POWER ANALYSIS IN LONGITUDINAL ONE-WAY CROSSOVER STUDIES

BY YIRUI HU

A dissertation submitted to the

Graduate School – New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Donald R Hoover

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2016

ABSTRACT OF THE DISSERTATION

A Power Analysis in Longitudinal One-Way Crossover Studies

by YIRUI HU

Dissertation Director:
Donald R Hoover

This dissertation develops a general power estimation framework to estimate the variance of the new intervention effect estimate for longitudinal one-way crossover designs. Orthogonalized decomposition is developed for compound symmetry correlation of repeated measurements over time. In particular, we merge conventional difference-in-differences (DD) and more newly developed general stepped-wedge (SW) studies for both randomized and non-randomized allocation of units to the intervention, and investigate on the optimality properties in terms of study power (i.e. minimum variance of the intervention effect estimate). For a fixed total number of repeated measurements, we quantitatively compare the efficiency in detecting new intervention effect using DD and SW designs using formulas for compound symmetry covariance structure and empirical calculations for more general Toeplitz correlations. For this we provide insights for researchers in planning longitudinal one-way crossover designs.

The following thesis is composed of three chapters represented by three manuscripts. The first chapter develops a unified power estimation approach for continuous outcomes

in randomized difference-in-differences (R-DD) studies for both compound symmetry and more general Toeplitz correlation structures that were observed empirically. Optimal number of pre-and post-intervention allocation is analyzed. The second chapter extends the GLS power estimation framework to the non-randomized difference-in-differences (NR-DD) studies and quantitatively compare the penalty of being non-randomized versus randomized for a DD study. Optimal pre-post allocation is also analyzed for NR-DD studies. The third chapter, further investigates on the more general stepped-wedge designs and develop an Orthogonalized Least Squares power estimation framework for both randomized and non-randomized SW (R-SW and NR-SW). The third chapter is research conducted during graduate studies that has been accepted for publication in *Statistical Methods in Medical Research* published by SAGE.

Dedication

This dissertation is dedicated to people who have meant and continue to mean so much to me.

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Donald R Hoover, for his constant support, encouragement and trust. Dr. Hoover is an excellent adviser who genuinely cares for his students, which allowed me to grow in both professional and personal levels. I sincerely appreciate and value all I learned as his student, and the great opportunities to work on such innovative and interesting projects.

This dissertation was supported by NIH Grants 6U01AI035004, 6U01AI096299-06 and R01NR014632-01A1. I am extremely grateful for the support that I received from NIH. I am also grateful for all of my professors and fellow Ph.D. students here at Department of Statistics and Biostatistics who helped me during my days at Rutgers. Specifically, I would like to thank Dr. John Kolassa for his support in the past five years.

The members of my dissertation committee, Dr. Ying Hung, Dr. Han Xiao and Dr. Qiuhu Shi, have generously given their time and expertise to better my work. I thank them for their contribution and their good-natured support.

This journey would not have been possible without the support of my family, professors, and friends. I am especially grateful to my parents, who supported me emotionally. I always knew that you believed in me and wanted the best for me.

The following chapter was previously published in a peer-reviewed journal:
Chapter 3: Hu, Yirui, and Donald R. Hoover. Non-randomized and randomized stepped-wedge designs using an orthogonalized least squares framework. *Statistical Methods in Medical Research* (2016): 0962280216657852.

Table of Contents

List of Figures

List of Tables

General Introduction and Overview

To evaluate new intervention effect in studies, repeated measures of longitudinal continuous outcomes are gathered for the same units before and after intervention over several periods. These "units" could be persons where for example the continuous longitudinal outcomes could be weight, blood pressure, or average number of cigarettes smoked per day in the prior six months. The units could also be health care facilities such as hospitals where for example the continuous longitudinal outcomes could be portion of patients who are depressed, died from surgery or readmitted in 60 days after discharge. The intervention would be something designed to improve these outcomes (i.e. reduce high blood pressure or portion of patients that die in surgery.) As described below, for this dissertation the intervention is something that once started cannot be removed (i.e. one-way crossover).

One-way crossover studies are useful in longitudinal data analysis where as noted above, units can only switch from control to intervention in one direction; once switched they cannot switch back. Reasons for this include that: i) once policies are implemented, they cannot be undone, ii) after people are educated for new behaviors, they will not forget and/or iii) ethical/logistical concerns of withdrawing a new support after is has been initiated. Based on the number of crossover time points when units switched onto intervention, one-way crossover studies have been further classified into difference-in-differences (DD) (only one switching time point) and the stepped-wedge (SW) designs (more than one switching time point).

The one-way crossover design that first used in practice is DD, where the all units that switch to the new intervention do so at one time, i.e. there is only one crossover point. Consider, for example, for 100 facilities where portions of patients that underwent surgery that died in January, February and March were recorded. Then 50 of the facilities were given a new intervention to reduce surgery mortality at the end of March. Portions of patients that underwent surgery that died in April, May and June were recorded in all 100 facilities and the changes in mortality from the first three months to the next three months were compared between the 50 facilities that did and the 50 that did not receive the new intervention. While mixed models are not often formally used to compare the "difference in difference" this is arguably the best approach and thus is what is done for this dissertation.

However, phased intervention is often preferable due to practical constraints. For example, in the previous example perhaps only enough resources exist to switch 20 new facilities to the new interventions at one time point and/or ethically/politically all 100 facilities must eventually receive the intervention. Thus starting in the late 1990s a more general one-way crossover design known as SW, where the intervention is delivered to new groups of units switched at sequential time orders, i.e. there are multiple crossover points. The longitudinal before / after intervention changes are then compared within the group as a whole. In the example of the previous paragraph one SW design would be to start all 100 facilities out untreated in January, then switch 20 onto treatment each in February, March, April, May and June so that by June all 100 are treated. A mixed model would then be used to compare the "cascading longitudinal changes" as more facilities are switched onto the intervention.

As the SW is a newer design, there is less literature on SW studies than on DD studies. There is also a growing awareness of different types of SW designs including a spectrum between DD and SW. However, due to practical constraints that limit to simpler designs, this spectrum is probably more theoretical than actual.

The mixed models used to compare both DD and SW designs take on standard Generalized Least squares forms of $\underline{\hat{\beta}} = (X'V^{-1}X)^{-1}X'V^{-1}Y$ where (as described in each of the three chapters presented later) $\underline{\hat{\beta}}$ is the vector of coefficients one of which is for the intervention effect, *X* is the design matrix, *Y* is the outcome vector and *V* is the covariance matrix for the repeated measurements within the same unit. The variance of the estimates is then $(X'V^{-1}X)^{-1}\sigma^2$. The choice of *V* is very important as often-normative data on repeated measure correlation is unknown. To our knowledge, the simplest correlation of compound symmetry (CS) is used in study planning power estimation with more complicated forms not having been investigated. A more general tenable form of *V* is the Toeplitz structure that is diagonal-constant, i.e., the correlation ($\rho_{jj'}$) decreases as the distance of two time points *j* and *j'* (i.e. $|j - j'|$) increases.

One important aspect of designing either a DD or SW study is whether or not to randomize which units receive the intervention. While randomization of crossover is always preferred as a gold standard to minimize bias and improve efficiency, it is not always logistically feasible in practice. The effect of randomization on the design matrix *X* is to remove column(s) that would otherwise be needed to account for any baseline pre-intervention differences that can exist when randomization is not used.

This dissertation focuses on expanding tools for power / sample size estimation for planning studies of one-way crossover designs such as DD and SW based on the

variances for the intervention effect from $(X'V^{-1}X)^{-1}\sigma^2$. Very general methodologies for the analysis of randomized longitudinal repeated measure studies of all types using mixed models have been developed in recent years. However, there is both a need for simple estimation tools in practical power calculation targeted towards DD and SW one way cross-over designs and exploration of more approaches for power / sample size estimation for DD and SW designs when assumptions needed for the simple tools do not hold. In particular, current literature gives less guidance on power and sample size calculation for non-randomized versus randomized one-way crossover designs.

Again, power and sample size calculation is crucial to evaluate the effectiveness of the new intervention effect in longitudinal studies including one-way crossover designs. My dissertation has thus focused on developing and evaluating simple and robust tools for power / sample size estimation of various randomized and non-randomized Difference-in-Difference and Stepped-Wedge designs with continuous outcomes that conform to the central limit theorem. Some empirical exploration is undertaken into whether the simple correlation structure that is typically assumed (compound symmetry) holds in practice versus the more general Toeplitz and if the true structure is not compound symmetry, whether a compound symmetry or some other simple approximations are tenable. In addition to the power estimation tools and corresponding knowledge for one way cross over study planning gained, the orthogonalized decomposition of interventions effect versus other design parameters that was implemented in the research is a salient characteristic.

Three-Paper Dissertation Structure

My dissertation is presented using a three-paper format, consisting of three independent, yet congruent chapters.

Chapter 1: Power / Sample Size Estimation for Randomized Two-Arm Pre-Post Intervention Trials with Repeated Longitudinal Outcomes

The first chapter, develops a unified approach for continuous outcomes in randomized difference-in-differences (R-DD) studies by modeling the intervention effect in a Generalized Least Squares framework based on covariance of repeated measures. For compound symmetry (CS) correlation, the optimal pre-post allocation is presented with the advantage of closed form formulas. However, CS may not always hold in practice as was the case in four examples from nursing homes and HIV infected patients we used. For these cases a more general Toeplitz correlation is tenable, but would be harder to obtain in practice under study planning settings. The power for these Toeplitz settings are approximated using CS structures, however, even "conservative" CS approximations overestimated the power. Thus two alternative conservative approaches are presented: the simple 1-1 allocation and partial compound symmetry (PCS) based on mean summary statistics, but these often substantially underestimated power.

The formulas presented here for R-DD designs can be easily implemented with current programming languages, which may promote further recognition, application and development of these issues. These results may enable investigators working on R-DD to i) perform needed sample size / power estimation using CS covariance structure; ii) provide alternative lower bunds for power approximation; iii) decide the optimal allocation of pre- to post-intervention time points in planning a study.

Chapter 2: Power / Sample Size Estimation for Non-Randomized Difference-in-Differences Studies

The second chapter, extends the GLS power estimation framework to the non-randomized difference-in-differences (NR-DD) and quantitatively compares the advantage of optimal R-DD over optimal NR-DD. The penalty of non-randomization versus randomization on power / required sample size of DD designs is quantitatively calculated for compound symmetry correlation and empirically computed for Toeplitz correlation. While randomized designs have better precision, the advantage is minor for high within-unit correlation and/or with more baseline than follow-up measurements. For the more general Toeplitz correlation that is harder to obtain, the power approximated using CS and mean summary statistics approaches are presented, and compared to the Toeplitz using computer program with real examples from New Jersey nursing home and 1012 Bronx HIV study.

These results may enable investigators working on NR-DD to i) perform needed sample size / power estimation using CS covariance structure; ii) provide lower bounds for power approximation using mean summary statistics; iii) decide the optimal allocation in planning a study. In particular, the formulas presented here for non-randomized (as well as some new formulas for randomized) DD designs can be easily implemented with current programming languages, which again may promote further recognition, application and development of these issues.

Chapter 3: Non-Randomized and Randomized Stepped-Wedge Designs using an Orthogonalized Least Squares Framework

The third chapter, develops a unified approach for continuous outcomes in SW studies by modeling the intervention effect in an Orthogonalized Least Squares framework. General closed form formulas for variance of the intervention effect are derived for SW studies and optimal R-SW and NR-SW designs to maximize power are further investigated for the balanced SW studies (where the same number of units are switched to the new intervention and the number of time periods before a switch is constant). The impact of non-randomization on the baseline value of the outcome is modeled using both fixed and random effects. The penalty of non-randomization (versus randomization) is quantified in terms of power / required sample size for stepped wedge designs. While randomized designs have better precision (particularly if the within-unit repeated measures correlation is $\rho \leqslant 0.30$), the advantage is minor when $\rho \geqslant 0.50$ as was the case in the examples of health outcomes in nursing homes from New Jersey. However, for a non-randomized design with $\rho \leqslant 0.30$, the random effects (versus fixed effects) approach may considerably reduce the variance of estimated intervention effect, albeit use of random effects may increase bias in this setting. In terms of optimality properties, optimally designed non-randomized SW designs tend to reduce variance of intervention effect estimates to about 75% of the best achievable with traditionally used difference-in-differences studies.

These results may enable investigators to i) perform needed sample size / power estimation for SW studies and ii) decide the best study design to use. In particular, the formulas presented here for non-randomized (as well as some new formulas for randomized) stepped-wedge designs can be easily implemented with current

programming languages, which promote further recognition, application and development of these issues.

Chapter 1 Power / Sample Size Estimation for Randomized Two-Arm Pre-Post Intervention Trials with Repeated Longitudinal Outcomes

Abstract

Intervention effect on normal continuous longitudinal processes is often estimated in randomized two-arm longitudinal clinical trials that have b ≥ 1 pre- and k ≥ 1 post-intervention measures. Power / sample size estimation methods for such studies that can be used with available normative data is often limited. We derive simple Generalized Least Squares (GLS) power and sample size estimation formulas for randomized clinical trials (RCT) using the following correlation structures for the repeated measures: Toeplitz (TP), compound symmetry (CS) and partitioned compound symmetry (PCS) based on mean summary statistics. We then applied the GLS power estimation framework to examples from longitudinal nursing hospital and HIV outcomes where $b + k = 7$. *In these examples with* $b + k = 7$, *setting* $b = 1$ *produced optimal or close to optimal results to minimize variance of the estimated intervention effect (which maximizes power to detect an intervention difference), but having b=2 or b=3 often performed nearly as well by this metric. When there is uncertainty about exact Toeplitz structure, CS approaches approximate the "unknown" variance of the estimated intervention effect well when b=1 but can greatly underestimate this variance when b > 1. To avoid overestimation in power, we presented two approaches: PCS approximation based on mean summary statistics can serve as a conservative lower bound for GLS power calculation but greatly underestimated the power in two of our examples. An alternative lower bound approach with T=2 longitudinal measures (b=1 and k=1) obtained nearly*

as precise estimates of the intervention effect as did any design with T=b+k=7 measures where b>1 in these two cases.

1. Introduction

In clinical trials and other modern experiments, researchers often evaluate repeated measurements of continuous outcomes on each unit at systematic time points before and after intervention [1]. In our nomenclature, units could be facilities such as nursing homes or persons such as HIV infected patients. These repeated measure clinical trials are particularly done to compare long-term impact of a new policy / intervention versus the existing policy. When possible, randomization of units into each intervention arm is preferred to improve precision and minimize potential for bias; the randomized controlled trial (RCT) is considered a general gold standard [2]. Investigators first observe the longitudinal outcomes on each unit over *b* sequential time points pre-intervention. Then the units are randomly divided into two arms: one with intervention started and one without the intervention started and the outcomes are measured over *k* sequential time points (which are after receiving intervention for the intervention arm and still without intervention for the other arm). In medical research, there is increasing focus on power calculation and sample size determination for such longitudinal randomized clinical trials [3].

Standard power calculations have been developed for various settings over the years: Overall and Doyle [4] discussed sample size determination for repeated measures models with two groups. A key characteristic for such designs including longitudinal measures is that repeated measures from same units are (typically positively) correlated [5]. Ignoring the correlations using standard linear models may introduce bias and/or inefficiency into power estimates [6]. The general linear model (GLM) takes correlation into account for

the normal distribution approximation, and generalized least squares (GLS) is the statistical method of choice which has the best linear unbiased estimator (BLUE) [7]. Self and Mauritsen [8] developed unified tools for sample size and power estimations using GLM. Liang and Zegar [9] proposed a general variance formula that incorporated the impact of randomization using a constrained longitudinal data analysis (cLDA) model in which the baseline "pre-intervention" outcome values as well as post-baseline outcome "post-intervention" values are modeled as longitudinal dependent variables. The 'constraint' is that the baseline mean from different intervention arms are assumed equal due to randomization. While these approaches have established useful frameworks, they can be difficult to follow by researchers in power / sample size estimation as the generic design necessitates complex input structures including often unknown correlation structure for repeated measures.

Our goal is to develop a simple power estimation framework based on generalized least squares estimate in pre-post randomized intervention longitudinal clinical trials with two intervention arms where central limit theorem normality holds. The chapter is organized as follows: Section 2 presents the general linear model (GLM) for longitudinal data with pre-post repeated measurements. Section 3 develops a generalized least squares (GLS) framework for estimation of the intervention effect and incorporates the GLS variance estimate into power / sample size estimation. As the variance of the intervention effect depends on the correlation structure of repeated measure, Section 4 introduces three correlation structures: the simplest CS for compound symmetry; a more general Toeplitz correlation structure that must be implemented on computer for variance estimation; and Partial Compound Symmetry (PCS) as a lower bound for Toeplitz

structures when there is uncertainty. Section 5 derives optimal *b: k* allocation for randomized longitudinal clinical trials for a fixed number of total time points for the CS correlation and discusses extension to more general Toeplitz structures. In Section 6, we extract Toeplitz patterns with $T=b+k=7$ and generally $\rho > 0.5$ from important longitudinal health care outcomes of nursing homes, hospitals and HIV infected patients. We then compare the actual Toeplitz variances for estimated intervention effect from varying *b: k*, and analyze how close CS, and partitioned compound symmetry (PCS) approximations estimate the true variance for these settings. Section 7 summaries and discusses possible future work.

2. The General Linear Model

For randomized longitudinal studies with two intervention arms, researchers encounter repeated measures of a quantitative outcome at $b+k$ systematic time points with *b* being before and *k* being after randomization to the intervention arms. Let *h* denote the intervention arm with *h=0* for placebo and *h=1* for the new intervention. For each group there are n_h units (n_0 for the placebo and n_1 for the new intervention) and $j = \{-b, -(b-1), \ldots, -1, 1, 2, \ldots, k\}$ denotes the ordered times with $\{-b, -(b-1), \ldots, -1\}$ prior to and $\{1, 2, \ldots, k\}$ being after the intervention onset. The goal is to assess the impact of the new intervention (i.e. versus control) on pre-post change in a longitudinal continuous outcome *Y* where Y_{1ij} represents measure *j* from unit *i* in the new intervention arm and $Y_{0i',j'}$ represents measure *j'* from unit *i'* in the placebo arm.

For example, consider a trial with $n_0 = n_1 = n = 30$ hospitals in each arm, let *i* denote hospitals (as "units") where *i=1, …,* n_h. For the intervention arm (*h=1*), "units" are followed for *T*=7 years total with *b*=2 years (2001 to 2002) prior and *k*=5 (2003 to 2007) after the intervention implementation. Thus $Y_{1,3,-2}$ and $Y_{0,17,3}$ respectively denote the measure taken in 2001 (2 years prior to start of the intervention) in the 3^{rd} hospital of the intervention arm and 2005 (3 years after the start of the intervention) in the 17^{th} hospital of the placebo arm, respectively. We assume complete data with $T=b+k$ measures observed on each unit, which, in particular, is reasonable when the units are facilities that are required by regulations to keep records of the outcomes of interest.

Now Y_{hij} can be decomposed as:

$$Y_{hij} = \alpha + \beta_j + \theta Z_{hj} + \varepsilon^*_{ij} \quad (1)$$

The overall means (α) for two intervention arms are assumed to be equal at baseline, which is reasonable due to randomization. The intervention effect (θ) only delivers to the intervention arm (*h=1*) on the *k* post-intervention measurements with the corresponding indicator $Z_{hj} = I_{\{h=1, j>0\}}$. Any random unit (i.e. *i* level) effects are subsumed into the within-unit error term ε^*_{ij}, where $\varepsilon^*_{ij} \sim N(0, \sigma^2 V)$ with the correlation matrix *V* defined in (2). We assume an immediate "jump effect" of size θ after the intervention begins at time *j* =1, that remains unchanged at subsequent time points. Note that other functions such as linear intervention effect increase $j * \theta Z_{hj}$ for $j \geq 1$ or threshold followed by exponential decay $e^{-j} * \theta Z_{hj}$ for $j \geq 1$ are possible. However, there may be settings where an immediate "jump effect" that continues forward unchanged is appropriate, such

as when the intervention is a process change at a medical facility that can be implemented quickly; a drug that the body does not develop resistance or acclimation to, or an immediately successful behavioral intervention. Even if the intervention impact was not "immediate jump", it could be close to this.

3. The GLS Power Estimate Framework

3.1 GLS Variance Estimate

The matrix form of (1) can be written as: $Y = X\underline{\beta} + \varepsilon^*$, where $\varepsilon_{ij}^* \sim N(0, \sigma^2 V)$. Here X represents the design matrix and Y is a vector of outcomes. For (1) with the general parameter vector $\underline{\beta}=(\alpha, \beta_{-(b-1)}, \dots, \beta_{-1}, \beta_1, \dots, \beta_k, \theta)$, the corresponding X has columns $(I, J_{-(b-1)}, \dots, J_{-1}, J_1, \dots, J_k, Z)$, with $N*T$ rows per column. Z is a column/vector of intervention indicator with Z_{hj} coded (0, 1) as defined above; $J_{-(b-1)}, \dots, J_{-1}, J_1, \dots, J_k$ are columns corresponding to $b+k-1$ independent time coded variables as follows: for $j=(-(b-1), -(b-2), \dots, -1, 1, 2, \dots k)$, J_j= {-1 at time $-b$ (reference); 1 at time j; and 0 at all other times}. There is no column for J_{-b} as $\beta_{-b} = -\sum_{j=-(b-1)}^{k} \beta_j$ under the fixed effects constraint $\sum_{j=-b}^{k} \beta_j = 0$. Appendix 1 presents the full expansion of design matrix for randomized setting.

