, 11(1):127–138, 2010.

[57] Canova F and Matthes C. A composite likelihood approach for dynamic structural models. *The Economic Journal*, 131(638):2447–2477, 2021.

[58] Bates D, Mächler M, Bolker B, and Walker S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

Appendix A

APPENDIX FOR PROJECT 2

A.1 Relative frequencies of roots

To understand the practical importance of the roots of Equation 2.3, we use simulations to examine the relative frequencies of different roots under a variety of circumstances. We used the `lme4` package[58] (lmer command) in `R` (version 3.6.1) [3] to fit mixed models, but these results may be different for other software. For example, one important characteristic of the `lme4` package (version 1.1-21) is that it allows fitted variance components to be exactly zero; other software packages that do not allow this may have a dramatically different distribution of roots. For these simulations, we used all the default settings in `lme4`, including fitting models using restricted maximum likelihoods. In this appendix, we present results from two designs: one minimal SWT design, and one large SWT design. The minimal design is a classic design with two sequences, six clusters per sequence, and two individuals per cluster per time point. The large design is a classic design with six sequences, ten clusters per sequence, and 100 individuals per cluster per time point. So the total number of clusters is 12 and 60 for the minimal and large designs, respectively; the total number of individual observations is 72 and 42,000. We also considered different true values of the variance components. Fixing $\sigma_t^2 = 1$, we allowed τ_t^2, γ_t^2, and η_t^2 to be either 0.001 or 0.125. For both cases, we examined all four combinations of large and small variance components. Throughout, we fit a model with a minimal number of fixed effects (see Section 2.2). For each setting, relative frequencies were calculated based on 1,000 replications.

The results of these simulations are presented in Figure A.1. For both misspecification cases, prevalence of Root 1 increased as the total sample size MNK increased. In some scenarios with sufficiently large total sample size, Root 1 was the only root observed. For