The covariance matrix V is made up with $(n_0 + n_1)$ times block T diagonal matrices $V_0's$ with all off-block diagonal matrix elements being 0. The most basic assumptions for the error term is that measures are independent between units, and within-unit correlation structure is invariant given two time points j and j', i.e., $\rho_{i,jj'} = \rho_{i',jj'}$ $(i \neq i', j \neq j')$ The

within-unit correlation structure ($\rho_{jj'}$) is often unknown in advance. Perhaps the correlation for any two time points would be monotonically non-increasing with $|j - j'|$, i.e., as two time points are further separated, they may become less correlated [10, 11]. We will make these types of restrictions later in the chapter.

$$V = \begin{pmatrix} V_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_0 \end{pmatrix}_{(n_0+n_1)T},$$

$$\text{where } V_0 = \begin{pmatrix} \rho_{11} & \rho_{12} & \rho_{13} & \cdots & \rho_{1,T-1} & \rho_{1,T} \\ \rho_{21} & \rho_{22} & \rho_{23} & \cdots & \rho_{2,T-1} & \rho_{2,T} \\ \rho_{31} & \rho_{32} & \rho_{33} & \cdots & \rho_{3,T-1} & \rho_{3,T} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{T-1,1} & \rho_{T-1,2} & \rho_{T-1,3} & \cdots & \rho_{T-1,T-1} & \rho_{T-1,T} \\ \rho_{T,1} & \rho_{T,2} & \rho_{T,3} & \cdots & \rho_{T,T-1} & \rho_{TT} \end{pmatrix}_T. \quad (2)$$

The GLS estimate for $\underline{\beta}$ is $\underline{\hat{\beta}}$ in (3), which is the best linear unbiased estimator (BLUE) for $\underline{\beta}$ and uniform minimum variance (UMVU) if Y_{hij} is normally distributed [7]. The GLS variance for $\underline{\hat{\beta}}$ is Λ in (4) being a square matrix of order T+1. The variance of $\hat{\theta}$ is

$$\underline{\hat{\beta}} = (X'V^{-1}X)^{-1}X'V^{-1}Y; \quad (3)$$

$$\Lambda = (X'V^{-1}X)^{-1}\sigma^2. \quad (4)$$

3.2 General Power Estimation Formula

We consider $H_0: \theta = 0$ versus $H_A: \theta = \pm\theta_A$. Where without loss of generality, $\delta = \frac{\theta_A}{\sigma}$ is a predefined clinically important effect size in terms of standard deviation, while α *and* β are Type I and Type II errors, respectively. For practical repeated measure

designs, the normal approximation of the non-central *t* distribution can be applied [12]. In specific, the two distributions are almost identical when degrees of freedom (DF) $\gamma > 30$ and we have the following equations of power $(1-\beta)$ using the notation from [1], in which $Var(\hat{\theta})$ as derived above in the GLS variance estimate in (4).

$$\theta_A = (z_{1-\frac{\alpha}{2}} + z_{1-\beta})\sqrt{Var(\hat{\theta})}. \quad (5)$$

For smaller sample sizes, it may be appropriate to approximate degrees of freedom (DF) (γ) in non-central *t* distribution for the mixture variance (for example, by Satterthwaite's [13], and Kenward-Roger's approximations [14]) and adjust (5) for this. However, the full details are beyond the scope of this chapter.

4. Covariance Matrices and Variance Formulas

As just noted, one main difficulty in parametric analysis of longitudinal data lies in specifying covariance structure [15, 16], i.e. estimating $\rho_{jj'}$ for $j \neq j'$. Typically, normative data from historical settings must be structured for application to future settings. In this section, we introduce several approximated correlation structures.

Table 1 describes three types of covariance matrices (V_0) for within-unit correlation structure considered here with$(b, k) = (3, 4)$. The most simple approximation is compound symmetry structure (V_{CS}) where correlations among repeated measures are assumed to be equal within the same unit; V_{CS} is often used in practice. But we believe Toeplitz structure (V_{TP}) where $\rho_{jj'} = \rho_{|j-j'|}$ is a reasonable estimate which may be

closer to true correlation structure, however, V_{TP}may be hard to estimate in practice. In that case, an approximation which we do not believe ever holds, but as we show later in the chapter gives conservative estimates for variance of the intervention effect estimates is partitioned compound symmetry structure (V_{PCS}) with four partitioned matrices divided by the time when intervention is delivered. Note that while Table 1 is constructed for $T = b + k = 7$, the subsequent formulations are generalizable to all pairs of (b, k). Further details now follow.

Table 1: Summary of three covariance matrices (*T*=7)

Structure	Example *(b, k)=(3, 4)*	# of Parameters
Compound Symmetry (CS)	$V_{\mathrm{CS}} = \begin{bmatrix} 1 & \rho & \rho & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & \rho & \rho & 1 \end{bmatrix}$	1
Toeplitz (TP)	$V_{\mathrm{TP}} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \rho_5 & \rho_6 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \rho_5 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_5 & \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_6 & \rho_5 & \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}$	*T-1*
Partitioned Compound Symmetry (PCS)	$V_{\mathrm{PCS}} = \begin{bmatrix} 1 & \rho_{\mathrm{pre}} & \rho_{\mathrm{pre}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} \\ \rho_{\mathrm{pre}} & 1 & \rho_{\mathrm{pre}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} \\ \rho_{\mathrm{pre}} & \rho_{\mathrm{pre}} & 1 & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} \\ \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & 1 & \rho_{\mathrm{post}} & \rho_{\mathrm{post}} & \rho_{\mathrm{post}} \\ \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{post}} & 1 & \rho_{\mathrm{post}} & \rho_{\mathrm{post}} \\ \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{post}} & \rho_{\mathrm{post}} & 1 & \rho_{\mathrm{post}} \\ \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{cross}} & \rho_{\mathrm{post}} & \rho_{\mathrm{post}} & \rho_{\mathrm{post}} & 1 \end{bmatrix}$	3

4.1. Compound Symmetry and a Simple Formula for Variance

Compound Symmetry (CS, also denoted sphericity or equi-correlation), is a commonly used covariance [15, 16] either as the true structure or as an approximation. It is considered reasonable to expect that the largest covariance component would be a main effect for the unit *i* with much smaller within-unit temporal changes from ε_{ij}^{*}. This results in $\rho_{|j-j'|} \approx \rho$ with perhaps a small but ignorable decrease as $|j - j'|$ increases. While surprisingly little empirical research has been done to confirm this structure holds given how often it is used, one study finds that CS was a reasonable simplification in quantitative planning of repeated measures trials for the examples used in that paper [1].

However, even if the true correlation structure is believed to be non-CS Toeplitz, a compound symmetry approximation might still be used for power / sample size estimation either because as we will see later CS formulas are easier to implement or there is uncertainty about the values of the Toeplitz parameters $\{\rho_1, \rho_2, \dots, \rho_{b+k-1}\}$. It this is done then is may be reasonable to use the weighted average of the observed or inferred V_{TP} to estimate the CS ρ, used in the approximation i.e. $\rho = \rho_{avg} = \frac{(b+k-1)\rho_1+(b+k-2)\rho_2+\cdots+\rho_{b+k-1}}{\sum_{i=1}^{b+k-1} i}$. However, as we will show later using $\rho_{min} = min\{\rho_1, \rho_2, \dots, \rho_{b+k-1}\}$ as the in the CS approximation will give a more conservative estimate. We explore later how good CS approximations using ρ_{avg} and ρ_{min} perform in real world settings in Section 7.1.

Under the assumption of CS, we derive a closed form GLS formula for $Var(\hat{\theta}_{CS})$ follows. The GLS estimator of $\underline{\beta}$ is therefore $\underline{\hat{\beta}} = \left(X' V^{-1} X\right)^{-1} X' V^{-1} Y$ and has variance $\Lambda = \left(X' V^{-1} X\right)^{-1} \sigma^2$ where Λ is a square matrix of order *b*+*k*+1. $Var(\hat{\theta}_{CS})$ is

the last diagonal element of Λ. Using the inverse formula for portioned matrix, we calculate for the following GLS variance estimate of intervention effect in Appendix 2:

$$Var(\hat{\theta}_{CS}) = (\frac{1}{n_0} + \frac{1}{n_1}) \frac{[1+(b+k-1)\rho](1-\rho)}{k[1+(b-1)\rho]} \sigma^2 \quad (6)$$

Therefore, $Var(\hat{\theta}_{CS})$ in (6) is a simple formula for GLS variance estimate that enables derivation of optimal properties in Section 5.

4.2. Toeplitz a General Structure with Variance that can be obtained by Computer

However, if CS does not hold it still seems that an essentially necessary assumption for estimabillity is Toeplitz (TP, also known as a diagonal-constant matrix), with correlations a function of the difference in *j* and *j'* and independent of chronological time namely that $\rho_{jj'} \equiv \rho_{|j-j'|}$(with |*j*-*j'*| = 1, 2, …., *T*-1) as otherwise it seems impossible to have any stationary estimability for $\rho_{jj'}$ from normative data [15, 16]. While, under this assumption, one would like to use V_{TP} in practice for variance estimation and power planning, typically, normative data to estimate $\{\rho_1, \rho_2, \dots, \rho_{b+k-1}\}$are not available and computation is difficult without sophisticated software to implement the estimation. The Restricted maximum likelihood (REML) is recommended for estimation of $\{\rho_1, \rho_2, \dots, \rho_{b+k-1}\}$ from normative data when interest lies in estimating accurate variance components of mixed models [17]. In fact, REML estimation is included as a default option in many current model-fitting software packages (e.g., Proc Mixed in SAS). There is no simple form for the variance of the estimated intervention effect under V_{TP}, rather $Var(\hat{\theta}_{TP})$ must be obtained by computer incorporating V_{TP} into (4).

4.3. Partitioned Compound Symmetry Gives a Simpler Upper Bound Variance

Partitioned compound symmetry (PCS) using mean summary (MS) statistics [1] is now presented to be used as an approximation to Toepltiz (or for that matter other covariance structures), since as we show later it is able to obtain an upper bound for variance or lower bound for power. PCS effectively assumes repeated measures from both pre- and post-partitioned block are equicorrelated within block partitions as follows; the *b* pre-intervention time points have equal correlations to each other denoted as ρ_{pre}, the *k* post-intervention time points have equal correlations to each other denoted $\rho_{\text{post}}$and the cross correlations between each of the *b* pre- and *k* post-intervention time points is the same denoted as ρ_{cross}. The common correlations of the partitions are calculated from "mean summary" statistics of the actual Toeplitz (or other) correlations as described below. If the correlation structures are Toeplitz as described above, then the mean summary statistics for $\rho_{\text{pre}} = \frac{(b-1)\rho_1+(b-2)\rho_2+\cdots+\rho_{b-1}}{(b-1)+(b-2)+\cdots+1}$ is the averaged correlation among the *b* pre-intervention time points, $\rho_{\text{post}} = \frac{(k-1)\rho_1+(k-2)\rho_2+\cdots+\rho_{k-1}}{(k-1)+(k-2)+\cdots+1}$ is the averaged correlation among the post-intervention time points and $\rho_{\text{cross}} = \frac{\sum_{i=1}^{k}(\rho_i+\rho_{i+1}+\ldots+\rho_{i+b-1})}{bk} = \frac{\sum_{i=1}^{k}\sum_{j=0}^{b-1}\rho_{i+j}}{bk}$ is the averaged correlation between the pre- and post-intervention measurements.

Frison & Pocock [1] proposed to use analogous mean summary statistics ($\bar{Y}_{hi.}^{post} = \frac{1}{k}\sum_{j=1}^{k} y_{hij}$ and $\overline{Y}_{hi.}^{pre} = \frac{1}{b}\sum_{j=-b}^{-1} y_{hij}$) to analyze repeated measurements in randomized trials with two intervention arms based on an Analysis of Covariance (ANCOVA) approach being used to model the data. Then the overall mean for the post-intervention is

$\bar{\mu}_{h..}^{post} = \frac{1}{n_h}\sum_{i=1}^{n_h} \bar{Y}_{hi.}^{post}$; and the overall mean for the pre-intervention is $\bar{\mu}_{...}^{pre} = \frac{1}{n_0+n_1}(\sum_{i=1}^{n_0} \bar{Y}_{0i.}^{pre} + \sum_{i=1}^{n_1} \bar{Y}_{1i.}^{pre})$, which is the same for both intervention arms due to randomization.

The idea of ANCOVA is to model the pre-intervention mean for each unit as a covariate in a linear model for intervention arm comparison of the post-intervention mean.

$$\bar{Y}_{hi.}^{post} = \bar{\mu}_{h..}^{post} + \theta(\bar{Y}_{hi.}^{pre} - \bar{\mu}_{...}^{pre}) + \varepsilon_i \quad (7)$$

Using the above PCS parameters, the variance estimate of $\hat{\theta}$ is obtained by least squares from ANCOVA [1, 18]:

$$\mathrm{Var}(\hat{\theta}_{PCS}) = \left(\frac{1}{n_0} + \frac{1}{n_1}\right)\left[\frac{1+(k-1)\rho_{\mathrm{post}}}{k} - \frac{b\rho_{cross}^2}{1+(b-1)\rho_{\mathrm{pre}}}\right]\sigma^2 \quad (8)$$

The ANCOVA estimate in (7) based on mean summary statistics using PCS is unbiased for θ, and the GLS estimate in (1) is a best linear unbiased estimator (BLUE) [7]. We can conclude from the Gauss-Markov theorem that the GLS variance estimate $Var(\hat{\theta})$ based on $V_0 = V_{TP}$in (4) is no greater than the ANCOVA variance estimate $\mathrm{Var}(\hat{\theta}_{PCS})$ in (8). Therefore, ANCOVA approach based on mean summary statistics can serve as an upper bound for the GLS variance of a given Toeplitz correlation (and a lower bound for power). It should be noted that under CS approximation formula (8) numerically gives the same results as formula (6) meaning the ANCOVA mean summary approach of Frison & Pocock and GLS produce the same $Var(\hat{\theta})$ estimate when CS holds.

5. Optimal Allocation of Pre-Post Intervention Measurements when Compound Symmetry is Easily Derived

We now present one important property of the longitudinal clinical trial we have been studying that is easily derived for CS covariance structure as $Var(\hat{\theta}_{CS})$has a simple closed form GLS variance formula (6). A repeated measures design may have a constrained total number of longitudinal times *T* (*T*=*b* + *k*) because of the budget and/or time constraints. Finding the optimal allocation of *T* into *b* and *k* to maximize power or minimize the sample size needed to obtain a given power would be important. From (6), for CS structure with constrained *T* given *ρ,* the optimal *b* with the local minimization of variance is:

$$b^* = round(\frac{T+1}{2} - \frac{1}{2\rho}). \quad (9)$$

Frison and Pocock [1] obtained this result using CS in their ANCOVA model. In general, the optimal *b* with 'minimum variance' becomes larger as the correlation coefficient ρ increases for a constrained total time points (*T*) because the pre-intervention measurements are of greater use.

For example, suppose *T*=7 and $\rho = 0.4$ for a randomized trial, then we calculate the optimal pre-intervention measurements $b^* = round\left(\frac{7+1}{2} - \frac{1}{2(0.4)}\right) = round(2.75) = 3.$ If *T*=6 and $\rho = 0.5$, then $b^* = round\left(\frac{6+1}{2} - \frac{1}{2(0.5)}\right) = 2.5 = 2$ or 3. Note $Y = round(X)$ rounds each element of *X* to the nearest integer. If an element is exactly between two integers, then *Y* can be either of the two integers.

The following contour lines in Figure 1 depict the distribution of optimal choice of pre-intervention measurements *b* for any given function of total time points *T* and

correlation ρ. Note: for any region between two contour lines, the optimal choice of pre-intervention measurements is determined by the lower contour line. If it lies exactly on a contour line, then the optimal is determined by that particular value. Now look up (T, ρ) = (7, 0.4) and (7, 0.6) in the contour plot below: (7, 0.4) lies exactly on the contour line labeled with 3, so the optimal choice of b is 3; (7, 0.6) lies between two contour lines whose values are 3 and 4, so the optimal is the lower value 3.

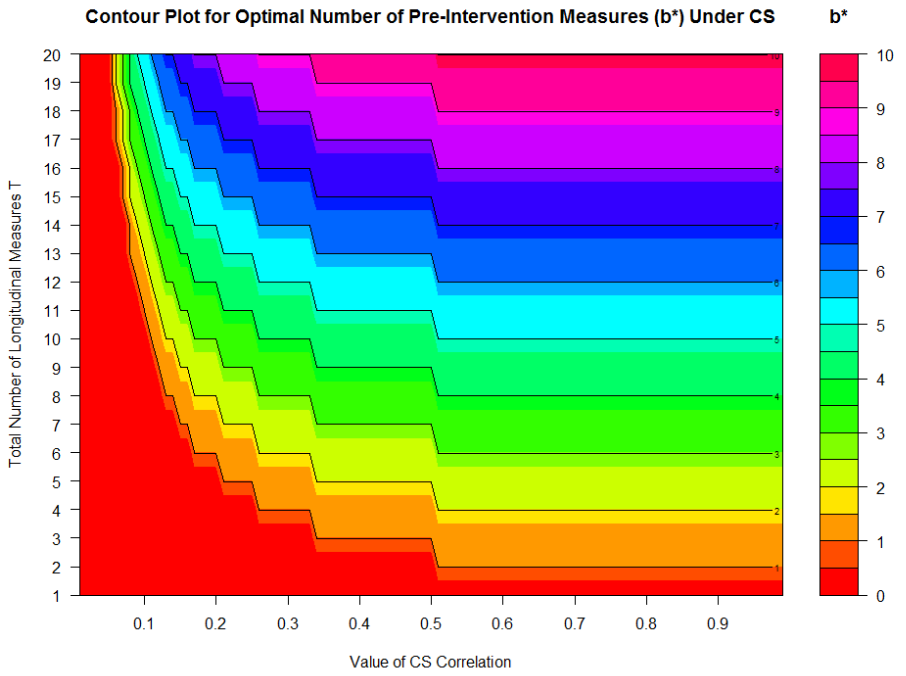

Figure 1: Optimal allocation for CS using contour plot

6. Empirical Examples, Properties with Toeplitz Correlation

6.1. Examples having Toeplitz Correlation

We now turn to the more general setting where the correlation structure is non-CS Toeplitz and begin with some real examples of where this happens. Our first two

examples are from data collected on 365 New Jersey nursing homes being monitored every three months from the second quarter of 2011 to the fourth quarter of 2012 (seven quarters total) in the Nursing Home Compare [19] for proportions of: 1) long stay residents with long term need for help with activities of daily living (LS_ADL); and 2) short term stay patients that reported moderate to severe Pain (SS_Pain). Higher levels of both LS_ADL and SS_Pain are undesirable and targeted for improvement at a facility level. The "unit" for these examples is the facility with the repeated measure being quarterly facility values. Thus, for example, in a future study, it is conceivable that all 365 New Jersey nursing homes could be followed for b baseline time points to obtain LS_ADL and/or SS_Pain proportions and then around 50% be moved to a facility intervention to improve one or both of these with k post-intervention measures obtained from both groups for comparison of change.

The next two examples are obtained from 1012 Bronx HIV infected women [18] who had complete data for their first seven semiannual visits for CD4 counts and CESD Depression scores [19]. Higher CD4 and lower CESD are desired and have been previously targeted for interventions. The repeated measures for these examples are from semiannual visits of patients. It is conceivable that in a future study these patients could be followed for b baseline visits to obtain CD4 and/or CESD scores and then around 50% be put on an intervention to improve one or both of these outcomes with k post-intervention measures obtained from both groups for comparison of change.

Note we chose $T=b+k=7$ for these examples which is reasonable not only for our examples but for trials conducted over 2-4 years with repeated measures at 3-6 months interval.

Table 2 and Figure 2 summarize the empirical Toeplitz correlation structures for the four outcomes described above estimated using the mixed procedure in SAS from the normative data. Visually, Figure 2 illustrates a range from starting correlations at ρ_1 of ~0.60 to ~0.87 and in slight to steep generally monotonic linear declines going down to ρ_6 ranging from ~0.34 to ~0.55.

Table 2: Toeplitz correlation structures (V_{TP}) from four examples

Time	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6
Among Quarterly Evaluations of 365 New Jersey Nursing Homes						
LS_ADL	0.59	0.47	0.41	0.39	0.40	0.34
SS_Pain	0.87	0.76	0.69	0.66	0.63	0.54
Among Semiannual Visits of 1012 HIV-Infected Bronx-WIHS Patients						
CD4	0.84	0.74	0.65	0.57	0.46	0.47
CESD	0.64	0.59	0.54	0.53	0.52	0.55

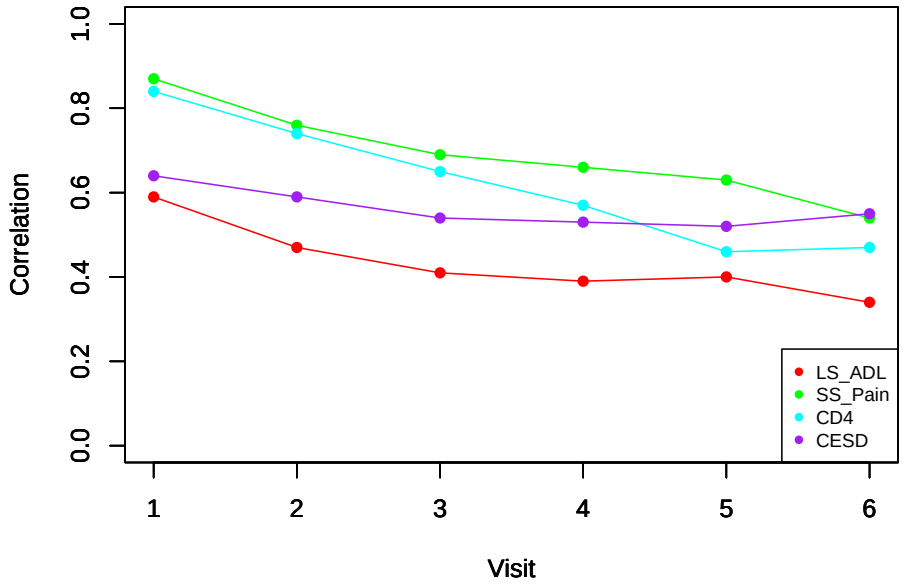

Figure 2: Visualization of Toeplitz correlation structures from real examples $T=b+k=7$

For example with LS_ADL, the within facility correlation between times (i.e. quarters) j and $j+1$ is $\rho_1 = 0.59$. But this drops to $\rho_6 = 0.34$ for between times j and $j+6$. More specifically, the correlations of CD4 and SS_Pain start higher at $\rho_1 \approx 0.85$ than do those of LS_ADL and CESD at $\rho_1 \approx 0.60$; the decline $(\rho_1 - \rho_6)$ for SS_Pain, LS_ADL and CD4 (≈0.30) is greater than that for CESD (≈0.10). But qualitatively we argue that none of these correlation structures are close to compound symmetry. Thus Section 6.2 presents power estimates and optimality properties for these four examples obtained by computer using (4) and (5) incorporating the V_{TP} structures in Table 2 and Figure 2.

6.2. Empirical Optimal *b: k* allocation for the Four Toeplitz Examples

In order to discover the optimal allocation with greatest power, we compared $\mathrm{Var}(\hat{\theta}_{TP})$ from $(X' V^{-1} X)^{-1}\sigma^2$ using the true Toeplitz correlation structure over all possible *b*: *k* allocations for each of the four examples assuming without loss of generality that the variance of each outcome was $\sigma = 1$ and $n_0 = n_1 = 30$. Table 3 illustrates the computed variance from (4) using the appropriate V_{TP} as shown in Table 2 and Figure 2 over all possible allocations of (*b, k*) for *T*= 7. The optimal *b*: *k* allocation for each example occurs at the minimum variance as indicated in bold.

Table 3: Toeplitz variances from four examples ($n_0 = n_1 = 30, \sigma = 1, T = b + k = 7$)

(*b*, *k*)	(0, 7)	(1, 6)	(2, 5)	(3, 4)	(4, 3)	(5, 2)	(6, 1)
LS_ADL	0.0360	0.0240	0.0231	0.0241	0.0263	0.0302	0.0400
SS_Pain	0.0501	0.0140	0.0141	0.0142	0.0147	0.0153	0.0149
CD4	0.0467	0.0149	0.0175	0.0175	0.0178	0.0181	0.0171
CESD	0.0424	0.0216	0.0192	0.0192	0.0204	0.0231	0.0326

From Table 3, we can see that *b*=0 performs particularly poorly for all examples and *b*=2 is the optimal choice for LS_ADL and CESD; *b*=1 is the optimal choice for SS_Pain and CD4. However, in many settings, minimizing *b* to maximize *k* and hence the ability to observe the long term intervention effects may be desired. For ethical considerations, earlier intervention is also preferred in research on individuals at clinical high risk. To that end, *b*=1 performed well in Table 3 in that: i) *b*=1 was much better than *b*=0, which indicates at least one baseline measurement is required. ii) *b*>1 was at best minimally

better than b=1, which implies multiple pre-intervention measurements will not help much in variance estimation.

While more comprehensive analyses for other values of T and V_{TP} is beyond the score of this chapter, we believe that: i) V_{TP} presented here are similar to those seen elsewhere [1] and thus likely to hold in many settings ii) $T \approx 7$ may be reasonable for many settings so this observation may be widely applicable.

7. Power Estimation using Simple Approximations to Toeplitz Structure

If the actual structure of V_{TP} can be identified and the needed software is available, it is ideal to use it in (4) for variance / power calculation. However, in practice, investigators often have limited access to normative data from which to obtain V_{TP} or access/skills to use needed software to generate $\mathrm{Var}\left(\hat{\theta}_{TP}\right)$ from (4) in the limited time that is typically available to apply for study funding. Furthermore, power/sample size estimates using V_{TP} could have unknown robustness properties against misspecification on $\{\rho_1, \ldots, \rho_{T-1}\}$. It thus may be of value to develop power estimate tools using simple approximations to the actual Toeplitz structure in longitudinal clinical trials in Section 7.1 or otherwise to obtain conservative (upper bound) variance estimates as described in Section 7.2.

7.1. Compound Symmetry Approximations to Toeplitz Variance

The compound symmetry structure with a common ρ is probably the simplest approximation if obtaining V_{TP} is impractical or impossible as (5) and (6) can be used to estimate power. However, which value of "approximated ρ" to use in (6) is not clear.

As discussed in Section 4.1, one reasonable approach is to estimate (i.e. what is believed to be) the equi-correlation ρ with the weighted average of all (i.e. estimated) intra-unit correlations among the *T* time points, in the substituted V_{CS} where $\rho = \rho_{avg} = \frac{(b+k-1)\rho_1+(b+k-2)\rho_2+\cdots+\rho_{b+k-1}}{\sum_{i=1}^{b+k-1} i}$. For example, for LS_ADL with (*b*, *k*) = (3, 4) using the observed correlations from Table 1, $\rho_{avg} = \frac{6\rho_1+5\rho_2+4\rho_3+3\rho_4+2\rho_5+\rho_6}{21} = 0.47$. The second approach is to let $\rho = \rho_{min}$ as the common correlation in the substituted $V_{CS}$where $\rho_{min}$is the minimum correlation in V_{TP}. The second approach is more conservative in power estimation than the first in that it obtains larger variances since the GLS-CS variance in (6) increases as ρ decreases. This ρ_{min}, typically would be $\rho_{1,b+k}$ if the correlations are decreasing with |*j*-*j'*|. For example, for LS_ADL with all values of (*b*, *k*), $\rho_{min} = \rho_6$ = 0.34. The first two columns of Table 4 give $\rho_{avg}$and ρ_{min} for *T*=*b*+*k*=7 based on the V_{TP} in all four examples of Figure 2.

Table 4: Calculated parameters for CS approximation and conservative approximations from the Toeplitz correlation structures in Table 2 for *(b, k)*=(3, 4) ($n_0 = n_1 = 30, \sigma = 1$)

Outcome in Table 2	Heuristics Approximations		Conservative Approximations			
	CS Parameters[1]		(*b*, *k*) = (1, 1)	PCS Parameters		
	ρ_{avg}	ρ_{min}	ρ_1	ρ_{pre}	ρ_{post}	ρ_{cross}
Among Quarterly Evaluations of 365 New Jersey Nursing Homes						
LS_ADL	0.47	0.34	0.59	0.55	0.52	0.42
SS_Pain	0.74	0.54	0.87	0.83	0.80	0.69
Among Semiannual Visits of 1012 HIV-Infected Bronx-WIHS Patients						

CD4	0.69	0.46	0.84	0.81	0.78	0.61
CESD	0.58	0.52	0.65	0.62	0.61	0.55

1. The CS Approximation parameters $\rho_{avg}$and ρ_{min} are invariant to (*b*, *k*)

While clearly one would prefer to use these V_{TP} directly in (4) id they and software for (4) were available it still is useful to see how close CS approximations from $\rho_{avg}$and ρ_{min} ain (6) are as these are much easier to obtain. Thus Figure 3 shows the results for the above four examples including the actual $\mathrm{Var}(\hat{\theta}_{TP})$ in the randomized design for all possible allocations of (*b*, *k*) from Table 3 compared to those produced by CS approximations in (6) with $\rho = \rho_{avg}$ and $\rho = \rho_{min}$ as shown in Table 4. For *b*=1 and *b*=6, CS using $\rho = \rho_{avg}$ performed well for all four examples never being anticonservative and being almost exact to CESD and LS_ADL. By contrast, CS with $\rho = \rho_{min}$ greatly overestimated the variances when *b*=1 and 6 for CESD and LS_ADL. However, for *b* ranging from 2 to 5, $\rho = \rho_{avg}$ often greatly underestimated the variance and for CESD and LS_ADL even using $\rho = \rho_{min}$ was anticonservative.

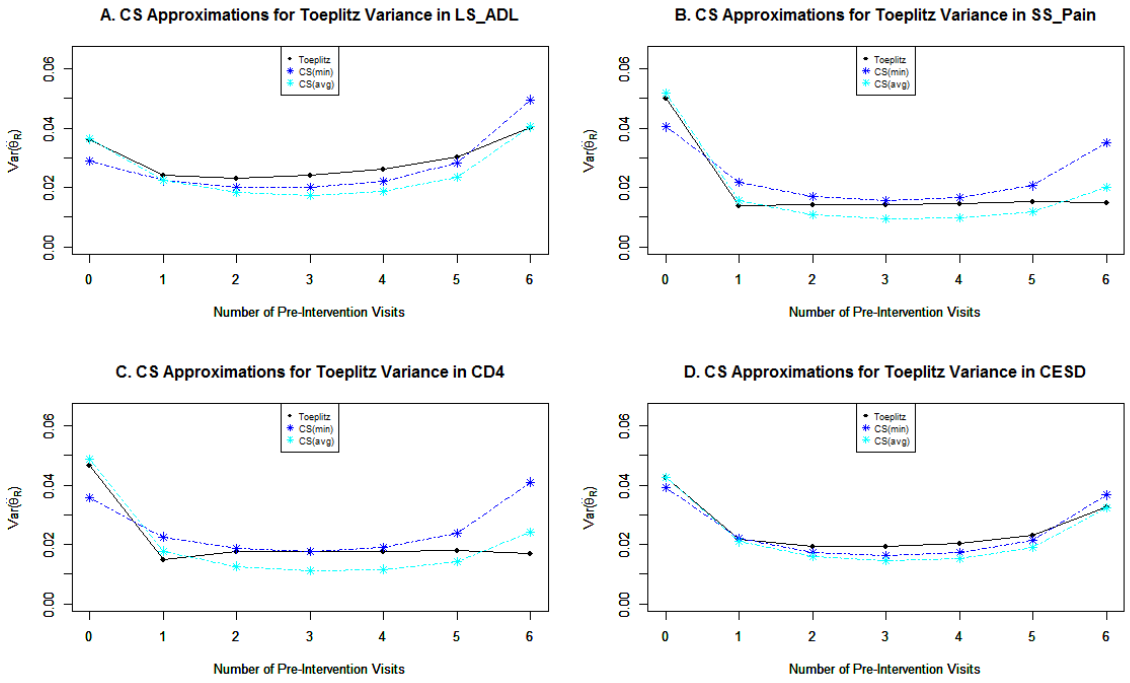

Figure 3: Variance approximations using CS compared to Toeplitz over all *b: k* allocations ($n_0 = n_1 = 30, \sigma = 1, T = b + k = 7$)

7.2. Two Conservative Approximations to Toeplitz Variance

The failure of even $\rho = \rho_{min}$ in a CS approximation to consistently produce conservative estimates of $\text{Var}(\hat{\theta}_{TP})$ motivates the need to have simple approaches to not underestimate $\text{Var}(\hat{\theta}_{TP})$ if direct calculation is not feasible.

If $b \geq 1$, perhaps the simplest conservative estimate for general T is to reduce the study to T=2 and (b, k) = (1, 1) with only one off-diagonal correlation, the correlation structure is by default V_{CS} with $\rho = \rho_1$ in (6). Clearly restricting the study to T=2 measures with one pre- and one post-intervention with this ρ should yield smaller variance than using all T time points with more than one pre-intervention measure ($T \geq 2$ and $b \geq 1$); and one would also expect that for all $T \geq 2$ the maximum off-diagonal

correlation (i.e. $\rho_{max} = \rho_1$). Note that if b=0 then this is not necessarily a conservative approximation.

Restricting to only 2 of T (i.e. T=7) measures (i.e. when $b \geq 1$) at first glance seems overly conservative which motivates need for another lower bound. To that end as described in Section 4.3, Frison & Pocock [1] proposed to use mean summary statistics ($\bar{Y}_{hi.}^{post} = \frac{1}{k}\sum_{j=1}^{k} y_{hij}$ and $\bar{Y}_{hi.}^{pre} = \frac{1}{b}\sum_{j=-b}^{-1} y_{hij}$) to analyze repeated measurements in randomized trials with two intervention arms. As we discussed in Section 4.3, $\mathrm{Var}\left(\hat{\theta}_{PCS}\right)$ in (8) can serve as an upper bound for $\mathrm{Var}\left(\hat{\theta}_{TP}\right)$ in (6).

Table 4, thus also describes four our 4 examples, implementation of the simple (b, k) = (1, 1) approximation when $b \geq 1$and the PCS approximation using mean summary statistics to derive upper bounds for $\mathrm{Var}\left(\hat{\theta}_{TP}\right)$ (and thus a lower bound for power). The third column of Table 4 presents ρ_1 for the (b, k) = (1, 1) conservative approximation. The last 3 columns of Table 4 present the values for PCS parameters with (b, k) = (3, 4): ρ_{pre}, ρ_{post} and

ρ_{cross} for LS_ADL, SS_Pain, CD4 and CESD outcomes described earlier with based on the longitudinal correlations shown in Table 1.

Thus, note that for LS_ADL, incorporating $\rho_{max} = \rho_1 = 0.59$ with (b, k) = (1, 1) into (6) gives $\left(\frac{1}{n_0} + \frac{1}{n_1}\right) * \frac{[1+(2-1)*0.59](1-0.59)}{1*[1+(1-1)*0.59]} = 0.65 * \left(\frac{1}{n_0} + \frac{1}{n_1}\right) = 0.043$; while incorporating$\rho_{pre} = 0.55$, $\rho_{post} = 0.52$ and ρ_{cross}=0.42 with (b, k) = (3, 4) into (8) gives $\left(\frac{1}{n_0} + \frac{1}{n_1}\right)\left[\frac{1+(4-1)*0.52}{4} - \frac{3*0.42^2}{1+(3-1)*0.55}\right] = 0.388 * \left(\frac{1}{n_0} + \frac{1}{n_1}\right) = 0.026$. Both numbers can serve as upper bound to the variance based on the actual Toeplitz structure in Table 3

where $Var(\hat{\theta}_{TP}) = 0.024$ for a randomized study of n_0 units in placebo and n_1 units in intervention with (b, k) = (3, 4) for LS_ADL.

Figure 4 presents the results for the above four examples in the randomized design including the actual $\mathrm{Var}(\hat{\theta}_{TP})$ compared to conservative estimates produced by i) simple approximation with (b, k) = (1, 1) and $\rho_{max} = \rho_1$in (6) for when $b \geq 1$and ii) PCS approximations (with ρ_{pre}, ρ_{post} and ρ_{cross}) in (8). For SS_Pain and CD4 where the within-unit correlation ρ_1 was very high at ~0.85 followed by rapid drop-off going to ρ_2 and beyond, the PCS approximation greatly overestimated the true variance as shown in Figure 4-B and Figure 4-C. However, surprisingly for all values of $b \geq 1$, the variance of the intervention effect was only barely smaller than from restricting to two time points (b, k) = (1, 1) with $\rho_{max} = \rho_1$. But for LS_ADL and CESD where the Toeplitz correlations were much closer to CS with much smaller drop-off from ρ_1 to ρ_6, the PCS upper bound was very close to the true Toeplitz variance for all values of b, while restricting to two time points with (b, k) = (1, 1) using ρ_1 greatly overestimated the variance form the Toeplitz correlation.

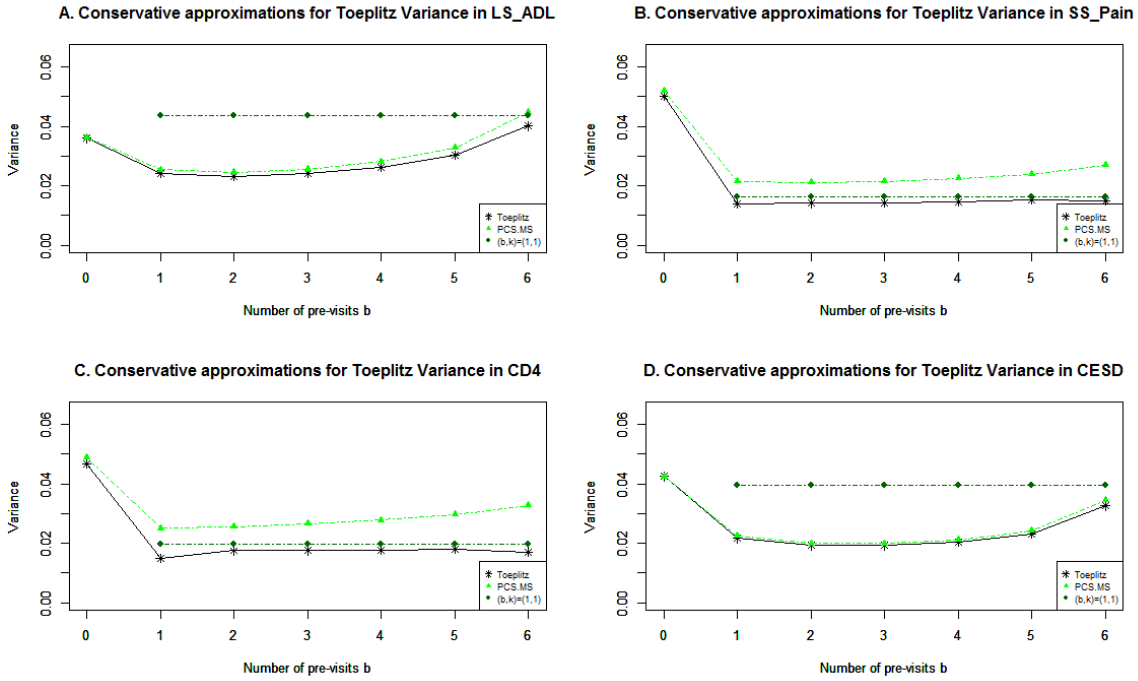

Figure 4: Conservative approximations for randomized designs over all *b: k* allocations ($\boldsymbol{n_0 = n_1 = 30, \sigma = 1, T = b + k = 7}$)

8. Concluding Remarks

Methodologies for the analysis of repeated measures have been developed in recent years based on general linear models. However, there is a need for simple estimation tools in practical power calculation. One aim of this chapter was to develop "usable" power and sample size estimation tools for researchers working on randomized before and after two-arm intervention designs with repeated longitudinal measurements. We developed tools for variance of the estimated intervention effect in randomized studies based on a Generalized Least Squares (GLS) framework. We first used compound symmetry structure for the within-unit correlation and derived a simple GLS formula for $Var(\hat{\theta}_{CS})$ in (6), which is easily calculated and implemented for power / sample size

estimation. With the advantage of closed form GLS formula based on CS, we explored the optimal value of *b* to minimize $Var(\hat{\theta}_{CS})$ for constrained *T*. In general, this optimal *b* was larger as the correlation coefficient ρ increased for a constrained *T* because the pre-intervention measurements became of greater use.

However, in our real data examples using outcomes from long term care facilities and HIV patients, the correlation structures were (sometimes very) different from compound symmetry suggesting further investigation on power estimation with the more general Toeplitz correlation was needed. We thus computed and analyzed $Var(\hat{\theta}_{TP})$with the empirical Toeplitz correlation structures of these examples. As closed form formulas for variance of estimated intervention effect are not directly available for Toeplitz correlations we numerically evaluated the properties of $Var(\hat{\theta}_{TP})$using computer software in (4). Our four examples suggest that *b*=1 gave close to optimal results although larger values of *b* up to 4 were often better. In addition, having at least one baseline pre-intervention measure is important as *b*=0 always did much worse.

In practice, investigators often neither have precise normative data on the Toeplitz variance parameters $\{\rho_1, \ldots, \rho_{T-1}\}$ of repeated measures nor the software/expertise to derive or implement V_{TP}. We thus investigated power approximation approaches using closed form formula variances in (6) for CS approximations to V_{TP} when $T=b+k=7$. The CS approximations using $\rho = \rho_{avg}$ sometimes substantially underestimated the true $Var(\hat{\theta}_{TP})$ of the estimated intervention effect and thus overestimated power. Moreover, even the more conservative CS approximations using $\rho = \rho_{min}$ sometimes resulted in substantial power overestimation.

We thus looked for approaches to derive upper bounds to $Var(\hat{\theta}_{TP})$ for our four examples and had some surprising results. The PCS approximation based on mean summary statistics provided an alternative upper bound for $Var(\hat{\theta}_{TP})$ (equivalently lower bound for power). Note that for the two examples with high ρ_1 that dropped off rapidly (SS_Pain and CD4), the PCS approximation greatly overestimated $Var(\hat{\theta}_{TP})$. Furthermore for these examples using (b, k) = (1, 1) produced only slightly lower variances than $Var(\hat{\theta}_{TP})$ when $b \geq 1$, T=7, meaning that having only T=2 time points may be sufficient and much less costly than T=7 for these settings. However, for the other two examples where ρ_1 was smaller and the drop-off between ρ_1 and $\rho_6$was smaller (LS_ADL and CESD) the PCS approximation only slightly overestimated $Var(\hat{\theta}_{TP})$ and restriction to T = 2 total time points with (b, k) = (1, 1), resulted in a large increase in variance. Thus it does not appear to be a simple way to obtain simple upper bounds for $Var(\hat{\theta}_{TP})$ that works in all settings. Thus, it does not appear to be a simple way to derive upper bounds for $Var(\hat{\theta}_{TP})$ that works in all settings.