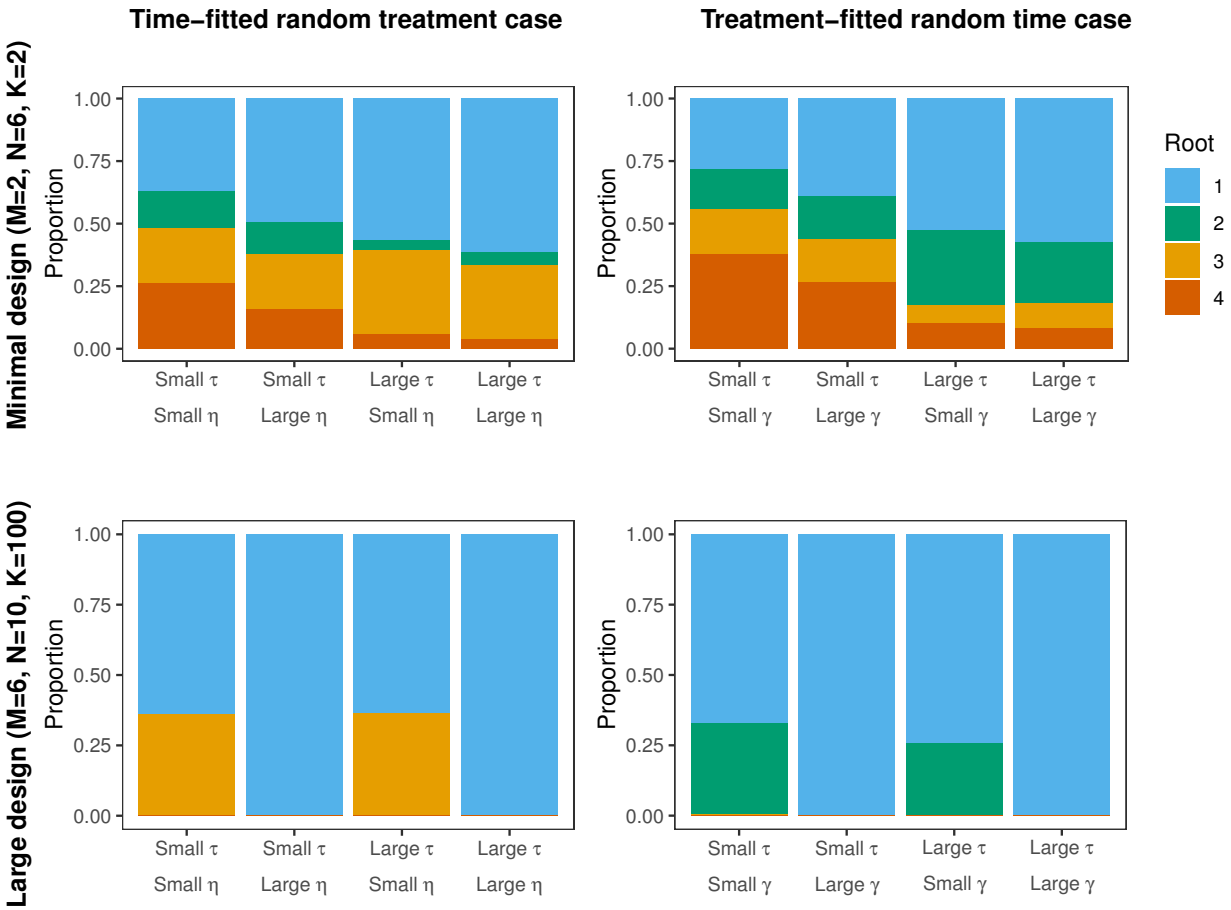

Figure A.1: Frequency at which each root is observed in simulations using `lme4`'s default settings. Throughout, $\sigma_t^2 = 1$. For the 'large' values of τ_t^2, γ_t^2, and η_t^2, we used 0.125. For the 'small' values, we used 0.001. Note that these correspond to ACCs ranging between 0.001 and 0.16 for the time-fitted random treatment case, and ranging between 0.002 and 0.20 for the treatment-fitted random time case.

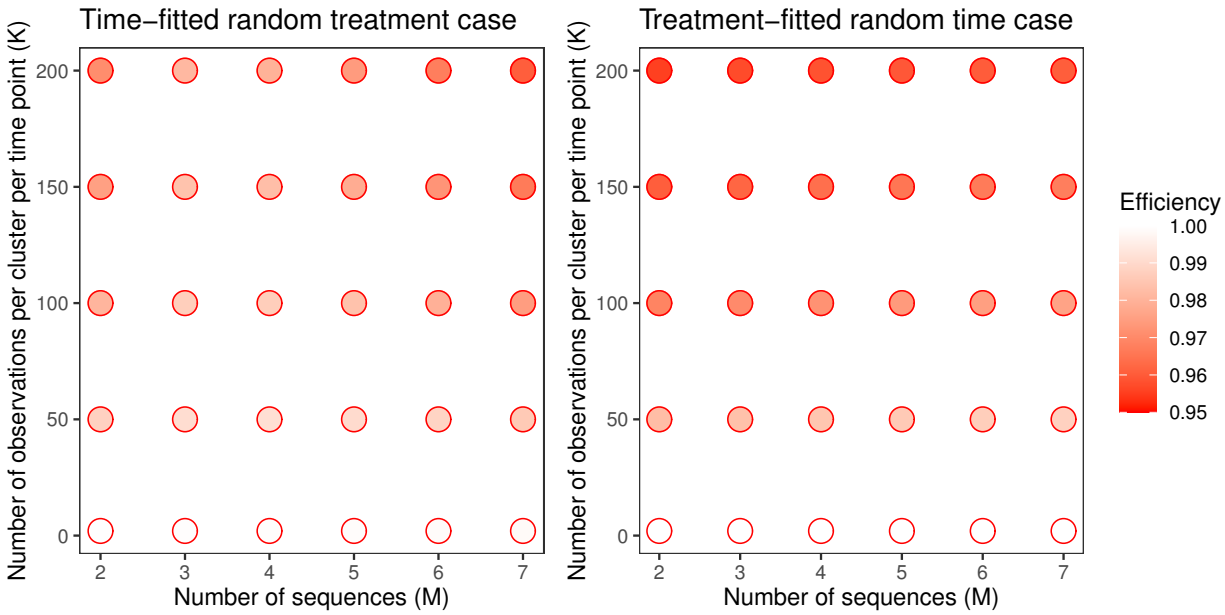

Figure A.2: Efficiency of Root 1 for both cases, for a variety of classic designs. For each case, $\sigma_t^2 = 5$ and $\tau_t^2 = 0.1$. To keep the ACC consistent, we chose $\eta_t^2 = 0.1$ and $\gamma_t^2 = 0.05$ for the time-fitted random treatment and treatment-fitted random time cases, respectively.

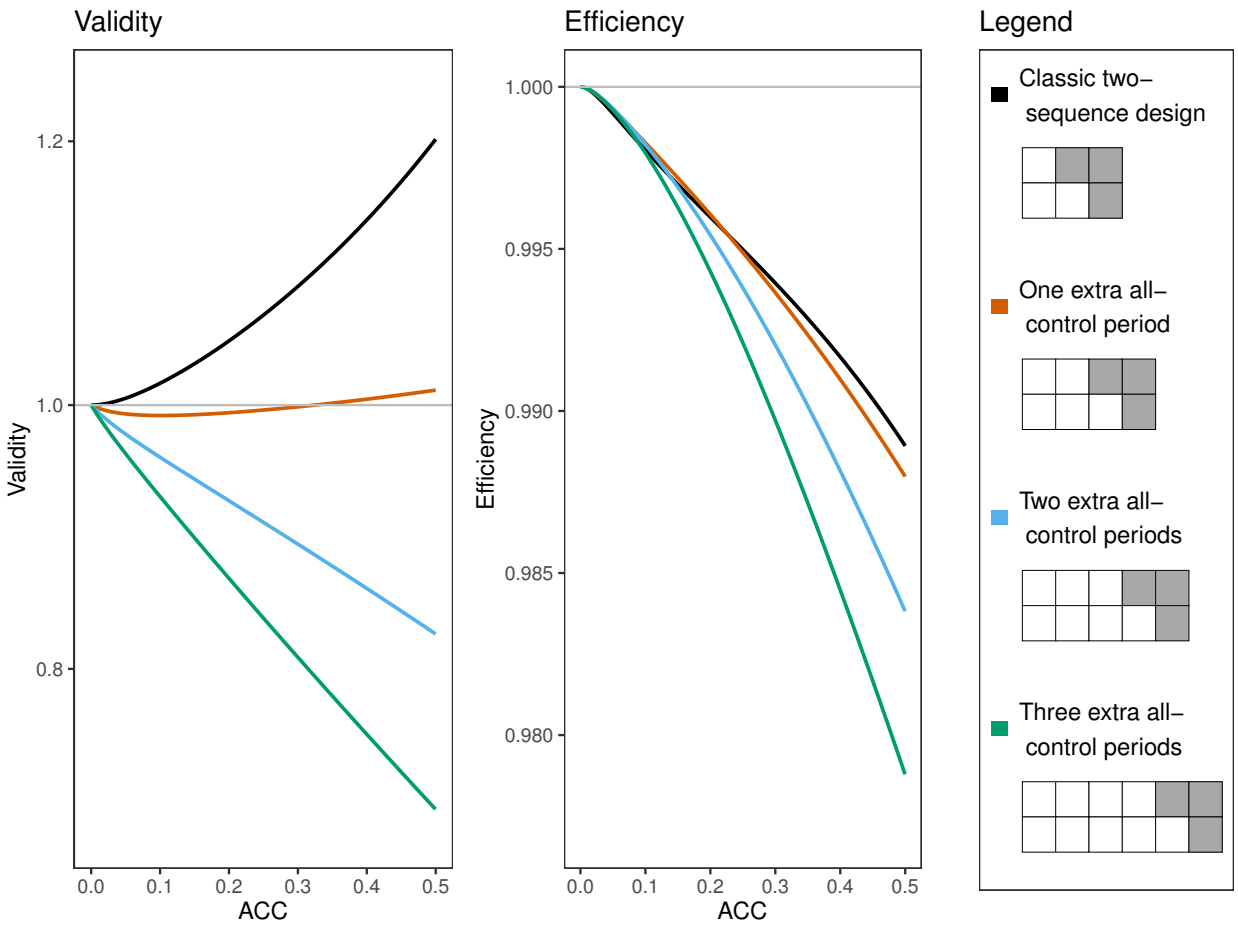

Figure A.3: Validity and efficiency of Root 1 for the time-fitted random treatment case, for four different SWT designs and a variety of true ACC's. The designs each have two sequences and $K = 5$ observations per cluster per time period. Throughout, $\sigma_t^2 = 1$ and the balance of τ_t^2 and η_t^2 is fixed at $\eta_t^2 = \tau_t^2/2$.

the time-fitted random treatment case when the random treatment effect was small, Root 3 was relatively common even when the total sample size was large. Roots 2 and 4 were only present when the total sample size was small. For the treatment-fitted random time case when the random time effect was small, Root 2 was relatively common even when the total sample size was large. Roots 3 and 4 were only present when the total sample size was small.

Through careful use of software settings and other strategies [19], it is possible to increase the prevalence of Root 1. Thus, these results are a 'lower bound' on how common Root 1 is. Additionally, we intentionally included settings which increase the presence of other roots (i.e. a very small minimal design, and a very small variance component lower bound). We would expect reasonable settings for most SWTs to be less extreme, and have a higher prevalence of Root 1. The results of these simulations support the decision to focus on Root 1 in this chapter.