There are some limitations in our work. We assumed an immediate one-time jump effect of the intervention, but in some settings the effect may be linear cumulative or some other pattern. The illustrative examples we used are limited with a fixed total time points ($T = 7$). While more comprehensive analyses for other values of T in general and other correlation structures is beyond the scope of this chapter, we believe that the correlation structures in the four examples presented here are likely generalizable and that $T \approx 7$ may be reasonable for many settings with repeated measures taken at 3-6 month intervals. Although we assumed static covariance (a minimum requisite to use historical data for correlation estimation), covariance could change over time from uncontrollable

mechanisms in practice. Relaxation of the above assumptions may likely lead to complicated settings that perhaps can only be addressed with simulation.

In conclusion, this chapter developed a power estimation framework based on covariance approximations and investigated optimal allocation of number of pre- (b) and post- (k) intervention measurements for constrained T for randomized longitudinal difference in differences studies. Under the assumption of compound symmetry correlation, we derived simple formulas for $Var(\hat{\theta}_{CS})$ in (6). However, CS may not always hold in the real world as shown in our examples. Our illustrative examples using observed Toeplitz correlations did not always empirically support similar properties for optimal allocation of b: k as were derived for CS using closed form formulas. When the exact Toeplitz structure is unknown or hard to apply, we presented conservative lower bounds for power based on simple (b, k) = (1, 1) approximation and PCS approximation using mean summary statistics. Thus, while it may be difficult for many investigators both to obtain normative data for Toeplitz correlation structure and to compute variances of intervention effect estimates based on Toeplitz variances, our efforts to identify simple and conservative approximations had mixed success.

References

1. Frison L, Pocock SJ. Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine* 1992; 11, 1685–1704.

2. Meldrum ML. A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. *Hematol Oncol Clin North Am* 2000; 14 (4): 745–60.

3. Meldrum ML. A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematol Oncol Clin North Am* 2000; 14 (4): 745–60.

4. Overall JE, Doyle SR. Estimating sample sizes for repeated measurement designs. *Controlled Clinical Trials* 1994; 15, 100-123.

5. Muller KE, Barton CN. Approximate power for repeated measures ANOVA lacking sphericity. *Journal of the American Statistical Association* 1989; 84, 549-555.

6. Liu A, Shih WJ, Gehan E. Sample size and power determination for clustered repeated measurements. *Stat Med* 2002; 21:1787–1801.

7. Aitken AC. On Least-squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh* 1934; 55: 42–48.

8. Self S Mauritsen R. Power/sample size calculations for generalized linear models. *Biometrics* 1988; 44, 79-86.

9. Zeger SL, Liang KY, Albert PS. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* 1988; Vol. 44, No. 4, pp. 1049-1060.

10. Galecki AT. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics, Theory and Methods* 1994: 23: 3105-3120.

11. Litttell RC, Pendergast J, Natarajam R. Tutorial in Biostatistics: modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* 2000; 19: 1793-1819. 12. Fisher RA. Applications of "Student's" distribution. *Metron* 1925; 5: 90–104.

13. Satterthwaite FE. Synthesis of Variance. *Psychometrika* 1941; 6, 309-316.

14. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; Vol. 53, No. 3: 983-997.

15. Litttell RC, Pendergast J, Natarajam R. Tutorial in Biostatistics: modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* 2000; 19: 1793-1819.

16. Wolfinger RD. Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics* 1996; Vol. 1, No. 2, 205-230.

17. http://www2.sas.com/proceedings/sugi29/188-29.pdf

18. Fleiss JL. Design and analysis of clinical experiments, Wiley, New York, 1986, Chapter 7.

19. Centers for Medicare and Medicaid Services Five Star Quality Rating System, https://www.cms.gov/medicare/provider-enrollment-and-certification/certificationandcomplianc /fsqrs.html

20. Pakker NG, Notermans DW, de Boer RJ, Roos MT, de Wolf F, Hill A, Leonard JM, Danner SA, Miedema F, Schellekens PT. Biphasic kinetics of peripheral blood T cells after triple combination therapy in HIV-1 infection: a composite of redistribution and proliferation. *Nat Med* 1998 Feb;4(2):208-14.

21. de la Rosa R, Leal M. Thymic involvement in recovery of immunity among HIV-infected adults on highly active antiretroviral therapy. *Journal of Antimicrobial Chemotherapy* 2003; 52, 155–158.

Appendix 1: Design Matrix

For (1) with the general parameter vector $\underline{\beta}=(\alpha, \beta_{-(b-1)}, \ldots, \beta_{-1}, \beta_1, \ldots, \beta_k, \theta)$, the corresponding design matrix has columns $(I, J_{-(b-1)}, \ldots, J_{-1}, J_1, \ldots, J_k, Z)$.

The design matrix X is made up of $(n_0 + n_1)$ row-stacked $X_{h,i}$'s, where $X_{h=0,i}$denotes the partial design matrix for each unit in the untreated group and $X_{h=1,i}$denotes for each unit in the treated group. Note the $(T+1)^{th}$column stands for intervention effect θ. That is:

$$X = \begin{bmatrix} X_{h=0,1} \\ \vdots \\ X_{h=0,n_0} \\ X_{h=1,1} \\ \vdots \\ X_{h=1,n_0} \end{bmatrix} \text{ where}$$

$$X_{h=0,i} = \begin{bmatrix} 1 & 1 & \ldots & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & -1 & \cdots & -1 & -1 & \cdots & -1 & 0 \end{bmatrix}_{T*(T+2)}$$

$$X_{h=1,i} = \begin{bmatrix} 1 & 1 & \ldots & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & -1 & \cdots & -1 & -1 & \cdots & -1 & 1 \end{bmatrix}_{T*(T+2)}$$

Appendix 2: Covariance Matrix and GLS Estimate

The goal is to find $(X'V^{-1}X)^{-1}$as the most lower right element of $(X'V^{-1}X)^{-1}\sigma^2$ is $Var(\hat{\theta})$. First under CS where $\rho_{jj'} \equiv \rho$, the covariance matrix in (2) reduces to $V_{CS} =$

$$\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}_T \text{ and } V^{-1} = \begin{pmatrix} {V_0}^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & {V_0}^{-1} \end{pmatrix}_{(n_0+n_1)T} \text{ with}$$

$${V_0}^{-1} = \frac{1}{[1+(T-1)\rho](1-\rho)} \begin{pmatrix} 1+(T-2)\rho & \cdots & -\rho \\ \vdots & \ddots & \vdots \\ -\rho & \cdots & 1+(T-2)\rho \end{pmatrix}_T .$$

Then we apply the technique for the inverse of the partitioned matrix.

$$(X'V^{-1}X)^{-1} = \begin{bmatrix} A_{11} & A_{21} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} B_{11} & B_{21} \\ B_{21} & B_{22} \end{bmatrix}$$

where $B_{22} = (A_{22} - A_{21}{A_{11}}^{-1}A_{12})^{-1}$and $Var(\hat{\theta})$is contained in B_{22} to derive this simple closed form formula for GLS-CS estimate of variance.

$$(X'V^{-1}X)^{-1} = \left(\frac{1}{n_0}+\frac{1}{n_1}\right)[1+(T-1)\rho](1-\rho)$$

$$\begin{bmatrix} 2T(1-\rho) & 2(1-\rho) & \cdots & 2(1-\rho) & 0 \\ 2(1-\rho) & 2[1+(T-2)\rho] & \cdots & -2\rho & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 2(1-\rho) & -2\rho & \cdots & 2[1+(T-2)\rho] & 0 \\ 0 & 0 & \cdots & 0 & k[1+(b-1)\rho] \end{bmatrix}^{-1}$$

$$= \left(\tfrac{1}{n_0}+\tfrac{1}{n_1}\right)[1+(T-1)\rho](1-\rho)\begin{bmatrix} A_{11} & A_{21} \\ A_{21} & A_{22} \end{bmatrix}^{-1},$$

where $A_{11} = \begin{bmatrix} 2\mathrm{T}(1-\rho) & 2(1-\rho) & \cdots & 2(1-\rho) \\ 2(1-\rho) & 2[1+(T-2)\rho] & \cdots & -2\rho \\ \vdots & \vdots & \ddots & \vdots \\ 2(1-\rho) & -2\rho & \cdots & 2[1+(T-2)\rho] \end{bmatrix}$,

$$A_{22} = \left[\frac{k[1+(b-1)\rho]}{2}\right] \text{ and } A_{21} = A_{12}' = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Because $A_{21}{A_{11}}^{-1}A_{12} = 0$,

$$B_{22} = (A_{22} - A_{21}{A_{11}}^{-1}A_{12})^{-1} = {A_{22}}^{-1} = \frac{1}{k[1+(b-1)\rho]}.$$

Thus $Var(\hat{\theta}_{R-CS}) = (\frac{1}{n_0} + \frac{1}{n_1}) \frac{[1+(T-1)\rho](1-\rho)}{k[1+(b-1)\rho]} \sigma^2$.

Chapter 2 Power / Sample Size Estimation for Non-Randomized Difference-in-Differences Studies

Abstract

Intervention effect on continuous chronic normal or normal approximated conditions is often estimated in two-arm longitudinal clinical trials with T=b+k total time points. One arm receives the intervention with b≥1 pre- and k≥1 post-intervention measures while the other arm is untreated for all T times. *Although randomization of which units receive treatment is preferred, non-randomized designs using Difference-in Differences (DD) analyses are often necessary for practical issues. Estimated variance of the intervention effect that incorporates the covariance structure of repeated measures are needed for power/sample size estimation of DD analyses. We develop Generalized Least Squares (GLS) based tools for variance of the intervention effect estimate in non-randomized DD studies using compound symmetry (CS) and Toeplitz covariance. For compound symmetry (CS) repeated measure correlation, a closed form variance of the estimated intervention effect was derived, and is minimized (hence power maximized) with equal number of pre-and post-intervention measurements (b=k) for T even and |b-k|=1 for T odd. While given the same b and k, randomized designs are superior, non-randomized designs deliver nearly as precise estimates of intervention effect for high within-unit correlation and/or with more baseline than follow-up measurements (b ≫ k). However, CS correlation structure did not hold for the four longitudinal nursing hospital and HIV examples we evaluated where T=7. We thus calculated the power directly for*

these examples using the observed Toeplitz covariance. Again randomization improved the precision of the intervention effect estimate for these examples. As normative data for Toeplitz correlation may be difficult to derive or implement in practice, we explored simpler approaches to approximation of the variance of the intervention effect in this setting. However, none of these approximations performed uniformly well. But one of these approach revealed that in some cases T=2 longitudinal measures with b=1 and k=1 obtained nearly as precise estimates of the intervention effect as did any design with T=b+k=7 measures.

1. Introduction

Clinical trials and other prospective studies often evaluate repeated measurements of continuous normal (or normal approximated) chronic outcomes on treated units at systematic time points before and after an intervention compared to the same times on controls whom are never treated [1-3]. In our nomenclature "units" could refer to "treatment/care facilities" such as nursing homes or refer to "persons". While randomization of which units are treated is preferred to improve precision, it is not always feasible; particularly in health economics and services research. Thus non-randomized designs known as Difference-in-Differences (DD) analyses are widely applied [2, 3] to estimate the impact of new interventions or policies introduced at a given time point into non-randomized facilities (or individuals), compared to controls continuing on the existing regimen. Units in both arms introduced to the intervention and the controls are measured at the same T longitudinal time points. The outcome being affected by the intervention is measured at b consecutive time points (denoted $-b, -(b-1) \ldots -1$) prior and k consecutive time points (denoted $1, 2 \ldots k$) after the intervention is introduced to the intervention arm with $b+k=T$. The difference in outcomes for the intervention arm during the b pre- and k post-intervention periods is compared to that for the control arm. Difference-in-differences analysis is best applied using a mixed model framework that adjusts for serial correlation of repeated measures within the same intervention facility or individual [4].

We assume for this chapter "non-randomized" allocation to intervention and control arms is done by convenience or other processes that are not purposely based on levels of the outcome over the first b time points. For example, maybe hospitals that are closer to

a university are assigned the intervention developed at that university. Still, the pre-intervention levels of the outcome may differ by an unknown amount between the intervention arms due to confounding from the criteria that such "circumstance" allocation was based on. For example, the hospitals closer to the University may have worse pre-existing levels of the outcome. Such differences will (i.e. with or without an intervention) continue into the *k* post-intervention measurements. This contrasts with a regression to the mean phenomenon [5, 6], which would be generated if units (i.e. hospitals) that were performing worst prior to the initiation of the intervention were deliberately over-selected (or under-selected) to be given the intervention. That setting is discussed in Appendix 1, but again is assumed not to exist in this chapter.

To design and plan a longitudinal study in evaluating a new intervention, it is important to estimate whether one has a large enough sample for adequate power to detect a reasonable intervention effect. This depends on what the variance of the intervention effect estimate will be, which among other things, depends on the often-unknown correlation structure between repeated measures of the same unit. We develop variances of the intervention effect estimate using generalized least squares (GLS) models based on i) the simplest repeat-measure correlation structure (compound symmetry) and ii) a more complex, but more empirically tenable Toeplitz correlation structure.

The chapter is organized as follows: Section 2 presents the general linear model (GLM) for DD analysis. Section 3 introduces general GLS variance/power formulas for intervention effect estimates from GLM. Section 4 provides insights on allocation of *T* into *b: k* and discusses the penalty of non-randomization (versus randomization) if

compound symmetry repeat-measure correlation (CS) holds. Section 5 investigates the Toeplitz repeat-measure correlation and discusses the optimal *b: k* allocation using illustrative examples from real data where CS does not hold. Section 6 develops and compares simple / conservative variance estimates for Toeplitz covariance when there is uncertainty about the actual Toeplitz structure. Section 7 summarizes and discusses possible future work.

2. Difference-in-Differences Design Examples and General Linear Model

In DD studies, Let *Y* denote the longitudinal continuous outcomes observed at *b* times before and *k* times after the intervention implementation in the new intervention arm and at all $T = b + k$ times in the control arm; *j* = {-*b*, -(*b*-1), …, -1, 1, 2, …, *k*} denotes the ordered times with before the intervention is implemented being {-*b*, -*(b-1),* …, -1} and after the intervention is implemented being {1, 2, …, *k*}; *h* denotes the intervention arm with *h=0* for control and *h=1* for the new intervention. There are n_0 units receiving the control and n_1 receiving the new intervention (or n in each if $n_0 = n_1$); unit (nested within intervention arm) is denoted by *i*. Thus Y_{1ij} represents the measure at time *j* from unit *i* in the new intervention arm and $Y_{0i',j'}$ represents the measure at j' from unit i' in the control arm. For example, consider a trial with $n_0 = n_1 = n = 30$ hospitals in each arm, let *i* denote hospitals (as "units") where *i=1, …,* n_h. For the intervention arm (*h=1*), "units" are followed for *T*=7 years total with *b*=2 years (2001 to 2002) prior and *k*=5 years (2003 to 2007) after the intervention implementation. Thus $Y_{1,3,-2}$ and $Y_{0,17,3}$ respectively denote the measure taken in 2001 (2 years prior to start of the

intervention) in the 3^{rd} hospital of the intervention arm and 2005 (3 years after the start of the intervention) in the 17^{th} hospital of the control arm, respectively. We assume complete data with $T=b+k$ measures observed on each unit, which, in particular, is reasonable when the units are facilities that are required by regulations to keep regular records of the outcomes of interest.

The Y_{hij} measure is decomposed as:

$$Y_{hij} = \alpha_0 + \alpha_1 I_{\{h=1\}} + \beta_j + \theta Z_{hj} + \varepsilon^*_{ij} \quad (1)$$

With $\alpha_0$denoting the average baseline value taken at *j=-b* for control units; $\alpha_1$denoting a potentially different central tendency at *j=-b* from the "circumstance" selection of *h=1* versus *h=0; a difference* that continues onto subsequent times; β_j denoting the fixed effects corresponding to time *j* (*j = -(b-1), …, -1, 1, …,k* relative to *j=-b*);. Now $Z_{hj} = I_{\{h=1, j>0\}}$ is an indicator of if the intervention is delivered in arm *h* at time *j* with 1=intervention delivered and 0=control delivered while θ is the size of the intervention effect on the outcome as described below. Any random unit (i.e. *i* level) effects are subsumed into the within-unit error term ε^*_{ij}, where $\varepsilon^*_{ij} \sim N(0, \sigma^2 V)$ with the correlation matrix *V* defined in (2).

We assume an immediate "jump effect" of size θ after the intervention begins that remains unchanged at subsequent time points, with $j \geq 1$. Note that other functions such as linear intervention effect increase $j * \theta Z_{hj}$ or threshold followed by exponential decay $e^{-\frac{j}{m}} * \theta Z_{hj}$ for some constant $m > 0$ are possible. However, there may be settings where

an immediate "jump effect" is appropriate such as when the intervention is a process change at a medical facility that is implemented quickly, a drug that the body does not develop resistance or acclimation to, or an immediately successful behavioral intervention. Even if the intervention impact (or differential intervention) was not exactly immediate jump it could be close to this.

3. GLS Variance / Power Estimate Framework

3.1. GLS variance estimate

The matrix form of (1) can be written as: $Y = X\underline{\beta} + \varepsilon^*$, where $\varepsilon_{ij}^* \sim N(0, \sigma^2 V)$. Here X represents the design matrix and Y is a vector of outcomes. For (1) with the general parameter vector $\underline{\beta}=(\alpha_0, \alpha_1, \beta_{-(b-1)}, \ldots, \beta_{-1}, \beta_1, \ldots, \beta_k, \theta)$, the corresponding X has columns $(I, I_{\{h=1\}}, J_{-(b-1)}, \ldots, J_{-1}, J_1, \ldots, J_k, Z)$, with $(n_0 + n_1)*T$ rows per column. Z is a column/vector of intervention indicators Z_{hj} coded (0, 1) as defined above; $J_{-(b-1)}, \ldots, J_{-1}, J_1, \ldots, J_k$ are columns corresponding to $b+k-1$ independent time coded variables as follows: for j= (-(b-1), -(b-2), …, -1, 1, 2, …k), J_j= {-1 at time –b (reference); 1 at time j; and 0 at all other times}. There is no column for J_{-b} as $\beta_{-b} = -\sum_{j=-(b-1)}^{k} \beta_j$ under the fixed effects constraint $\sum_{j=-b}^{k} \beta_j = 0$. Thus X (where $X_{h,i}$ as defined in Appendix 2) and V can be written as:

$$X = \begin{pmatrix} X_{h=0,i=1} \\ \vdots \\ X_{h=0,i=n_0} \\ X_{h=1,i=1} \\ \vdots \\ X_{h=1,i=n_0} \end{pmatrix}_{(n_0+n_1)\,T*(T+1)} , V = \begin{pmatrix} V_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_0 \end{pmatrix}_{(n_0+n_1)T},$$

$$\text{where } V_0 = \begin{pmatrix} \rho_{11} & \rho_{12} & \rho_{13} & \cdots & \rho_{1,T-1} & \rho_{1,T} \\ \rho_{21} & \rho_{22} & \rho_{23} & \cdots & \rho_{2,T-1} & \rho_{2,T} \\ \rho_{31} & \rho_{32} & \rho_{33} & \cdots & \rho_{3,T-1} & \rho_{3,T} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{T-1,1} & \rho_{T-1,2} & \rho_{T-1,3} & \cdots & \rho_{T-1,T-1} & \rho_{T-1,T} \\ \rho_{T,1} & \rho_{T,2} & \rho_{T,3} & \cdots & \rho_{T,T-1} & \rho_{TT} \end{pmatrix}_T . \quad (2)$$

The covariance matrix V is made up with $(n_0 + n_1)$ block T diagonal matrices $V_0's$ with all off-block diagonal matrix elements being 0. The most basic assumptions for the error term is that measures are independent between units, and correlation structure is invariant given two time points j and j' for any unit, i.e., $\rho_{i,jj'} = \rho_{i',jj'}$ $(i \neq i', j \neq j')$ The within-unit correlation structure ($\rho_{jj'}$) is often unknown in advance. Typically, the correlation for any two time points is generally non-increasing, i.e., the closer the two time points are, the higher the correlation is; as they are further away, they become less correlated [10, 11]. We will restrict correlation assumptions further later in the chapter.

Note that for the randomized setting where units are randomized into the intervention arm, the baseline measurements from both groups are the same due to randomization where $\alpha_1 = 0$, and thus renaming the only intercept parameter α_0 as α. The parameter vector reduces to $\underline{\beta}=(\alpha, \beta_{-(b-1)}, \dots, \beta_{-1}, \beta_1, \dots, \beta_k, \theta)$ where $\alpha \equiv 1$ and the other parameters as defined before. Appendix 2 presents the full expansion of design matrix for the non-randomized (NR) as well as the randomized (R) setting.

The GLS estimate for $\underline{\beta}$ is $\underline{\hat{\beta}}$, which is the best linear unbiased estimator (BLUE) for $\underline{\beta}$ and uniform minimum variance (UMVU) if Y_{hij} is normally distributed [9]. The GLS variance for $\underline{\hat{\beta}}$ is Λ being a square matrix of order T+1. The variance of $\hat{\theta}$ is the lowest-right diagonal element of Λ.

$$\underline{\hat{\beta}} = (X'V^{-1}X)^{-1}X'V^{-1}Y; \tag{3}$$

$$\Lambda = (X'V^{-1}X)^{-1}\sigma^2. \tag{4}$$

3.2. General Power Estimation Approach

We consider $H_0: \theta = 0$ versus $H_A: \theta = \pm\theta_A$. Without loss of generality, $\delta = \frac{\theta_A}{\sigma}$ is a predefined clinically important effect size in terms of standard deviation, while $\alpha \; and \; \beta$ are Type I and Type II errors, respectively. We have the following equations of power $(1-\beta)$ using the notation from [8], in which $Var(\hat{\theta})$ is derived from the GLS variance estimate in equation (4).

$$\theta_A = (z_{1-\frac{\alpha}{2}} + z_{1-\beta})\sqrt{Var(\hat{\theta})}. \tag{5}$$

For practical repeated measure designs, the normal approximation of the non-central *t* distribution is applied for studies with relative large sample sizes [14]. In specific, the two distributions are almost identical when degrees of freedom (DF) $\gamma > 30$. For smaller sample sizes, it may be appropriate to approximate degrees of freedom (DF) (γ) in non-central *t* distribution for the mixture variance (for example, by Satterthwaite's (1946) [15], and Kenward-Roger's (1997) approximations [16]) and adjust equation (5) for this. But the full details are beyond the scope of this chapter.

4. Power Estimation Framework based on Compound Symmetry

4.1. GLS variance under compound symmetry

Compound symmetry (CS), also denoted sphericity or equi-correlation [11, 12], is a commonly used within-unit covariance matrix for $\boldsymbol{\varepsilon}^*$ in this type of setting [8] as CS only requires assumption of one unknown correlation parameter. It can be reasonable to expect that the largest (and perhaps only) covariance component of $\boldsymbol{\varepsilon}_{ij}^*$ within the same unit *i*, would be a main effect for the unit *i* with smaller ignorable within-unit temporal changes over *j*. While little empirical research has been done to confirm this structure holds, one study finds that CS was a reasonable simplification in quantitative planning of repeated measures clinical trials [8]. In this section, we derive the GLS estimate replacing V_0 in equation (2) with V_{CS} having a common ρ. Note that $\rho = \frac{\tau^2}{\tau^2+\sigma_e^2}$ where τ^2 is a unit (*i*-level) variance and σ_e^2 is an independent unit-visit (*ij*-level) variance as described in Appendix 4. For *T*=7, $V_0 = V_{CS} = \begin{bmatrix} 1 & \rho & \rho & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & \rho & \rho & 1 \end{bmatrix}$.

Under the assumption of CS we derive a closed form GLS-CS formula for the variance of $\hat{\theta}$ as follows. Using the inverse formula for a partitioned matrix described in Appendix 3, the following GLS-CS variance estimate depends on *b*, *k*, σ and ρ.

For non-randomized designs (NR-DD) under CS in equation (1),

$$Var(\hat{\theta}_{NR-CS}) = (\frac{1}{n_0} + \frac{1}{n_1}) \frac{(b+k)(1-\rho)}{bk} \sigma^2; \quad (6)$$

For randomized designs (R-DD) under CS,

$$Var(\hat{\theta}_{R-CS}) = (\frac{1}{n_0} + \frac{1}{n_1}) \frac{[1+(b+k-1)\rho](1-\rho)}{k[1+(b-1)\rho]} \sigma^2. \quad (7)$$

Note that as Appendix 4 shows, compared to fitting any R-DD, the NR-DD with non-randomization effects under the same setting will result in a possibly lower model within-population measurement variance on Y_{ij} ($\sigma^2_{NR} \leq \sigma^2_R$) together with a possibly smaller within-unit correlation of Y_{ij} andY_{ij}, ($\rho_{NR} \leq \rho_R$) due to elimination of variance from about a common α to that about dispersed intercepts, α_0 and $\alpha_0 + \alpha_1$. However, from equation (6), $Var(\hat{\theta}_{NR-CS})$ only depends on $\sigma^2$and ρ through the product $(1-\rho)\sigma^2$. To that end Appendix 4 shows this product is unchanged by application of the fixed effects NR-DD design in that it turns out that always, $(1-\rho_{NR})\sigma^2_{NR} = (1-\rho_R)\sigma^2_R$. This invariance property means that *under compound symmetry*, the "pre non-randomization study design" effect parameters for $\sigma^2$and ρ can be used in equation (6) no matter what the impact of the fixed effects NR-DD on the final $\sigma^2$and ρ is.

To calculate the number of units needed in each arm to achieve a given power $(1-\beta)$, if CS correlation structure is used then plugging equations (6-7) into (5) and reorganizing, gives the number of units (n_{NR-CS}) needed to obtain given power for NR-DD studies:

$$n_{NR-CS} = \frac{(\mathrm{b+k})(1-\rho)}{bk\delta^2} (z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2; \quad (8)$$

and the number of units (n_{NR-CS}) needed to obtain given power using R-DD studies is:

$$n_{R-CS} = \frac{[1+(b+k-1)\rho](1-\rho)}{k[1+(b-1)\rho]\delta^2}(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2. \quad (9)$$

Note that equation (8) is symmetric on *b* and *k* in that *(b', k')=(k, b)* gives the same variance as (*b*, *k)*. This symmetry follows from the symmetry of V_0 and can also be shown through a reformulation of the problem that reverses the timescales and thus (*b'*, *k'*) replaces α_1 with $\alpha_1 + \theta$ and θ with $-\theta$.

4.2. Non-Randomized Versus Randomized Designs under Compound Symmetry

While it is known that randomization is superior to non-randomization, as randomized studies can be more costly and logistically more difficult to conduct, the relative superiority may be important to know. Under CS, the ratio of the variances of non-randomized to randomized DD studies is calculated below using equation (6) and equation (7) with the randomized setting as a reference.

$$\frac{Var(\hat{\theta}_{NR-CS})}{Var(\hat{\theta}_{R-CS})} = \frac{(b+k)[1+(b-1)\rho]}{b[1+(b+k-1)\rho]} = 1 + \frac{k(1-\rho)}{b[1+(b+k-1)\rho]} > 1. \quad (10)$$

The ratio in equation (10) indicates $Var(\hat{\theta}_{NR-CS}) > Var(\hat{\theta}_{R-CS})$, and thus confirms that power in randomized design is greater than the non-randomized setting with the same design parameters. As $\rho \to 1$, the ratio goes to 1, meaning the randomized design behaves similar to, but still better than the non-randomized design when ρ is close to 1. As $\rho \to 0$, the ratio reduces to $1 + \frac{k}{b}$, meaning an NR-DD requires $(1 + \frac{k}{b})$ times more units than the comparable R-DD design to achieve the same power. Thus increasing *k* or decreasing *b* (with all other parameters fixed) can lead to more advantages in conducting randomization. For $b \geq k$, the ratio lies within (1, 2); for very small $\frac{k}{b}$, the ratio is close to

1, meaning that randomization does not significantly reduce variance of the intervention effect estimate.

Figure 1 provides examples with the number of pre-visits (b) varying between 1 and 6 for T=7. Note that here and elsewhere, we chose the total number of visits to be 7 as it seems reasonable for the four examples presented later in Section 5 and other settings where trials would be conducted over periods of 2-3 years with repeated measures at 3-6 month intervals. For $\rho \geq 0.6$ and $b \geq 2$, non-randomization performs close to randomization as the variance ratio is less than 1.22. But for b=1 variance from non-randomization did not approach that from randomization until $\rho > 0.8$ where the variance ratio was 1.21.

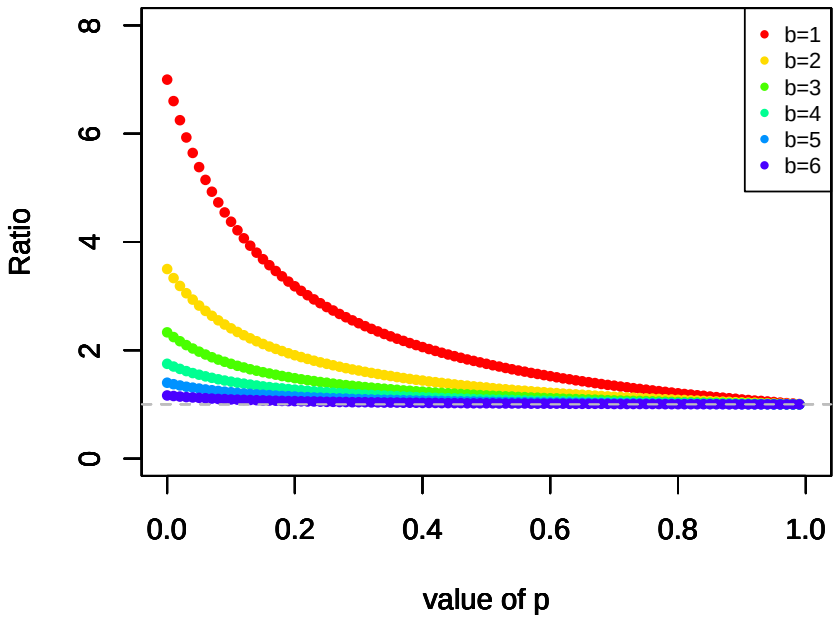

Figure 1: Ratio of variances in NR-DD versus R-DD assuming CS structure ($T = b + k = 7$)

4.3. Power by *b*: *k* Allocation for Non-Randomized Designs under Compound Symmetry

Focusing now on the non-randomized design (the main subject of this chapter), in practice, the total number of visits *T (T=b + k)* will likely be fixed because of budget and/or time constraints. To investigate the optimal allocation of *b: k* that maximizes power of NR-DD designs by minimizing $Var(\hat{\theta}_{R-CS})$, we find the optimal b^* where $b^* = \arg\underbrace{\max}_{b} \text{Power} = \arg\underbrace{\min}_{b} \text{Var}(\hat{\theta}_{R-CS})$ under the constraint of $T=b + k$. For the CS, setting the derivative of $\log(Var(\hat{\theta}_{NR-CS}))$ in equation (6) with respect to b be 0 yields $\left(\frac{1}{T-b} - \frac{1}{b}\right) = 0$ which occurs when $b = k$. Optimal allocation to minimize the variance and maximize the power is $b^* \approx k^*$. Thus if *T* is even, then $b^* = k^* = \frac{T}{2}$; if *T* is odd, then $b^* = \frac{T-1}{2}$ *or* $\frac{T+1}{2}$. For example, if *T*=6, then $b^* = k^* = \frac{T}{2} = 3$; if *T*=7, then $(b^*, k^*) = (3, 4)$ or equivalently $(b^*, k^*) = (4, 3)$.

We can also quantitatively measure the impact on $Var(\hat{\theta}_{NR-CS})$ by comparing the ratio of this variance from any given (b, k) versus the optimal (b^*, k^*) assuming constant *T*, i.e., $\frac{\text{Var}(\hat{\theta}_{NR-CS})|(b,k)}{\text{Var}(\hat{\theta}_{NR-CS})|(b^*,k^*)} = \frac{b^*k^*}{bk}$ using equation (6). For example (*T=7*), when $(b, k) = (1, 6)$, the ratio of variance is $\frac{b^*k^*}{bk} = \frac{3*4}{1*6} = 2$, meaning that $Var(\hat{\theta}_{NR-CS})$ for $(b, k) = (1, 6)$ is twice than that for the optimal $(b^*, k^*) = (3, 4)$; but when $(b, k) = (2, 5)$, the ratio of variance is only $\frac{b^*k^*}{bk} = \frac{3*4}{2*5} = \frac{6}{5}$ = 1.20. Thus if b is close to b^*, the ratio is close to 1 and differences in power estimation are small. However, as b goes further away from b^*, the ratio gets larger and differences in power estimation become meaningful.

5. GLS Power Estimation using Toeplitz Structure

While compound symmetry has led to simple and useful closed form variance and power formulas, the issue of how well this structure fits in practice needs to be broached. The first step in this process is to have an alternate and usable correlation. We propose the Toeplitz (TP) which is more general than CS, and present power estimation with four illustrative examples.

5.1. GLS variance estimate given Toeplitz correlation

In loosing the assumption beyond CS, an essentially necessary assumption for estimability is that correlations be stationary over chronological time and thus a function of the difference in j and j' ($\rho_{jj'} \equiv \rho_{|j-j'|}$) [11, 12] as otherwise it is impossible to project historical normative data to a future study. This is known as Toeplitz structure (V_{TP}) with correlations denoted ρ_1 for $|j-j'|=1$, ρ_2 for $|j-j'|=2, \ldots, \rho_{b+k-1}$ for $|j-j'|=b+k-1$ as illustrated in (11) for $T=7$.

$$V_0 = V_{TP} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \rho_5 & \rho_6 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \rho_5 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_5 & \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_6 & \rho_5 & \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}. \quad (11)$$

The generalized least squares (GLS) estimate of all the parameters in (1) is then given in equations (4, 5), with $V_0 = V_{TP}$. Now ρ_j is generally non-increasing with $|j-j'|$ where CS is the special case of $\rho_j = \rho$; and for $T=7$, if ρ_6 is not too much less than ρ_1 then perhaps CS as an approximation would be good.

We now estimate Toeplitz correlation structures for four illustrative examples. The first two are from data collected on 365 New Jersey nursing homes being monitored every three months from the second quarter of 2011 to the fourth quarter of 2012 (seven quarters total) in the Nursing Home Compare [17] for proportions of: 1) long stay residents with long term need for help with activities of daily living (LS_ADL); and 2) short term stay patients that experience moderate to severe (SS_Pain). Higher levels of both LS_ADL and SS_Pain are undesirable and targeted for improvement at a facility level. The "unit" for these examples is the facility with the repeated measure being quarterly facility averaged values. Thus, for example, in a future study it is conceivable that all 365 facilities could be followed for b baseline time points to obtain LS_ADL and/or SS_Pain proportions and then around 50% be moved to a facility intervention to improve one or both of these with k post-intervention measures obtained from both groups for comparison of change.

The second two examples are obtained from 224 Bronx HIV infected women [18] who had complete data for their first seven semiannual visits for CD4 counts and CESD Depression scores [19]. Higher CD4 and lower CESD are desired and have been previously targeted for interventions. The repeated measures for these examples are from semiannual visits of patients. It is conceivable that in a future study these patients could be followed for b baseline visits to obtain CD4 and / or CESD scores and then around 50% be put on an intervention to improve one or both of these with k post-intervention measures obtained from both groups for comparison of change. Again, we chose $T=b+k=7$ which is reasonable not only for our examples but also for trials conducted over 2-4 years with repeated measures at 3-6 months interval.

Table 1 and Figure 2 summarize the Toeplitz correlation structures for the four outcomes described above estimated using the mixed procedure in SAS from the normative data. Visually, Figure 2 illustrates a range from starting correlations at ρ_1 of ~0.60 to ~0.87 and in slight to steep generally monotonic linear declines going down to ρ_6 ranging from ~0.34 to ~0.55.

Table 1: Toeplitz correlation structures from four examples

Time	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6
Among Quarterly Evaluations of 365 New Jersey Nursing Homes						
LS_ADL	0.59	0.47	0.41	0.39	0.40	0.34
SS_Pain	0.87	0.76	0.69	0.66	0.63	0.54
Among Semiannual Visits of 1012 HIV-Infected Bronx-WIHS Patients						
CD4	0.84	0.74	0.65	0.57	0.46	0.47
CESD	0.64	0.59	0.54	0.53	0.52	0.55

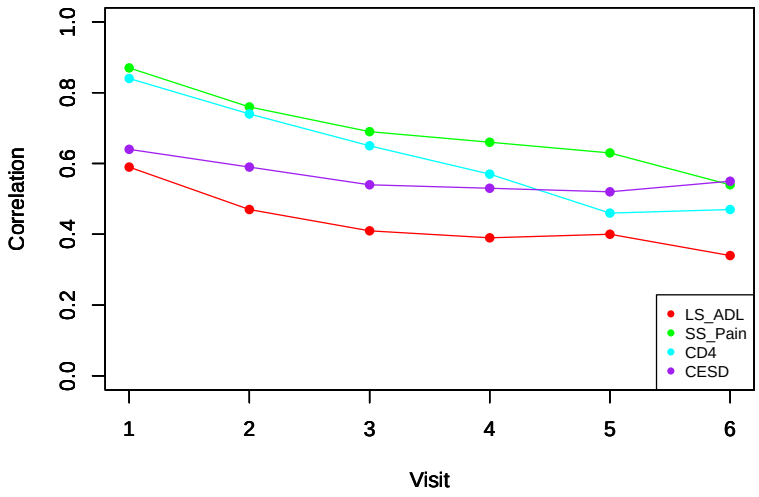

Figure 2: Visualization of Toeplitz correlation structures from real examples ($T = b + k = 7$)

For example, the correlations of CD4 and SS_Pain start higher at $\rho_1 \approx 0.85$ than do those of LS_ADL and CESD at $\rho_1 \approx 0.60$. For CESD, the decline is $\rho_1 - \rho_6 \approx 0.10$ which might be qualitativley close to CS. The decline for SS_Pain, LS_ADL and CD4 is $\rho_1 - \rho_6 \approx 0.35$, thus these correlation structures are not close to compound symmetry. The remainder of Section 5 presents power estimates and properties for these four examples based on the empirical Toeplitz correlation structures in Table 1 and Figure 2. In Section 6 we evaluate whether compound symmetry or another simple approach can be used to get good estimates of variance and power for these examples in the case of where the Toeplitz structure could not be estimated or where a robust approach is needed.

5.2. Power by *b*: *k* Allocation for Given Toeplitz Correlation

As we could not derive a simple variance estimate for the general Toeplitz structure, we begin by presenting the empirical Toeplitz variance estimates for the four examples with the added goal of replicating Section 4.2 comparing randomized versus non-randomized variance by *b*: *k* allocation for the given Toeplitz structures in Table 1. We computed the $Var(\hat{\theta})$ in equation (4) by computer for the given Toeplitz correlation structure over all possible *b*: *k* allocations for each of the four examples. Without loss of generality and to facilitate comparisons, these computations assumed the variance of each outcome was $\sigma = 1$ and $n_0 = n_1 = 30$. Figure 3 presents variances for both non-randomized and randomized designs; the number of pre- and post-intervention (*b*, *k*) were allowed to be (1, 6); (2: 5); (3, 4); (4, 3); (5, 2); (6, 1). The solid lines stand for the variance under V_{TP} of non-randomized designs, $Var\left(\hat{\theta}_{NR-TP}\right)$, and the dotted lines for

that from randomized designs $Var(\hat{\theta}_{R-TP})$. The optimal *b*: *k* allocation to minimize $Var(\hat{\theta}_{NR-TP})$ in each example occurs at the value of *b* where the solid line hits the minimum.

It should be noted that while comparisons of $Var(\hat{\theta}_{NR-TP})$ to $Var(\hat{\theta}_{R-TP})$ in Figure 3 assume that the NR and R designs has the same repeated measure correlation / variance structure, we acknowledge that the imposition of NR designs may change the correlation / variance structure. Appendix 4 showed that for CS repeated measure correlation, the impact on the variance and correlation structure from absorption of variance by imposing an NR design on an R setting canceled out in terms of $Var(\hat{\theta}_{NR-CS})$ in equation (6). But we are not able to derive this same result for Toeplitz correlation in general. So while we believe the comparisons of $Var(\hat{\theta}_{NR-TP})$ to $Var(\hat{\theta}_{R-TP})$ given Toeplitz in Figure 3 are qualitatively meaningful, we acknowledge this limitation.