A.2 Non-classic designs

For simplicity, we have focused primarily on classic designs. However, there are a wide variety of SWT designs that do not follow the standard crossover schedule that typifies a 'classic' SWT. For example, some SWTs may observe all sequences in the control condition for multiple time periods before beginning crossover. Figure A.3 compares the validity and efficiency for some non-classic designs that have differing numbers of all-control periods as an example of how relatively simple design changes can impact inference. Because the design elements of a SWT are inexorably linked, it is difficult to isolate the effects of design choices. For example, in Figure A.3, adding extra all-control periods also increases the total number of time points, increases the total number of observations, and changes how much of the total variation is attributed to the random treatment effect (by changing the proportion of cluster-periods that are assigned to treatment $\frac{1}{JM}\Sigma_{m=1}^{M}T_m$). One plausible explanation for the association between worsening validity and adding extra control periods in Figure A.3 is that when $\frac{1}{JM}\Sigma_{m=1}^{M}T_m$ is close to zero, most cluster-periods have no additional variation beyond a random intercept, which causes the variance of a fitted random time effect to

be close to zero. In contrast, in a classic design where $\frac{1}{JM}\Sigma_{m=1}^{M} T_m = \frac{1}{2}$, half of the cluster-periods have no additional variation and half have some additional variation from the random treatment effect, so we might expect a fitted random time effect to be a more balanced average between those states that has more flexibility to account for variation coming from the random treatment effect. Although it is difficult to identify universal trends in non-classic designs because of their complexity, the supplemental materials provide code that allows researchers to explore many non-classic designs. One important type of non-classic design which is not covered in the supplemental materials is a design in which the number of observations per cluster-period (K) varies. Simulations suggest that having a sample size that differs by sequence and/or time period can affect validity and efficiency, possibly by affecting the relative weights of cluster-periods on treatment and control.

Appendix B

APPENDIX FOR PROJECT 3

B.1 Simulation setting details

In this hypothetical SWT, each cluster is a county, and each time period is one month long. We would be studying a treatment implemented at the county level (e.g. a policy change or public health intervention) that attempts to reduce the spread of COVID-19. We primarily use confirmed cases from Minnesota in 2021, originally reported as daily counts per county. For each of the 87 counties, we averaged the daily counts over each month, so one cluster-period is one county-month. We log-transformed the outcome after aggregating to remove a right-skew. We then fit several models to this data to obtain realistic parameter estimates, described below. Each setting is simulated under the (weak) null ($\theta = 0$), and under an alternative ($\theta < 0$), where the magnitude of the effect is chosen to have roughly 80% power under Equation 3.2.

Note that in this hypothetical SWT, the assumption of an instantaneous and constant treatment effect θ is not plausible due to the nature of infectious diseases. A more realistic design might be to have washout periods, or have more distance between time periods (e.g. only measuring the first month of each quarter). Alternatively, researchers could analyze the trial accounting for a time-varying treatment effect; this is relatively straightforward for mixed models, GEE's, and composite likelihoods [44].

To inform the settings in scenario A, we fit the model from Equation 3.2 to the Minnesota COVID-19 data. We used the parameter estimates from this model for the simulation settings: $\mu = 1.86$, $\beta_2, \ldots, \beta_{12} = -0.59, -0.32, 0.12, -0.72, -2.73, -2.14, -0.18, 0.57, 0.79, 1.12, 0.73$, $\sigma = 0.43$, and $\tau = 1.27$. Throughout, for the 3-sequence design we used the first month of each quarter for the time

trends (January, April, July, and October). We simulate the cluster-period means directly, so we do not specify the cluster-period size K. For context, these variance components are consistent with an individual-level intra-cluster correlation (ICC) of .47 (if $K = 10$), .08 (if $K = 100$), or .02 (if $K = 500$).

To inform the settings in scenario B, we fit a model like Equation 3.2, but with an AR(1) correlation structure within clusters. The parameter estimates from this model for the fixed effects, σ, and τ were very similar to those from setting A, so for simplicity we kept those settings the same as scenario A. The model we fitted to the COVID-19 data estimated that in the AR(1) correlation, $\phi = 0.31$. So for two observations from the same cluster t time points apart, $corr(Y_{ij}, Y_{ij+t}|u_i) = \phi^t$.