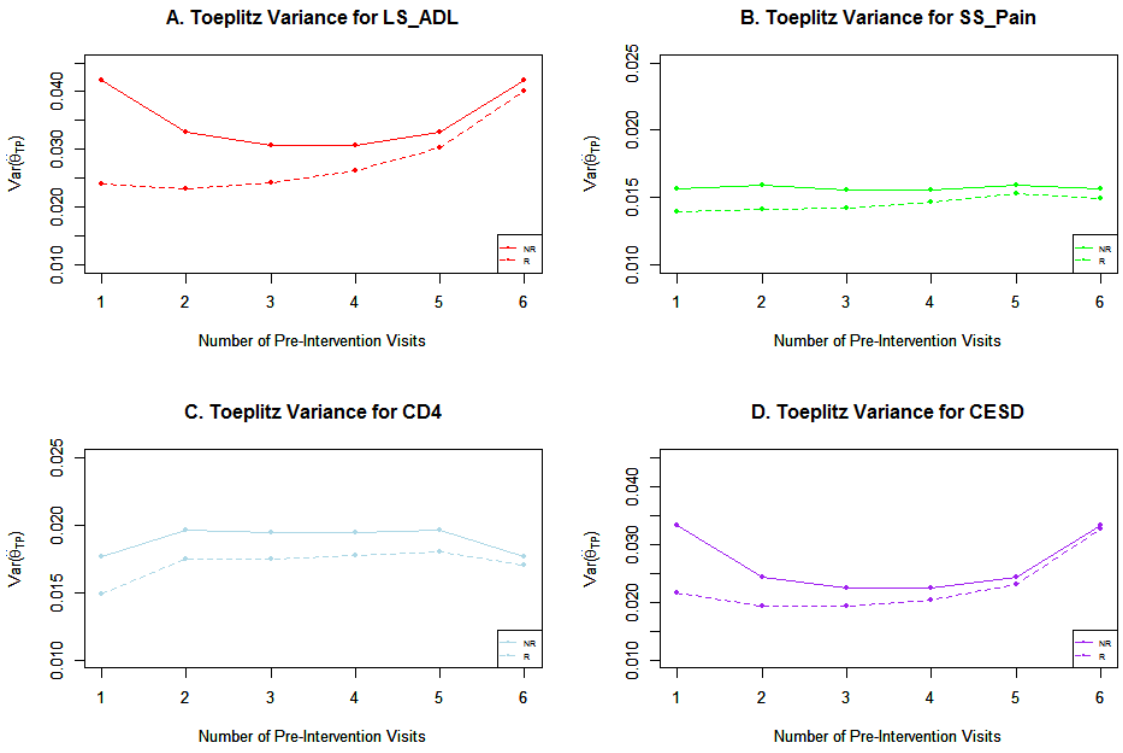

Figure 3: Ratio of variances in NR-DD versus R-DD from Toeplitz examples over all *b: k* allocations ($n_0 = n_1 = 30, \sigma = 1, T = b + k = 7$)

The examples in Figure 3 show that: i) The same optimal allocation of *b* and *k* under CS ($b \approx k$) held for the Toeplitz with three of the examples where (*b*, *k*) = (3, 4) and (*b*, *k*) = (4, 3) for NR-DD designs symmetrically had the smallest variances with *T=7*. But for CD4 in Figure 3-C, the variances from (*b*, *k*) = (1, 6) and (*b*, *k*) = (6, 1) are smaller than those from $b \approx k$. ii) Not surprisingly, the variances for the randomized setting were always smaller than those for non-randomized designs over all possible allocations. As *b* became larger the advantages for randomization decreased and were arguably minimal for $b \geq 2$ for CD4 (with ratio $\frac{Var(\hat{\theta}_{NR-TP})}{Var(\hat{\theta}_{R-TP})} = 1.12$ at *b = 2*) and SS_Pain (with

ratio $\frac{Var(\hat{\theta}_{NR-TP})}{Var(\hat{\theta}_{R-TP})} = 1.12$ at $b = 2$) and $b \geq 4$ for CESD (with ratio $\frac{Var(\hat{\theta}_{NR-TP})}{Var(\hat{\theta}_{R-TP})} = 1.10$ at $b = 4$) and LS_ADL (with ratio $\frac{Var(\hat{\theta}_{NR-TP})}{Var(\hat{\theta}_{R-TP})} = 1.16$ at $b = 4$); iii) As was described in the last paragraph of Section 4.1 because the Toeplitz correlation structure is symmetric, $Var(\hat{\theta}_{NR-TP})$ is identical for (b, k) and (b', k') where (b', k')=(k, b).

6. Power Estimation for Non-Randomized Designs using Simple Approximations to Toeplitz Correlation

If the actual structure of V_{TP} can be identified from normative data, it is ideal to use it as V_0 in $(X'V^{-1}X)^{-1}\sigma^2$ for variance / power calculation. However, in practice, investigators often have limited access or software to obtain V_{TP} from normative data in the limited time that is typically available to apply for study funding which may make estimation of power based on V_{TP} prohibitive. Furthermore, power/sample size estimates using V_{TP} could have unknown robustness properties against misspecification on $(\rho_1, \ldots, \rho_{T-1})$.

With so much complexity, uncertainly and difficulty in deriving parameters investigator may consider using heuristics approximations with less unknown parameters to estimate power and variance for non-randomized DD studies with Toeplitz covariance. We thus compared variances from several heuristic approximations to the actual $Var(\hat{\theta}_{NR-TP})$.

6.1. Compound Symmetry-Heuristics Approximation to Toeplitz Correlation

A compound symmetry structure with a common ρ is probably the simplest approximation for $Var(\hat{\theta}_{NR-TP})$ if obtaining V_{TP} seemed impractical or impossible as this ρ can be input into (6) or (8) as heuristically estimates to $Var(\hat{\theta}_{NR-TP})$.

However, which value of "approximatedρ" to use in (6) or (8) is not clear. One reasonable approach is to estimate (i.e. what is believed to be) the equi-correlation ρ with the weighted average of all (i.e. estimated) intra-unit correlations among the *T* time points, in the V_{CS} matrix that is being substituted for V_{TP}. That is use $\rho = \rho_{avg} = \frac{(b+k-1)\rho_1+(b+k-2)\rho_2+\cdots+\rho_{b+k-1}}{\sum_{i=1}^{b+k-1} i}$. For example, for LS_ADL with (*b*, *k*) = (3, 4) using the observed correlations from Table 1, $\rho_{avg} = \frac{6\rho_1+5\rho_2+4\rho_3+3\rho_4+2\rho_5+\rho_6}{21} = 0.47$.

A second approach that is more conservative is to let $\rho = \rho_{min}$ as the common correlation in the substituted $V_{CS}$where $\rho_{min}$is the minimum correlation in V_{TP}. The second approach is more conservative in power estimation than the first in that it obtains larger variances since the GLS-CS variance in (6) increases as ρ decreases. This ρ_{min}, typically would be $\rho_{1,b+k}$ if the correlations are decreasing with |*j-j'*|. For example, for LS_ADL with all values of (*b*, *k*), $\rho_{min} = \rho_6$ = 0.34.

The first two columns of Table 2, respectively, present CS parameters (ρ_{avg} and ρ_{min}) in V_{CS} approximation of $Var(\hat{\theta}_{NR-TP})$ using the observed correlations in all four examples of Figure 3. Again, by symmetry as described at the end of Section 4.1, ρ_{avg} in the first column also holds for (*b*, *k*)=(3, 4). The ρ_{min} given in the second column holds for all *T*=*b*+*k*=7.

Table 2: Calculated parameters for CS approximations and conservative approximations from the Toeplitz correlation structures in Table 1 for (b, k) = (3, 4) ($n_0 = n_1 = 30, \sigma = 1$)

Outcome in Table 1	Heuristics Approximations		Conservative Approximations			
	CS Parameters[1]		(b, k) = (1, 1)	PCS Parameters		
	ρ_{avg}	ρ_{min}	ρ_1	ρ_{pre}	ρ_{post}	ρ_{cross}
Among Quarterly Evaluations of 365 New Jersey Nursing Homes						
LS_ADL	0.47	0.34	0.59	0.55	0.52	0.42
SS_Pain	0.74	0.54	0.87	0.83	0.80	0.69
Among Semiannual Visits of 1012 HIV-Infected Bronx-WIHS Patients						
CD4	0.69	0.46	0.84	0.81	0.78	0.61
CESD	0.58	0.52	0.65	0.62	0.61	0.55

1. The CS Approximation parameters $\rho_{avg}$and ρ_{min} are invariant to (b, k)

Figure 4 compares the approximated variances from V_{CS} using $\rho = \rho_{avg}$ and $\rho = \rho_{min}$ for the four examples in Table 1 to the actual $Var(\hat{\theta}_{NR-TP})$ for all possible allocations of (b, k). Without loss of generality for making comparisons, we again assume that $n_0 = n_1 = 30, \sigma = 1, T = b + k = 7$. For b=1 and b=6, the V_{CS}approximation to $Var(\hat{\theta}_{NR-TP})$ using $\rho = \rho_{avg}$ performed well for all four examples never being anticonservative with both values almost exactly equaling each other for CESD and LS_ADL. By contrast, the V_{CS} approximation with $\rho = \rho_{min}$ greatly overestimated $Var(\hat{\theta}_{NR-TP})$ when b=1 and 6 for CESD and LS_ADL. However, for b ranging from 2 to 5, the V_{CS} approximation using $\rho = \rho_{avg}$ often greatly underestimated $Var(\hat{\theta}_{NR-TP})$ and for CESD and LS_ADL even the V_{CS} approximation using $\rho = \rho_{min}$

often underestimated $Var(\hat{\theta}_{NR-TP})$ which would result in anticonservative power calculations.

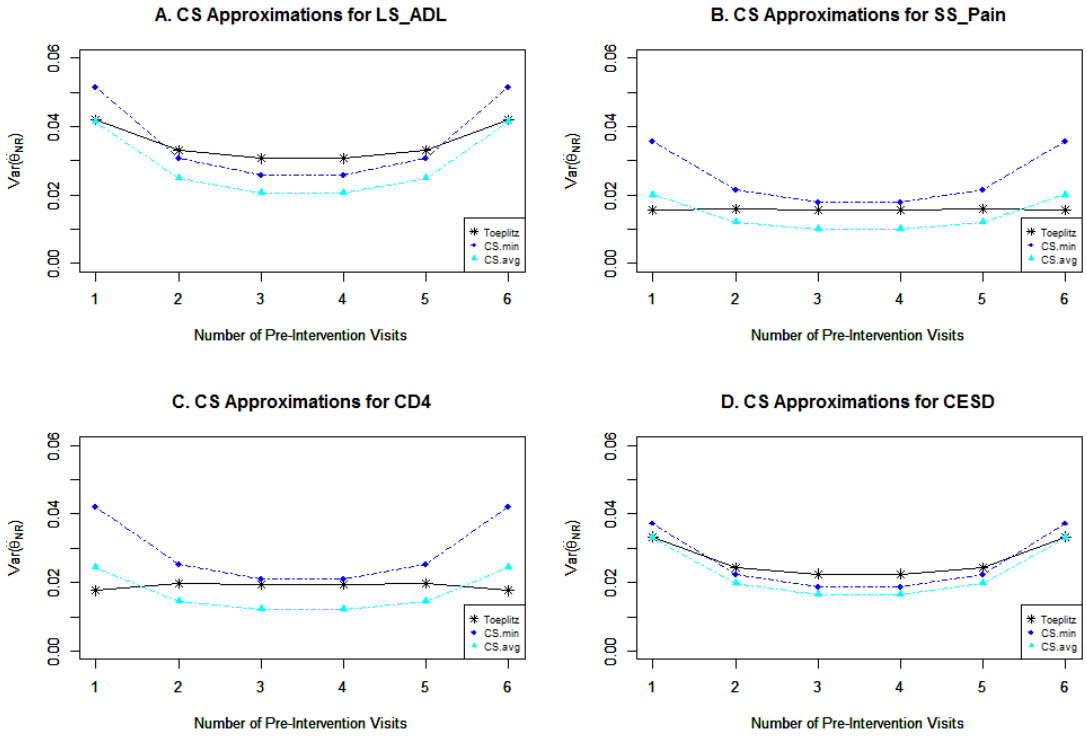

Figure 4: Variance approximations using CS ($\boldsymbol{\rho = \rho_{min}}$ and $\boldsymbol{\rho = \rho_{avg}}$) compared to Toeplitz over all *b: k* allocations ($\boldsymbol{n_0 = n_1 = 30, \sigma = 1, T = b + k = 7}$)

6.2. Two Conservative Approximations to Toeplitz Variance

The fact that using V_{CS} approximations even with $\rho = \rho_{min}$ often underestimated $Var(\hat{\theta}_{NR-TP})$ suggests that simple approaches which conservatively estimate (i.e. not underestimate) $Var(\hat{\theta}_{NR-TP})$ might be better. Perhaps the simplest conservative estimate for general *T* is to reduce the study to *T*=2 and (*b*, *k*) = (1, 1) with only one off-diagonal correlation, the correlation structure is by default V_{CS} with $\rho = \rho_1$ in (6) and (8). Clearly

restricting the study to T=2 measures with one pre- and one post-intervention with this ρ should yield smaller variance than using all T ($T \geq 2$) timepoints; and ii) one would also expect that for all $T \geq 2$ the maximum off diagonal correlation (i.e.$\rho_{max} = \rho_1$).

However, restricting to only 2 of T (i.e. T=7) measures at first glance seems overly conservative which motivates need for another lower bound. Frison & Pocock [8] proposed to use mean summary statistics ($\bar{Y}_{hi.}^{post} = \frac{1}{k}\sum_{j=1}^{k} y_{hij}$ and $\bar{Y}_{hi.}^{pre} = \frac{1}{b}\sum_{j=-b}^{-1} y_{hij}$) to analyze repeated measurements in randomized trials with two intervention arms. Using the same idea for a non-randomized study for each unit, the summary statistic is the mean change: $\bar{Y}_{hi.}^{post} - \bar{Y}_{hi.}^{pre}$. Then the overall intervention difference in these mean changes is: $\frac{1}{n_0}\sum_{i=1}^{n_0}(\bar{Y}_{0i.}^{post} - \bar{Y}_{0i.}^{pre}) - \frac{1}{n_1}\sum_{i=1}^{n_1}(\bar{Y}_{1i.}^{post} - \bar{Y}_{1i.}^{pre}) = (\bar{Y}_{0..}^{post} - \bar{Y}_{0..}^{pre}) - (\bar{Y}_{1..}^{post} - \bar{Y}_{1..}^{pre})$, which has expected value $(\bar{\mu}_0^{post} - \bar{\mu}_0^{pre}) - (\bar{\mu}_1^{post} - \bar{\mu}_1^{pre})$ and variance:

$$Var(\hat{\theta}_{NR-MS}) = \text{Var}[(\bar{Y}_{0..}^{post} - \bar{Y}_{0..}^{pre}) - (\bar{Y}_{1..}^{post} - \bar{Y}_{1..}^{pre})]$$

$$= (\frac{1}{n_0} + \frac{1}{n_1})[\frac{1+(k-1)\rho_{post}}{k} + \frac{1+(b-1)\rho_{pre}}{b} - 2\rho_{cross}]\sigma^2 \qquad (12)$$

Where i) $\rho_{pre}$is the averaged correlation among the pre-intervention timepoints and if the correlation structure is V_{TP} then $\rho_{pre} = \frac{(b-1)\rho_1+(b-2)\rho_2+\cdots+\rho_{b-1}}{(b-1)+(b-2)+\cdots+1}$; ii) ρ_{post} is the averaged correlation among the post-intervention timepoints and if the correlation structure is V_{TP} then $\rho_{post} = \frac{(k-1)\rho_1+(k-2)\rho_2+\cdots+\rho_{k-1}}{(k-1)+(k-2)+\cdots+1}$ and iii) ρ_{cross} is the averaged correlation between the pre- and post-intervention time points and if the correlation structure is V_{TP} then $\rho_{cross} = \frac{\sum_{i=1}^{k}(\rho_i+\rho_{i+1}+\ldots+\rho_{i+b-1})}{bk} = \frac{\sum_{i=1}^{k}\sum_{j=0}^{b-1}\rho_{i+j}}{bk}$.

The GLS estimator is a best linear unbiased estimator (BLUE) [8, 9, 13] and the mean change above is unbiased for $\hat{\theta}$. We can thus conclude from the Gauss-Markov theorem that GLS variance estimate in (4) based on $V_0 = V_{TP}$ is no greater than the MS variance estimate in (12). We thus employ the mean summary statistics approach to derive an upper bound for the GLS variance of a given Toeplitz correlation.

Table 2 in the last 4 columns also presents the values described in this section for the examples in Table 1 (LS_ADL, SS_Pain, CD4 and CESD) based on their given empirical Toeplitz correlations in Table1 for *T*=7 with (*b, k*) = (3, 4): i) Using PCS parameters (ρ_{pre}, ρ_{post} and ρ_{cross}) for *T*=7 with (*b, k*) = (3, 4).

For example, looking at LS_ADL based on Table 2, incorporating $\rho_{max} = \rho_1 = 0.59$ with (*b, k*) = (1, 1) into (6) gives and approximated variance of $\left(\frac{1}{n_0} + \frac{1}{n_1}\right) * \frac{2*(1-0.59)}{1*1}\sigma^2 = 0.82 * \left(\frac{1}{n_0} + \frac{1}{n_1}\right)\sigma^2$ as an upper bound for $Var\left(\hat{\theta}_{NR-TP}\right)$ not only for (*b, k*) = (3, 4), but also for all values of ($b \geq 1, k \geq 1$). Similarly, again for LS_ADL but now specifically for (*b, k*) = (3, 4) incorporating $\rho_{pre} = 0.55$, $\rho_{post} = 0.52$and $\rho_{cross} = 0.42$ from Table 2 into (12) gives $\left(\frac{1}{n_0} + \frac{1}{n_1}\right)\left[\frac{1+(4-1)*0.52}{4} + \frac{1+(3-1)*0.55}{3} - 2 * 0.42\right]\sigma^2 = 0.50 * \left(\frac{1}{n_0} + \frac{1}{n_1}\right)\sigma^2$ as an upper bound for $Var\left(\hat{\theta}_{NR-TP}\right)$ for LS_ADL when *T*=7 and (*b, k*) = (3, 4) and also symmetrically for (*b, k*) = (4, 3).

In Figure 5, for LS_ADL, SS_Pain, CD4 and CESD in the non-randomized designs among all possible values of (*b, k*) (i.e. satisfying *T*=*b*+*k* =7), we present the actual $Var\left(\hat{\theta}_{NR-TP}\right)$ compared to the upper bounds for this produced by i) *T*=2 with (*b, k*) = (1, 1) and $\rho_{max} = \rho_1$in (6) and ii) MS approximations (with ρ_{pre}, ρ_{post} and ρ_{cross}) in (12). As before, we set $\sigma = 1, n_0 = n_1 = 30$.

For SS_Pain and CD4 where ρ_1 were very high (i.e. ~0.85 in Table 1 and Figure 2) followed by rapid drop to $\rho_2, \ldots, \rho_{b+k-1}$, the MS approximation greatly overestimated the true variance as shown in Figures 5-B and 5-C. However, for these outcomes surprisingly the true $Var(\hat{\theta}_{NR-TP})$ were only barely smaller than the simple approximation from restricting to T=2, (b, k) = (1, 1) with $\rho_{max} = \rho_1$.

For LS_ADL and CESD where $\rho_1$were lower (i.e. ~0.60 in Table 1 and Figure 2) and the drop to $\rho_2, \ldots, \rho_{b+k-1}$were smaller (especially for CESD), the MS upper bounds were very close to $Var(\hat{\theta}_{NR-TP})$ at all values of b. However, for these outcomes as shown in Figures 5-A and 5-D, restricting to two timepoints with (b, k) = (1, 1) using ρ_1 greatly overestimated $Var(\hat{\theta}_{NR-TP})$.

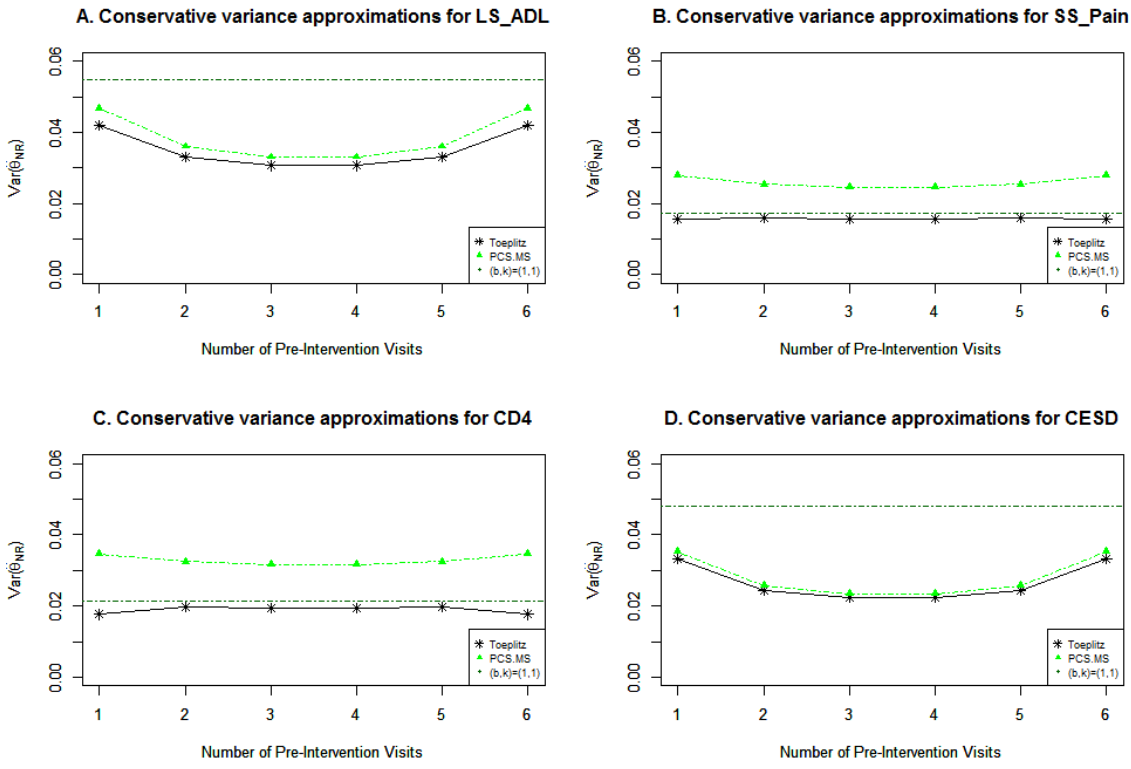

Figure 5: Conservative variance approximations compared to Toeplitz over all *b: k* allocations ($n_0 = n_1 = 30, \sigma = 1, T = b + k = 7$)

7. Concluding Remarks

This chapter developed estimation tools for variance of the intervention effect estimate in non-randomized difference-in-differences studies based on a Generalized Least Squares (GLS) framework. We first used compound symmetry structure for the within-unit correlation and derived closed form GLS variance estimate formulas. The closed form formulas in (6) and (8) are easily calculated and implemented in (5) and elsewhere for power / sample size estimation. We explored the optimal allocation into pre- and post-intervention (*b: k*) under CS when (*T=b+k*) is constrained but the investigators can choose *b* and *k*. Not surprisingly, equal (or closest to equal) number of pre- and post-intervention measurements *(b=k* for *T* even and *b=k±1* for *T* odd) minimized $Var(\hat{\theta}_{NR-CS})$. We then quantitatively compared randomized to non-randomized DD studies in terms of $Var(\hat{\theta}_{R-CS})/Var(\hat{\theta}_{NR-CS})$. Although $Var(\hat{\theta}_{R-CS}) < Var(\hat{\theta}_{NR-CS})$ for the same design parameters (*b*, *k* and ρ), non-randomization can work nearly as well in compound symmetry settings if within-unit correlation ρ is high and/or $b \geq k$.

However, in our real data examples using outcomes from long term care facilities and HIV patients, the correlation structures were (sometimes very) different from compound symmetry suggesting further investigation on power estimation with the more general Toeplitz correlation was needed. As simple closed form formulas for $Var(\hat{\theta}_{NR-CS})$ were

not feasible, we presented the GLS variance estimates from computer calculation using the empirical Toeplitz correlation structures of these examples. The same optimal allocation of *b* and *k* under CS, ($b \approx k$) held for the Toeplitz for only two of the examples where *T*=7. Again while R-DD designs yielded lower variance than NR-DD designs, the advantage of randomization was less when correlations were larger and as *b* increased.

In practice, investigators often neither have precise normative data on the Toeplitz variance parameters ($\rho_1, \ldots, \rho_{T-1}$) of repeated measures nor the software/expertise to derive variances from this Toeplitz structure. We thus investigated approximations to $Var\left(\hat{\theta}_{NR-TP}\right)$ by closed form formula variances using CS for $T=b+k=7$. The CS approximations to $Var\left(\hat{\theta}_{NR-TP}\right)$using the average correlation (i.e. V_{CS} using $\rho = \rho_{avg}$) performed well when (*b*, *k*) = (1, 6) and (*b*, *k*) = (6, 1). But for other values of (*b*, *k*), V_{CS} using $\rho = \rho_{avg}$ sometimes substantially underestimated $Var\left(\hat{\theta}_{NR-TP}\right)$ and thus overestimated power. Even CS approximations to the Toeplitz structure that seemed conservative (i.e. $V_{CS}$with $\rho = \rho_{min}$) resulted in substantial power overestimation. Thus while investigators might be tempted to do so as it is easy, misuse of CS with $\rho = \rho_{avg}$ causes even greater power overestimation.

We then studied approaches to derive upper bounds to $Var\left(\hat{\theta}_{NR-TP}\right)$ for our four examples and had some surprising results. In two examples where correlation was high at ρ_1 and dropped off rapidly (SS_Pain and CD4), it turned out that just using simple approximation with $T = 2$ total time points with one pre- and one post-intervention measurement resulted in a variance for the intervention effect that was only slightly larger than $Var\left(\hat{\theta}_{NR-TP}\right)$ with all *T*=7 timepoints. This is, of course premised on our

assumption that the intervention effect is an immediate jump and might not be the case if the intervention effect was cumulatively increasing over time. However, for at least one of the outcomes we studied, interventions to increase CD4 count in fact often do have close to a short term jump effect that is mostly fully manifested by three months [20, 21].

The MS approximation provided an alternative upper bound for $Var(\hat{\theta}_{NR-TP})$ (equivalently lower bound for power). Note that for the same two examples with high ρ_1 that dropped off rapidly (SS_Pain and CD4), the MS upper bound for variance greatly overestimated $Var(\hat{\theta}_{NR-TP})$. However, for the other two examples where ρ_1 was smaller and the drop-off between ρ_1 and $\rho_6$was smaller (LS_ADL and CESD) the MS approximation upper bounds only slightly overestimated $Var(\hat{\theta}_{NR-TP})$ and restriction to $T = 2$ total time points with $(b, k) = (1, 1)$, resulted in a large increase in variance. Thus it does not appear to be a simple way to derive upper bounds for $Var(\hat{\theta}_{NR-TP})$ that works in all settings.

There are some limitations in our work. We assumed an immediate jump effect of the intervention but in some settings the effect may be linear cumulative or some other pattern. The illustrative examples we used are limited with a fixed total visits ($T = 7$). While more comprehensive analyses for other values of T and other correlation structures is beyond the scope of this chapter, we believe that the correlation structures presented here are likely generalizable and that $T \approx 7$ may be reasonable for many settings. Although we assumed static covariance (a minimum requisite to use historical data for correlation estimation), covariance could change over time from uncontrollable mechanisms in practice. Non-randomized designs could lead to potential regression to the mean biases (Appendix 1) if units were deliberately chosen to receive (or to not

receive) the interventions based on poor (or good) baseline performance over the *b* pre-intervention visits. So investigators must ensure that this does not happen in the intervention arm allocation. The above limitations lead to complicated settings whose statistical properties perhaps can only be studied with simulation.

In conclusion, we derived closed form GLS formulas for variance of the estimated intervention effect and investigated optimal designs for non-randomized difference in difference studies based on compound symmetry correlation structure for repeated measures within the unit. For DD studies with CS correlation, the penalty from non-randomization (versus randomization) on variance of the estimated intervention effect was lessened by having larger numbers of pre-intervention measures relative to numbed of post-intervention measures and with larger ρ. However, CS may not always hold in the real world as shown in our examples. Our illustrative examples using observed Toeplitz correlations did not always empirically support similar properties as were derived for CS using closed form formulas such as $b \approx k$ minimizes the variance of the estimated intervention effect. Furthermore, in some empirical settings, T=2 measures with (b, k) = (1, 1) may be almost as powerful as having T=7 measures. While it may be difficult for many investigators both to obtain normative data for Toeplitz correlation structure and to compute variances of intervention effect estimates based on Toeplitz variances, our efforts to identify simple and conservative approximations had mixed success.

References

1. Pocock SJ, Hughes MD, Lee RL. Statistical problems in the reporting of clinical trials: a survey of three major medical journals. *New England Journal of Medicine* 1987; 317, 426-432

2. Abadie A. Semiparametric difference-in-differences estimators. *Review of Economic Studies* 2005; 72 (1): 1–19.

3. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: The difference-in-differences approach. *JAMA guide to statistics and methods* 2014; Volume 312, Number 22.

4. Bertrand M, Duflo E, Mullainathan S. How much should we trust Differences-in-Differences estimates? *Quarterly Journal of Economics* 2004; 119 (1): 249–275.

5. Galton F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* 1886; Vol. 15: 246-263

6. Stigler SM. Regression towards the mean, historically considered. *Statist Meth Med Res* 1997; 6:103–14

7. Barnett AG, Van Der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol* 2005; 34:215-20.

8. Frison L, Pocock SJ. Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Stat Med* 1992; 11: 1685–704.

9. Aitken AC. On Least-squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh* 1934; 55: 42–48.

10. Galecki AT. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics, Theory and Methods* 1994: 23: 3105-3120.

11. Litttell RC, Pendergast J, Natarajam R. Tutorial in Biostatistics: modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* 2000; 19: 1793-1819.

12. Wolfinger RD. Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics* 1996; Vol. 1, No. 2, 205-230.

13. Van Breukelen GJP. ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *J Clin Epidemiol* 2006; 59: 920–25.

14. Fisher RA. Applications of "Student's" distribution. *Metron* 1925; 5: 90–104.

15. Satterthwaite FE. Synthesis of Variance. *Psychometrika* 1941; 6, 309-316.

16. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; Vol. 53, No. 3: 983-997.

17. Centers for Medicare and Medicaid Services Five Star Quality Rating System, https://www.cms.gov/medicare/provider-enrollment-and-certification/certificationandcomplianc /fsqrs.html

18. Barkan SE, Melnick SL, Preston-Martin S., Weber K., Kalish L. A., Miotti P., Young M., Greenblatt R., Sacks H., Feldman J. The Women's Interagency HIV Study. WIHS Collaborative Study Group. *Epidemiology* 1998; 9(1), 117-25.

19. Radloff L. The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement 1977; 1, 385-401.

20. Pakker NG, Notermans DW, de Boer RJ, Roos MT, de Wolf F, Hill A, Leonard JM, Danner SA, Miedema F, Schellekens PT. Biphasic kinetics of peripheral blood T cells after triple combination therapy in HIV-1 infection: a composite of redistribution and proliferation. *Nat Med* 1998 Feb;4(2):208-14.

21. de la Rosa R, Leal M. Thymic involvement in recovery of immunity among HIV-infected adults on highly active antiretroviral therapy. *Journal of Antimicrobial Chemotherapy* 2003; 52, 155–158.

Appendix 1: Gradient Non-Randomized Designs

There may be ethical pressure to use a process that we define here as "gradient randomization" to give treatment to the most-needy units. In other words, units are selected to be given the new intervention based on the levels of the outcome at time *j=-1*, most likely under a gradient that those with worse levels at that time being more likely to be chosen for the intervention. Unfortunately, a "regression to the mean (RTM)" phenomenon [5, 6] is likely to occur in such gradient non-randomized designs. In general, high (or low) measurements in a longitudinal process are likely to be followed by less extreme ones that are closer to the unit's true mean at subsequent times. Thus many of those "poor-outcome" sites that were given preference for the intervention by performing badly at *j=-1*, are likely to regress back (i.e. improve) on their own even without the intervention.

The practical problem of RTM is to distinguish the intervention change from the expected change due to the natural variation [21]. This can, perhaps, be modeled with both long-term (α_1) and short-term (α_2) components for baseline differences between facilities that are and are not randomized to receive interventions shown in (A). The long-term pre-intervention gradient $\alpha_1$captures overall treatment arm differences at *j=–b, -(b-1), …, -1* coming from the fact that facilities that perform worse in general are more likely to perform worse at *j=-1*. The short-term immediate pre-intervention gradient (α_2) captures the selection effect from those facilities that have a directional shift at *j=-1* being selected on this basis into the intervention arm.

$$Y_{hij} = \alpha_0 + \alpha_1 I_{\{h=1\}} + \alpha_2 I_{\{h=1, j=-1\}} + \beta_j + \theta Z_{hj} + \varepsilon^*_{ij} \qquad \text{(A)}$$

Again β_j denotes main effects of times *j = -(b-1), …, -1, 1, …,k* relative to *j=-b*; as before α_0 denotes the average baseline value taken at *j=-b* for control units; $\alpha_1$denotes any long-term gradient effect of *h=1* versus *h=0* that lead to selection into the intervention arm; α_2 captures any short-term gradient effect of *h=1* versus *h=0* that lead to selection for intervention at *j=-1*; θ denotes a constant relative post-intervention change from treatment *h=1* (as opposed to *h=0*) after *j=0*.

We are assuming that the short-term intervention selection gradient manifests only at *j=-1*. If so, and we exclude the time point *j=-1*, from the study, then *b* becomes *b-1*, α_2 drops out of (A), and the model becomes the same as (1) in the main chapter with parameters (*b-1, k*) under CS. But note if CS does not hold then the setting is more complicated, which is beyond the scope of this chapter. Also, note that it could be argued that the error term at *j=-1* is altered by the gradient selection bias which manifests in part through subsuming the error term into α_1. However, even if so, the approach described above to exclude the timepoint *j*=-1from the analysis should be valid.

Appendix 2: Design Matrix

For (1) with the general parameter vector $\underline{\beta}=(\alpha_0, \alpha_1, \beta_{-(b-1)}, \dots, \beta_{-1}, \beta_1, \dots, \beta_k, \theta)$, the corresponding design matrix has columns $(I, I_{\{h=1\}}, J_{-(b-1)}, \dots, J_{-1}, J_1, \dots, J_k, Z)$. To simplify the calculation in $(X'V^{-1}X)^{-1}$ in Appendix 3, we reorder the parameter vector to put $\alpha_1$and θ together with $X = (I, J_{-(b-1)}, \dots, J_{-1}, J_1, \dots, J_k, I_{\{h=1\}}, Z)$.

As shown in (B), the general design matrix *X* is made up of $(n_0 + n_1)$ times $X_{h,i}$'s, where$X_{h=0,i}$denotes the partial design matrix for each unit in the untreated group

and $X_{h=1,i}$ stands for each unit in the treated group. Note the $(T+1)^{th}$column indicates a long-term pre-intervention gradient corresponding to $\alpha_1$due to non-randomization, and the $(T+2)^{th}$column stands for intervention effect θ. Therefore, for the NR-DD, the design matrix is $X = X_{NR}$.

$$X_{NR} = \begin{bmatrix} X_{h=0,1} \\ \vdots \\ X_{h=0,n_0} \\ X_{h=1,1} \\ \vdots \\ X_{h=1,n_0} \end{bmatrix} \text{where}$$

$$X_{h=0,i} = \begin{bmatrix} 1 & 1 & \ldots & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & \cdots & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & -1 & \cdots & -1 & -1 & \cdots & -1 & 0 & 0 \end{bmatrix}_{T*(T+2)} ;$$

$$X_{h=1,i} = \begin{bmatrix} 1 & 1 & \ldots & 0 & 0 & \cdots & 0 & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 & 0 & \cdots & 0 & 1 & 0 \\ 1 & 0 & \cdots & 0 & 1 & \cdots & 0 & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & -1 & \cdots & -1 & -1 & \cdots & -1 & 1 & 1 \end{bmatrix}_{T*(T+2)} . \quad \text{(B)}$$

Now for the randomized design the design matrix X_R is the same as X_{NR} shown above except that the second to last column corresponding to $\alpha_1$is removed.

Appendix 3: GLS Variance Estimate

The goal is to find $(X'V^{-1}X)^{-1}$as the lower right element of $(X'V^{-1}X)^{-1}\sigma^2$ is $Var(\hat{\theta})$ where *X* is the design matrix described in Appendix 2. First under CS where $\rho_{jj'} \equiv \rho$, the covariance matrix in (2) reduces to $V_{CS} = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}_T$ and $V^{-1} =$

$$\begin{pmatrix} {V_0}^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & {V_0}^{-1} \end{pmatrix}_{(n_0+n_1)T} \quad \text{with}$$

$${V_0}^{-1} = \frac{1}{[1+(T-1)\rho](1-\rho)} \begin{pmatrix} 1+(T-2)\rho & \cdots & -\rho \\ \vdots & \ddots & \vdots \\ -\rho & \cdots & 1+(T-2)\rho \end{pmatrix}_T .$$

Then we apply the technique for the inverse of the partitioned matrix.

$$(X'V^{-1}X)^{-1} = \begin{bmatrix} A_{11} & A_{21} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} B_{11} & B_{21} \\ B_{21} & B_{22} \end{bmatrix}$$

where $B_{22} = (A_{22} - A_{21}{A_{11}}^{-1}A_{12})^{-1}$and $Var(\hat{\theta})$is contained in B_{22}. We then derive this simple closed form formula for GLS-CS estimate of variance.

Here we take NR design as an example. For NR-DD,

$$\left(X_{NR}^{'}V^{-1}X_{NR}\right)^{-1} = \left(\frac{1}{n_0}+\frac{1}{n_1}\right)[1+(T-1)\rho](1-\rho)$$

$$\begin{bmatrix} 2\mathrm{T}(1-\rho) & 2(1-\rho) & \cdots & 2(1-\rho) & 0 & 0 \\ 2(1-\rho) & 2[1+(T-2)\rho] & \cdots & -2\rho & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 2(1-\rho) & -2\rho & \cdots & 2[1+(T-2)\rho] & 0 & 0 \\ & & & & \frac{\mathrm{T}(1-\rho)}{2} & \frac{1-\rho}{2} \\ 0 & 0 & \cdots & 0 & & \\ 0 & 0 & \cdots & 0 & \frac{1-\rho}{2} & \frac{[1+(T-2)\rho]}{2} \end{bmatrix}^{-1}$$

$$= \left(\tfrac{1}{n_0}+\tfrac{1}{n_1}\right)[1+(T-1)\rho](1-\rho)\begin{bmatrix} A_{11} & A_{21} \\ A_{21} & A_{22} \end{bmatrix}^{-1}.$$

In this partitioned matrix,

$$A_{11} = \begin{bmatrix} 2\mathrm{T}(1-\rho) & 2(1-\rho) & \cdots & 2(1-\rho) \\ 2(1-\rho) & 2[1+(T-2)\rho] & \cdots & -2\rho \\ \vdots & \vdots & \ddots & \vdots \\ 2(1-\rho) & -2\rho & \cdots & 2[1+(T-2)\rho] \end{bmatrix} =$$

$\begin{bmatrix} 2\mathrm{T}(1-\rho) & 2(1-\rho) & \cdots & 2(1-\rho) \\ 2(1-\rho) & & & \\ \vdots & & C_{11} & \\ 2(1-\rho) & & & \end{bmatrix}$ where C_{11} is compound symmetry with $2[1+$ $(\mathrm{T}-2)\rho]$ on diagonal and -2ρ off diagonal and the same meaning for this same pattern holds when it occurs in the other matrices presented in Appendix 3;

$$A_{22} = \begin{bmatrix} \frac{\mathrm{T}(1-\rho)}{2} & \frac{1-\rho}{2} \\ \frac{1-\rho}{2} & \frac{[1+(T-2)\rho]}{2} \end{bmatrix} \text{ and } A_{21} = A_{12}^{'} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}.$$

Because $A_{21}{A_{11}}^{-1}A_{12} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$,

$$B_{22} = (A_{22} - A_{21}{A_{11}}^{-1}A_{12})^{-1} = {A_{22}}^{-1} = \begin{bmatrix} \frac{\mathrm{T}(1-\rho)}{2} & \frac{1-\rho}{2} \\ \frac{1-\rho}{2} & \frac{[1+(T-2)\rho]}{2} \end{bmatrix}^{-1}.$$

Upon inverting A_{22}, the lower right element in B_{22} implies the the GLS variance estimate for non-randomized designs.

$$Var(\hat{\theta}_{NR-CS}) = (\frac{1}{n_0} + \frac{1}{n_1})\frac{\mathrm{T}(1-\rho)}{\mathrm{bk}}\sigma^2.$$

Similarly, to calculate $\left(X_R^{'}V^{-1}X_R\right)^{-1}$ using the partitioned matrix inverse approach described above, the GLS variance estimate for randomized designs is

$$Var(\hat{\theta}_{R-CS}) = (\frac{1}{n_0} + \frac{1}{n_1})\frac{[1+(T-1)\rho](1-\rho)}{k[1+(b-1)\rho]}\sigma^2.$$

Appendix 4: Variance is Invariant to Absorption from Non-Randomization Dispersion under Compound Symmetry

For the NR-DD model as defined in (1), the strata effects ($\alpha_0 + \alpha_1 I_{\{h=1\}}$) are treated as fixed, i.e., α_0 for the control arm and $\alpha_0 + \alpha_1$ for the intervention arm. Using compound symmetry, ε_{ij}^* can be decomposed as ($\mu_i + \varepsilon_{ij}$) in (B) with μ_i being a main error for the unit *i*, and ε_{ij} is independent random error at each time *j* within-unit *i*.

$$Y_{hij} = \alpha_0 + \alpha_1 I_{\{h=1\}} + (\mu_i - \alpha_0 - \alpha_1 I_{\{h=1\}}) + \beta_j + \theta Z_{hj} + \varepsilon_{ij} \quad \text{(B)}$$

However, had randomization been used in this setting, then there would be a common intercept α with the model being (C).

$$Y_{hij} = \alpha + (\mu_i - \alpha) + \beta_j + \theta Z_{hj} + \varepsilon_{ij} \quad \text{(C)}$$

where for example with $n_0 = n_1$, $\alpha = \alpha_0 + \frac{\alpha_1}{2}$. In the R setting,

$Var(\mu_i) = \tau^2, Var(\varepsilon_{ij}) = \sigma_e^2$ with $\tau^2 + \sigma_e^2 = \sigma^2$, and $\rho = \frac{\tau^2}{\tau^2 + \sigma_e^2}$.

But for the NR-DD model imposed in the same setting, the strata effects are treated as fixed. The within person common error now is ($\mu_i - \alpha_0 - \alpha_1 I_{\{h=1\}}$) with the variance ($\tau^2 - \sigma_h^2$) where $\sigma_h^2 = \frac{\alpha_1^2}{4}$. The $Var(\varepsilon_{ij}) = \sigma_e^2$ is unchanged by the NR design. For the NR-DD model applied to this setting, the overall variance of an observation is $\sigma_{NR}^2 = (\tau^2 - \sigma_h^2) + \sigma_e^2$ and the within person repeated measure correlation is $\rho_{NR} = \frac{\tau^2 - \sigma_h^2}{\sigma_{NR}^2} = \frac{\tau^2 - \sigma_h^2}{(\tau^2 - \sigma_h^2) + \sigma_e^2}$ thus $(1 - \rho_{NR}) = \frac{\sigma_e^2}{(\tau^2 - \sigma_h^2) + \sigma_e^2}$.

However, from (6) the $Var(\hat{\theta}_{NR-CS})$ in NR-DD only depends on $\sigma^2$and ρ through the product$(1 - \rho)\sigma^2$. To that end, this product is unchanged by application of the NR-DD design in that$(1 - \rho_{NR})\sigma_{NR}^2 = (1 - \rho_R)\sigma_R^2 = \sigma_e^2$. Thus the parameters for σ^2 and

ρ from the "pre-nonrandomized" study design population can be used in (6) no matter what the impact of the NR-DD on the final $\sigma^2$and ρ are.