For scenario C, since there was no actual SWT in the COVID-19 data we subjectively chose $\eta^2 = \tau^2/2$ so that the random treatment effect would be sizable, but not larger than the random cluster intercept effect.

For scenario D, we repeated all the data processing steps for counties in Florida, and fit the model from Equation 3.2 to data from Florida. From that model, we used the estimated time trend for our settings, which was $\mu = 4.16$ and $\beta_2, \ldots, \beta_{12} = -0.81, -1.50, -1.36, -1.85, -2.39, -0.46, 0.52, 0.15, -1.30, -2.55, -0.76$. The variance parameters from the Florida data were broadly similar to those from the Minnesota data, so we kept those settings the same as scenario A for simplicity.

B.2 Simulations: varying sample sizes

In order to understand whether it is appropriate to apply these methods to the LIRE trial data, we perform some limited simulations with varying cluster-period sample sizes. We use the simulation settings from scenario A described above, but simulate cluster-period sample sizes from a lognormal distribution with mean 3 and standard deviation between .001 (essentially constant cluster-period size) and .9 (very right-skewed cluster-period size, like the LIRE trial). We use the 3-sequence design, with 22 clusters per sequence (as in the original simulation settings) and 33 clusters per sequence (a total number of clusters

similar to the LIRE trial). Figure B.1 shows the results of these simulations under the null hypothesis ($\theta = 0$). There does not appear to be a strong relationship between the Type I error rate and the variability of cluster-period size for any of the methods. However, this is not an exhaustive analysis and many more complicated data-generating scenarios could be explored in future research (e.g. cluster size correlated with time, or cluster size correlated with random intercepts).

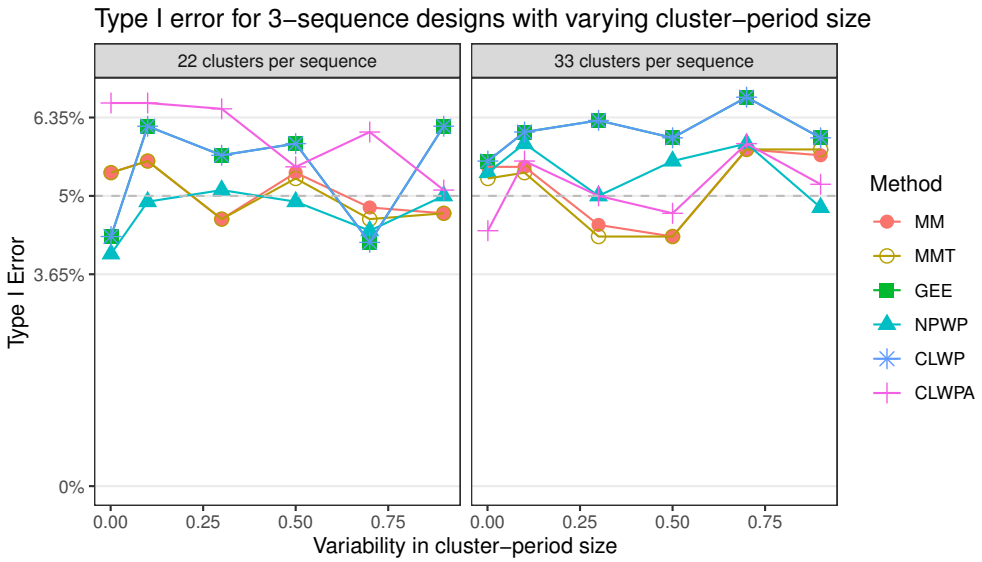

Figure B.1: Type I error from a SWT with 3 sequences and either 22 or 33 clusters per sequence. The simulated data had random cluster intercepts (scenario A), and no treatment effect ($\theta = 0$). Data was analyzed using a mixed model with random cluster intercepts (MM), mixed model with random cluster intercepts and random treatment (MMT), GEE with working independence (GEE), non-parametric within-period (NPWP), composite likelihood within-period model (CLWP), and CLWP adjusted for baseline (CLWPA).

B.3 Possible extensions

B.3.1 Modeling varying sample sizes

In the scenarios above, we considered trials that had a constant number of individuals per cluster-period, K. In reality, the number of individuals measured often varies both between clusters and over time. Let K_{ij} be the number of individuals measured in cluster i at time j. Then assuming the residual variance at the individual level σ^2_{ind} is constant, the residual variance for each cluster-period mean would be σ^2_{ind}/K_{ij}. The CL models in Equations 3.5 and 3.6 assume that variance is constant over all vertical contrasts, which is no longer true. If differences in sample size are very small across cluster-periods, this simple model might still perform well. But if cluster-period sample sizes vary widely, we might want to account for that in our CL model. In addition to allowing for varying cluster-period sizes in the likelihood components, weighting the likelihood components according to the sample size they represent could improve efficiency.

Note that under an exchangeable correlation matrix (i.e. Equation 3.2), the variance of an arbitrary vertical contrast is

$$\begin{aligned} Var(Y_{ij} - Y_{i'j}) &= Var(Y_{ij}) + Var(Y_{i'j}) \\ &= \sigma^2_{ind}/K_{ij} + \tau^2 + \sigma^2_{ind}/K_{i'j} + \tau^2 \\ &= (1/K_{ij} + 1/K_{i'j})\sigma^2_{ind} + 2\tau^2 \end{aligned}$$

More complicated data-generating models can be considered, but many will fall into a similar form of the sum of several terms, only one of which ($(1/K_{ij} + 1/K_{i'j})\sigma^2_{ind}$) involves the cluster-period sample sizes. For example, under Equation 3.3 $Var(Y_{ij} - Y_{i'j}) = (1/K_{ij} + 1/K_{i'j})\sigma^2_{ind} + 2\tau^2 + \eta^2$. Thus we consider Equation 3.2 for simplicity, but expect these would work for a larger family of data-generating models with independent clusters. Characteristics that would violate this form include correlated clusters and random effects that depend on cluster size.

Stratify

One option for handling a variable cluster-period sample size would be to stratify vertical contrasts into groups with similar $1/K_{ij} + 1/K_{i'j}$, and allow each group to have its own variance parameter. For example, suppose we decide to divide vertical contrasts into two groups based on some cut point k_{cut}. Then we can revise Equation 3.5 so that each group has its own estimated variance parameter, e.g.

$$\mathscr{L}_{CLWP}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} f(\Phi; Y_{ij} - Y_{i'j})$$

where $f(\Phi; Y_{ij} - Y_{i'j})$ is a Normal density with mean θ and some variance σ_1^2 if $1/K_{ij} + 1/K_{i'j} < k_{cut}$ and variance σ_2^2 if $1/K_{ij} + 1/K_{i'j} \geq k_{cut}$. We would expect this model to perform well if $1/K_{ij} + 1/K_{i'j}$ does not vary too much within a group, and there are not too many variance parameters. Choosing the number of groups and location of cut points may be challenging a priori.

Parametric model

Another option is to partially specify the form of the variance of vertical contrasts, and explicitly include sample sizes. For example,

$$\mathscr{L}_{CLWP}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} f(\Phi; Y_{ij} - Y_{i'j})$$

where $f(\Phi; Y_{ij} - Y_{i'j})$ is a Normal density with mean θ and variance $(1/K_{ij} + 1/K_{i'j})\sigma_3^2 + \sigma_4^2$. We would expect the MCLE estimates of σ_3^2 and σ_4^2 to approximate the individual-level residual variance and some function of the random effects, respectively. Disadvantages of this option include: it does not work for every data-generating model; the calculation of the standard errors is somewhat more complicated, as it now involves varying cluster-period sample sizes; and it requires estimating two variance parameters which may be very close to a boundary, which may cause convergence issues.

B.3.2 Using more SW-specific prior info

Most recent information

$$\mathscr{L}_{CLWPA}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} f(\Phi; Y_{ij} - Y_{i'j} | Y_{ij'} - Y_{i'j'}) \tag{B.1}$$

where $f(\Phi; Y_{ij} - Y_{i'j} | Y_{ij'} - Y_{i'j'})$ is a Normal density with mean $\theta + \gamma(Y_{ij'} - Y_{i'j'})$ and some variance σ^2_{CLWPA}. Here, j' is the most recent time period in which $X_{ij'} = X_{i'j'} = 0$. Using the most recent information is appealing if correlations between cluster-periods decay over time within a cluster.

If the decay is strong enough, the correlation between $Y_{ij} - Y_{i'j}$ and $Y_{ij'} - Y_{i'j'}$ might depend heavily on $j - j'$. The same issue might occur in the model using baseline control adjustment (Equation 3.6), with dependence on $j - 1$. In either case, it might be possible to create a more accurate model by fitting separate coefficients based on $j - j'$ or $j - 1$. However, if the decay is not very strong (i.e. the correlation structure is close to exchangeable), the cost of fitting additional parameters might outweigh the increase in precision from adjusting for a more accurate model of previous control contrasts. It might also be possible to reparameterize γ into a formula involving more parameters and $|j - j'|$ or $|j - 1|$. Depending on the trial design and suspected correlation structure, this might involve fewer parameters than fitting separate unstructured coefficients.

Average information

If the data-generating model is close to exchangeable, the amount of information from different time periods with $X_{ij} = X_{i'j} = 0$ might be very similar. In this case, it might be appealing to incorporate all the information by averaging all the previous control contrasts, e.g.

$$\mathscr{L}_{CLWPA}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} f(\Phi; Y_{ij} - Y_{i'j} | \frac{1}{j_{last}} \sum_{j'=1}^{j'=j_{last}} Y_{ij'} - Y_{i'j'}) \tag{B.2}$$

where j_{last} is the maximum j' such that $X_{ij'} = X_{i'j'} = 0$, so $f(\Phi; Y_{ij} - Y_{i'j}|\frac{1}{j_{last}} \sum_{j'=1}^{j'=j_{last}} Y_{ij'} - Y_{i'j'})$ is a Normal density with mean $\theta+\gamma(\frac{1}{j_{last}} \sum_{j'=1}^{j'=j_{last}} Y_{ij'}-Y_{i'j'})$ and some variance σ^2_{CLWPA}.

Variable number of adjustments

One characteristic of SWTs is that each sequence has a different number of observed cluster-periods in the control condition. If the correlation structure is exchangeable, it may be sensible to aggregate these (see above). If the correlation structure is autoregressive and decays very quickly, adjusting for the most recent control may be desirable (see above). Suppose we believe our correlation structure is autoregressive, but we have a very long trial and don't believe that there would be meaningful correlation between time points more than 2 units apart. We might construct the following CL:

$$\mathscr{L}_{CLWPA}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} f(\Phi; Y_{ij} - Y_{i'j}|Y_{ij'} - Y_{i'j'}) \tag{B.3}$$

where j' is the most recent time period in which $X_{ij'} = X_{i'j'} = 0$ with $j - j' \leq 2$. $f(\Phi; Y_{ij} - Y_{i'j}|Y_{ij'} - Y_{i'j'})$ is a Normal density with mean $\theta+\gamma_1(Y_{ij'} - Y_{i'j'})1_{j-j'=1}+\gamma_2(Y_{ij'} - Y_{i'j'})1_{j-j'=2}$ and variance σ_1^2 or σ_2^2 for $j - j' = 1$ or $j - j' = 2$, respectively. If there is no time period with $X_{ij'} = X_{i'j'} = 0$ in the desired range, then an unadjusted density is used, so the component likelihood is a Normal density with mean θ and variance σ_0^2. In this model, some vertical contrasts (i.e. those from time periods well after the relevant crossover period) will have no additional covariates in their mean model, while others will have a γ_1 (meaning that the treated cluster just crossed over) or γ_2 (meaning that the treated cluster crossed over two time periods ago).

The advantage of this is that it uses only the most relevant horizontal information, making minimal assumptions about the form of the correlation. Depending on the correlation structure of the data, this may be more efficient than e.g. adjusting for baseline as we did in the CLWPA.

B.3.3 Individual-level vertical CL

The models in this paper focused on cluster-period means, which is a valid method of inference and may be computationally attractive if there are many observations in each cluster-period. However, there may be scenarios in which researchers want to use an individual-level model. For example, perhaps there are important individual-level precision variables or effect modifiers.

As an example of how an individual-level CL might be constructed, here is a CL with working independence:

$$\mathscr{L}_{CLWPI}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} \prod_{k_i=1}^{K} \prod_{k_{i'}=1}^{K} f(\Phi; Y_{ijk_i} - Y_{i'jk_{i'}})$$

where $f(\Phi; Y_{ijk_i} - Y_{i'jk_{i'}})$ is a Normal density with mean θ and some variance σ^2_{CLWPI}, and Φ is the vector of all parameters in the model.

The individual-level model with working independence also assumes working independence between individuals from the same cluster-period. This is fairly unrealistic, and there might be a way to improve efficiency by modeling the joint distribution of individuals within a cluster-period and assuming e.g. exchangeability of individuals within cluster-periods.

B.3.4 Pairwise composite likelihood

In this paper, we use conditional likelihoods to link observations across time from the same cluster. Specifically, we included an adjustment for baseline differences in the mean model. An alternative way to link observations across time from the same cluster is through joint distributions. Using only pairwise distributions allows less of the covariance matrix to be specified in the composite likelihood, which imposes fewer assumptions compared to using joint distributions of e.g. all observations from a particular cluster. We write out one possible way of constructing a pairwise vertical CL here, but acknowledge that there are many ways to do this and further research would be needed to understand these options.

This is a pairwise CL where each component likelihood links vertical contrasts from adjacent periods of time:

$$\mathscr{L}_{CLWP}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} f(\Phi; Y_{ij} - Y_{i'j}, Y_{ij-1} - Y_{i'j-1}) \tag{B.4}$$

where $f(\Phi; Y_{ij} - Y_{i'j}, Y_{ij-1} - Y_{i'j-1})$ is a multivariate Normal density with mean $(\theta, \theta X_{ij-1})^T$ and a covariance matrix

$$\sigma^2_{CLWP} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Note that this does not allow for a random treatment effect, as the correlation between adjacent vertical contrasts is the same regardless of whether the contrast in time $j-1$ is between two untreated clusters ($X_{ij-1} = X_{i'j-1} = 0$) or a treated and control cluster ($X_{ij-1} = 1, X_{i'j-1} = 0$). This could be relaxed by adding more variance parameters, or restricting the pairwise likelihoods to only include certain types of pairs. Modeling the correlation between time-adjacent vertical contrasts could improve the efficiency of the CL compared to a full likelihood. Depending on the strength of correlation between time points, it could even be helpful to include component likelihoods connecting partially matching pairs, e.g. $Y_{ij-1} - Y_{i''j-1}$, $Y_{i''j-1} - Y_{i'j-1}$, and $Y_{i'j-1} - Y_{i''j-1}$. However, this will come at a cost of either making more assumptions about the correlation structure, or adding more variance parameters.

A pairwise model more analogous to Equation 3.6 would have pairwise components linking all the vertical contrasts to the baseline contrast, e.g.

$$\mathscr{L}_{CLWP}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} f(\Phi; Y_{ij} - Y_{i'j}, Y_{i1} - Y_{i'1}) \tag{B.5}$$

where $f(\Phi; Y_{ij} - Y_{i'j}, Y_{i1} - Y_{i'1})$ is a multivariate Normal density with mean $(\theta, 0)^T$ and a covariance matrix

$$\sigma^2_{CLWP} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Note that this correlation structure is only reasonable if cluster-periods are exchangeable within clusters, since the elapsed time since baseline varies. This pairwise method of accounting for baseline may have some advantages over the conditional method in Equation 3.6; for example, if the clusters vary in size this could be accounted for directly in the joint covariance matrix.

B.3.5 Time-varying treatment effect

Sometimes, researchers may be concerned about a treatment effect that varies over time, e.g. [44]. To adapt composite likelihood models to account for a time-varying treatment effect, we replace θ with θ_s, where s is the time since cross-over of the treated cluster. So, for example, a cluster with treatment assignment $(X_{i1}, X_{i2}, X_{i3}, X_{i4}) = (0, 1, 1, 1)$ would have a treatment effect in the mean model of $(0, \theta_1, \theta_2, \theta_3)$. So, if we adapted Equation 3.5 for a time-varying treatment effect, it would be

$$\mathscr{L}_{CLWP}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':X_{ij}=1,X_{i'j}=0} f(\Phi; Y_{ij} - Y_{i'j}) \tag{B.6}$$

where $f(\Phi; Y_{ij} - Y_{i'j})$ is a Normal density with mean θ_s where $s = \sum_{k=1}^{k=j} X_{ik}$ and some variance σ^2_{CLWP}.

Note that for classic designs, the model above cannot use information from the Jth time period because all clusters have crossed over. This means it cannot estimate θ_{J-1}, which is observed only in the first sequence in the last time period. Suppose we believe there are no random treatment effects (so variance is constant regardless of treatment assignment), and we want more information about the longer-term treatment effect. Then we could use vertical contrasts between treated cluster-periods which crossed over at different times. Let s_{ij} be the time since crossover for cluster i at time j, and let $s_{ij} = 0$ and $\theta_{s_{ij}} = 0$ for cluster-periods on control. Then the model below leverages more vertical contrasts:

$$\mathscr{L}_{CLWP}(\Phi; y) = \prod_{j=1}^{J} \prod_{i,i':s_{ij}>s_{i'j}} f(\Phi; Y_{ij} - Y_{i'j}) \tag{B.7}$$

where $f(\Phi; Y_{ij} - Y_{i'j})$ is a Normal density with mean $\theta_{s_{ij}} - \theta_{s_{i'j}}$ and some variance σ^2_{CLWP}. Unlike the original Equation 3.5 and other vertical methods, this model uses information from the Jth time period.