We should note that a similar decomposition for the general Toeplitz covariance matrix is too complicated to present at this stage and may not even have a well-defined formulation. In our comparisons of Randomized and Non-randomized designs in terms of $Var(\hat{\theta}_{NR-TP})$ and $Var(\hat{\theta}_{R-TP})$, we assumed that any impact of imposing a DD model on the variances and underlying correlations of repeated measures in the NR (versus the R design) will roughly cancel out in the determination of $Var(\hat{\theta}_{NR-TP})$. But this remains to be verified empirically.

Chapter 3 Non-Randomized and Randomized Stepped-Wedge Designs using an Orthogonalized Least Squares Framework

Abstract

Randomized stepped-wedge (R-SW) designs are increasingly used to evaluate interventions targeting continuous longitudinal outcomes measured at T fixed time points. Typically, all units start out untreated, and randomly chosen units switch to intervention at sequential time points until all receive intervention. As randomization is not always feasible, non-randomized stepped-wedge (NR-SW) designs (units switching to intervention are not randomly chosen) have attracted researchers. We develop an orthogonlized generalized least squares framework for both R-SW and NR-SW designs. The variance of the intervention effect estimate depends on the number of steps (S), length of step sizes (t_s) and number of units (n_s) switched at each step (s=1, …, S). If all other design parameters are equal, this variance is higher for the NR-SW than for the equivalent R-SW design (particularly if the intercepts of non-randomly stepped switching strata are analyzed as fixed effects). We focus on balanced SW (BR-SW, BNR-SW) designs (where t_s and n_s remain constant across s) to obtain insights into optimality for variance of the estimated intervention effect. As previously observed for the BR-SW, the optimal choice for number of time points at each step is also $t_s \equiv 1$ for the BNR-SW. In our examples, when compared to BR-SW designs, equivalent BNR-SW designs even with intercepts of non-randomly stepped switching strata analyzed using fixed effects sacrifice little efficiency given an intra-unit repeated measure correlation $\rho \geq 0.50$. Compared to traditional difference-in-differences designs, optimal BNR-SW designs are more efficient

with the ratio of variances of these designs converging to 0.75 when T>10. We illustrate these findings using longitudinal outcomes in long-term care facilities.

1. Introduction

Recently developed randomized stepped wedge (SW) designs [1, 2] are applied to longitudinal outcomes repeatedly measured at *T* fixed time points in *N* units being placed on the new intervention over time using a staggered schedule. For the examples used in this chapter, a unit is a single medical facility undergoing a facility-wide intervention with facility-level measurements taken over time. Although we do not have the person-level data within these units, we extend to such designs in Appendix 1 where units could in fact be individual persons undergoing person-level interventions against chronic conditions being measured over time. Typically, at the first time point all units are not on the new intervention (i.e. untreated). More units (who then remain on the intervention until end of study) are switched onto the new intervention (i.e. treated) at subsequent time points. A pooled comparison of the study outcome for "treated" versus "untreated" unit-measures that adjusts for secular time effect is made. The SW designs are increasingly implemented in diverse areas including: cardiovascular disease [3, 4], cancer [5, 6], HIV [7], respiratory disease [8], nutrition [9], maternal and child health [10], and health care financing [11].

Shifting of units onto intervention at different times complicates SW analysis in that we need to account for secular time effects and intra-unit correlation [1, 12]. Often the randomized stepped wedge (R-SW) is applied to "cluster randomized trials" where the "unit" is a cluster of a fixed $m>1$ individuals with a shared intercept measured at fixed time points [1, 2, 12, 13]. The analysis is based on the means of all individuals in the cluster at each time. However, the R-SW is also used for cases where there is only 1 measure at each time point (i.e. $m=1$) sometimes referred to as a wait list design [13].

This is also the case for the examples in our analysis so our derivations are for $m=1$ as is typically the case in facility-level analyses for health care settings [14, 15]. Appendix 1 gives a conversion between our setting and cluster-randomized trials with $m>1$. Hussey and Hughes [12] provided approaches to sample size and power calculations for R-SW implementation of cluster-randomized trials. The R-SW often has greater statistical power than traditional designs including: randomized longitudinal parallel designs [16] where randomly chosen participants treated at all time points are compared to those untreated at all time points and randomized longitudinal difference in difference (R-DD) designs [17] where all units start out untreated and are measured for a fixed number of time points, then at the same fixed time point a randomly chosen subset of units are switched to the intervention and remain switched until the end.

While randomization of units to intervention arms is preferred as a gold standard to reduce bias and improve efficiency [18], it is not always feasible, particularly in health service settings, because of resource and logistical constraints [19]. Thus, for example, non-randomized difference-in-differences (NR-DD) studies are applied to estimate impact of new interventions or policies [20]. In NR-DD designs, all units start out untreated and are followed for a fixed number of visits; then non-randomly chosen units are switched to the new policy or intervention and compared to controls continuing to be untreated. However, non-randomized stepped-wedge (NR-SW) designs have recently attracted researchers [21, 22]. While it is known that generalized linear mixed models [12, 23] can evaluate intervention effects on continuous outcomes under normal approximation, these have not been formally applied to NR-SW design including for power and sample size estimation.

This chapter develops a unifying orthogonalized framework for stepped-wedge designs and obtains simple formulas for variance of intervention effects on continuous outcomes. Section 2 develops a general linear model for both R-SW and NR-SW designs and proposes an orthogonalized design matrix to simplify derivation of variance estimates. Section 3 presents general least squares (GLS) estimates for variance of intervention effect using a within-unit compound symmetry repeated measure correlation and discusses the general framework for power estimation. Section 4 focuses on a special but common case of stepped wedge designs where equal numbers of units are switched at equally spaced times that we denote as balanced designs (BR-SW for randomized and BNR-SW for non-randomized) and derives simple closed form solutions for variance of the intervention effect estimate for these designs. Variances of optimal designs for BR-SW and BNR-SW studies are compared. Section 5 compares the optimal BNR-SW design to the optimal NR-DD studies. Section 6 presents illustrative examples from long-term care facilities and Section 7 summarizes and discusses possible future work.

2. Stepped-Wedge Models

2.1. Notations

Let T be the number of measured time points, S be the number of these time points at which one or more units is transitioned onto intervention (i.e. "steps") and N be the total number of units. Typically, at the first step, all of the units start in the control (untreated) condition and baseline measurements are taken, although we expand this to allow some units treated at baseline. Once switched onto the intervention, units remain treated until the end. At the last step, often all units have switched to the intervention, but we expand

to allow some units to remain untreated. Let *s = 1, …, S* enumerate the ordered stratum (of units) that is switched to intervention per step. Here a stratum is a group of units that share a common characteristic, i.e., the time when the intervention is first delivered. Let A_s denote the s^{th} "shifting strata" (the subset of units that switched to the intervention at step *s*) and n_s be the number of units in A_s.

Let *i = 1, …, N* enumerate the units with the enumeration ordered by the stratum. For instance, A_1contains $\{1, 2, \dots, n_1\}$, i.e., stratum *s=1* with units $i \in A_1$ switched to the intervention at the first step; A_2 contains $\{n_1 + 1, n_1 + 2, \dots, n_1 + n_2\}$, i.e., stratum *s=2* with units $i \in A_2$ switched to the intervention at the second step and so on. The number of consecutive time periods or step size per step is denoted $(t_0, t_1, \dots, t_S)$ with $\sum_{s=0}^{S} t_s = T$, but note that $t_0 = 0$ if some (i.e. n_1) units are already treated when the study starts. If all units have not been shifted to treatment by the end of the study, $t_S = 0$ and n_S denotes the number of units never shifted onto treatment. Let $j_s = \sum_{l=0}^{s-1} t_l$ denote the first time point that units in the s^{th}stratum are treated (with $j_s = \infty$ if some units are never treated); and Z_{ij} denote if unit *i* is treated at time *j* (0=no, 1=yes) with $Z_{ij} = 0$ if $j < j_s$ and $Z_{ij} = 1$ if $j \geq j_s$ where *s* is the ordered stratum that unit *i* belongs to. Figure 1 illustrates the general SW study where all units start out untreated ($t_0 > 0$), and all units are treated after the last step ($t_S > 0$).

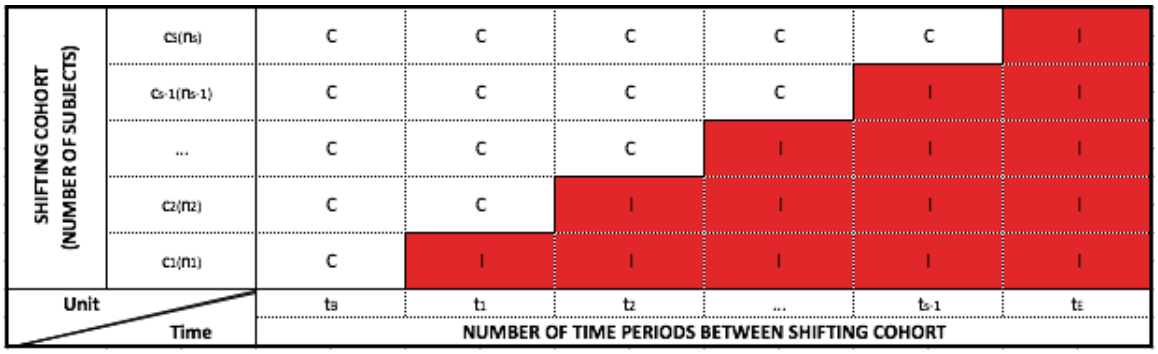

Figure 1: Overview for general Stepped-Wedge designs

We expand the general SW design to three special subcases: 1) Not all units are shifted onto treatment (or $t_S = 0$), even though the last n_S units are not shifted onto treatment, they do constitute a "shifting strata" and step "*S*" can be thought of as "never shifted". Note when $S = 2$, this subcase reduces to a DD design. 2) The study does not begin until after the first n_1 unit had been put on treatment. The fact that these units are never untreated is captured by $t_0 = 0$. 3) Both previous conditions $t_0 = t_S = 0$ happen with some units started on treatment and some units never shifted to treatment. Note this subcase reduces to a parallel design when $S = 2$.

2.2. Statistical model and orthogonal coding for design matrix

Let Y_{ij} be the measurement at the j^{th}time point from unit *i*. Again for this chapter we only have single facility-level measures over time. If there are *m* patient-level measures nested within each facility then $\bar{Y}_{ij}$ can be analyzed using the conversion in Appendix 1.

For any given unit *i* at time *j*, we can model the outcome of interest as $Y_{ij} = \mu_i + \beta_j + \theta Z_{ij} + \varepsilon_{ij}$. Here μ_i is the main effect for unit *i*, β_j is the main effect for time *j*, θ is the effect of the intervention and ε_{ij} is random error. The intervention effect (θ) is

modeled as an "immediate jump" effect and remains constant in the post-intervention measurements. Now $\mu_i \sim N(\alpha_0, \tau^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_e^2)$ with all previous terms being independent. Randomization, also known as random allocation of the units into shifting strata, results in each unit having an equal probability of being assigned to each of the *S* shifting strata. The purpose of randomization is to eliminate allocation bias and achieve shifting strata similar in baseline characteristics [24]. If we subsume the random unit deviation $(\mu_i - \alpha_0)$ into the error term ε_{ij}^* (i.e. $\varepsilon_{ij}^* = (\mu_i - \alpha_0) + \varepsilon_{ij}$), the R-SW model is:

$$Y_{ij} = \alpha_0 + \beta_j + \theta Z_{ij} + \varepsilon_{ij}^* \quad (1)$$

Now $\varepsilon_{ij}^* \sim N(0, \sigma^2)$ where $\sigma^2 = \tau^2 + \sigma_e^2$ and is independent between different units as shown in (6) but has correlation $\rho = \frac{\tau^2}{\tau^2 + \sigma_e^2}$ within two timepoints *j* and *j'* within the same unit *i*.

As discussed in Appendix 1, we should caution that our notation for ρ uses the wait list design or an already averaged unit as a single observation. This differs from that used in most cluster randomized stepped wedge design papers as our response is a single measure (Y_{ij}) as opposed to the average of *m* independent observations for a cluster *i* at time *j* ($\bar{Y}_{ij}$). To convert between the two notations, our $\rho = \frac{\tilde{\rho}}{\tilde{\rho} + \frac{1-\tilde{\rho}}{m}}$ where $\tilde{\rho}$ denotes the "ρ" used in "cluster randomization notation" papers [12, 25-27].

For a non-randomized design with *S* steps (and thus *S* shifting strata), we assume that the non-randomization is associated with the central tendency (mean) of the observations within the strata, but not otherwise with the trajectories. The mean effect is no longer a

common α_0, but differs by $\alpha_s (s = 1, \ldots, S)$ that captures the non-randomization displacement for being in shifting stratum *s*. The shared random error of the unit effect (i.e. $\mu_i - \alpha_s$) is again subsumed into the variance and within-unit covariance of ε_{ij}^* (i.e. $\varepsilon_{ij}^* = (\mu_i - \alpha_s) + \varepsilon_{ij}$). The model being fit for NR-SW is thus:

$$Y_{ij} = \alpha_s I_{\{i \in A_s\}} + \beta_j + \theta Z_{ij} + \varepsilon_{ij}^*. \quad (2)$$

We consider in (2) the strata effects (i.e. $\alpha_s - \alpha_0$) to be "fixed" rather than "random" effects [28]. In many NR-SW settings it may be hard to argue that strata effects are random (for example with respect to strata order and hence number of time points treated), normally distributed, and/or that a random-effects state model is numerically stable [29]. The Hausman Test [28] for this admissibility of random effects models could be used in such settings. Or correlation of α_s with *s* could be examined, and random effects models not be used for non-zero correlation.

However, if the strata effects are considered random in the NR-SW model, then the assumption on the covariance of ε_{ij}^* will be different from that in (6). For the NR-SW with the strata effects considered random, measurements from units in different strata are independent, and the repeated measure correlation within the same unit is ρ, but now the correlation of the error ε_{ij} and $\varepsilon_{i',j}$ from two different units in the same stratum is ρ_s where $0 \leq \rho_s \leq \rho$ as shown in (7).

For both R-SW and NR-SW in (1) and (2), the coding for intervention effect (Z_{ij}) is effectively (0, 1) with 0 for control and 1 for intervention. For t_0 (possibly zero) baseline measures, all units stay in control and the coding is 0. In the "build-up" steps, the coding switches from 0 to 1 sequentially as $j \geq j_s$. Eventually for the last step (unless $t_S = 0$),

every unit receives intervention and the coding is 1. However, to obtain an orthogonalized decomposition of the intervention parameters from the time parameters, we re-parameterize (1) and (2) as below.

For R-SW,

$$Y_{ij} = \alpha_0 + \beta_j^* + \theta Z_{ij}^* + \varepsilon_{ij}^*; \tag{3}$$

For NR-SW,

$$Y_{ij} = \alpha_s I_{\{i \in A_s\}} + \beta_j^* + \theta Z_{ij}^* + \varepsilon_{ij}^*; \tag{4}$$

where $Z_{ij} = 0$ for if $j < j_1$ or if $j \geq j_S$ as all units are in the same treatment condition, otherwise for $j_s \leq j < j_{s+1}$,

$$Z_{ij}^* = \begin{cases} -\frac{\sum_{l=1}^{s} n_l}{N}, & if\ Z_{ij} = 0 \\ \frac{N - \sum_{l=1}^{s} n_l}{N}, & if\ Z_{ij} = 1 \end{cases}. \tag{5}$$

Equivalently, $Z_{ij}^* = Z_{ij} - \frac{\sum_{l=1}^{s} n_l}{N}$ and thus $\beta_j^* = \beta_j + \frac{\sum_{l=1}^{s} n_l}{N}\theta$.

The orthogonal coding of intervention effect (5) for unit-time can be found in Appendix 2. The advantage of the proposed orthogonal coding is it simplifies the solution for the GLS estimates obtained in Section 3.

3. GLS Variance Formula and Power Estimation

3.1. General formula for GLS estimate

The matrix forms of (3) and (4) can be written as: $Y = X\underline{\beta} + \epsilon$, where $\varepsilon \sim N(0, \sigma^2 V)$. Here *X* represents the design matrix and *Y* is a vector of outcomes. For (4) with the

general parameter vector $\underline{\beta}=(\alpha_1, \ldots, \alpha_S, \beta_1^*, \ldots, \beta_{T-1}^*, \theta)$, the corresponding *X* has columns $(I_{\{i\in A_1\}}, \ldots, I_{\{i\in A_S\}}, J_1, \ldots, J_{T-1}, Z^*)$ with *N* rows per column. Z^* is a column with orthogonal coding in (5) (as shown in Appendix 2) and $J_1, \ldots, J_{T-1}$ are columns corresponding to dummy *T-1* independent time coded (-1, 1), so here and elsewhere with for *j*=*T*, $\beta_T = -\sum_{j=1}^{T-1}\beta_j$ under the fixed effects constraint $\sum_{j=1}^{T}\beta_j = 0$. Similarly for (3) except that $\underline{\beta}=(\alpha_0, \beta_1^*, \ldots, \beta_{T-1}^*, \theta)$.

As shown below for the R-SW in (3) and the fixed effects NR-SW in (4), the covariance matrix *V* is the overall correlation matrix of ε_{ij}^*, which is made up of *N* block diagonals of V_0 with all off-block diagonal matrix elements being 0. Each V_0 is the correlation matrix of repeated measures within each single unit, with dimension *T*.

$$V = \begin{pmatrix} V_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_0 \end{pmatrix}_{N\,T}, \text{ where } V_0 = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}_{T}. \tag{6}$$

The NR-SW with the strata effects treated as random uses (3) but with *V* now being a block diagonal of shifting stratum variances V_s (*s = 1, … , S*) where the random strata effects are subsumed in ρ_s.

$$V = \begin{pmatrix} V_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_S \end{pmatrix}_{N\,T} \text{ with } V_{\mathrm{s}} = \begin{pmatrix} V_0 & \cdots & \rho_s \\ \vdots & \ddots & \vdots \\ \rho_s & \cdots & V_0 \end{pmatrix}_{n_s\mathrm{T}}. \tag{7}$$

With V_0 as defined above and all V_{s} correlations not in the V_0 being ρ_s, the intra-stratum correlation is mediated by the non-randomization selection effect. The GLS estimate for $\underline{\beta}$ is $\underline{\hat{\beta}} = (X'V^{-1}X)^{-1}X'V^{-1}Y$ and has variance $\Lambda = (X'V^{-1}X)^{-1}\sigma^2$ where Λ

is a square matrix of order *(S+T)* for NR-SW and *(T+1)* for R-SW. The variance of $\hat{\theta}$ is the last diagonal element of Λ. This $\underline{\hat{\beta}}$ is the best linear unbiased estimator (BLUE) for $\underline{\beta}$ and uniform minimum variance (UMVU) if Y_{ij} is normally distributed [30].

Although the inverse of Λ is complicated, the orthogonalized coded $= (I_{\{i\in A_1\}}, \ldots, I_{\{i\in A_S\}}, J_1, \ldots, J_{T-1}, Z^*)$, for the fixed effects NR-SW based on (4) with *V* as defined by (6) simplifies $X'V^{-1}X$ with most cross-products being zero which simplifies derivation of $(X'V^{-1}X)^{-1}$.

As Appendix 3 proves, $Var(\hat{\theta}) = \frac{1}{OTD-IDP}\sigma^2$, where $OTD = \sum_{s=1}^{S} n_s\, Z_s^{*\prime} V_0^{-1} Z_s^*$ stands for the "orthogonalized treatment dispersion" (Z_s^* is the orthogonal coding for the s^{th}stratum in Table A as defined Appendix 2); $IDP =$ $\sum_{s=1}^{S} \frac{[1+(T-1)\rho]}{n_s T} (I'_{\{i\in A_s\}} V^{-1} Z^*)^2$ denotes "intercept dispersion penalty" and reflects reduction from OTD by the treatment term being dispersed about different intercepts α_s (rather than always about a common α_0) due to the fixed effects non-randomization. Full expansions of OTD and IDP are presented in Appendix 3.

For the fixed effects NR-SW, $Var(\hat{\theta}_{NR-SW}) = \frac{1}{OTD-IDP}\sigma^2$

$$= \frac{(1-\rho)\sigma^2}{\sum_{l=1}^{S-1} \frac{t_l\left(\sum_{h=1}^{l} n_h\right)\left(N-\sum_{h=1}^{l} n_h\right)}{N} - \frac{1}{T}\sum_{s=1}^{S} n_s [\sum_{l=1}^{S-1} \frac{t_l\left(N-\sum_{h=1}^{l} n_h\right)}{N} - \sum_{h=1}^{S-1} t_h]^2}. \quad (8)$$

Note that here and elsewhere, $Var(\hat{\theta}_{NR-SW})$ denotes the variance for "fixed effects" modeling.

For the R-SW, IDP=0, therefore the variance reduces to, $Var\left(\hat{\theta}_{R-SW}\right) = \frac{1}{OTD}\sigma^2$

$$= \frac{(1-\rho)\sigma^2}{\frac{1}{(1-\rho)}\sum_{l=1}^{S-1}\frac{t_l\left(\sum_{h=1}^{l} n_h\right)\left(N-\sum_{h=1}^{l} n_h\right)}{N} - \frac{\rho}{[1+(T-1)\rho]}\sum_{s=1}^{S} n_s\left[\sum_{l=1}^{S-1}\frac{t_l\left(N-\sum_{h=1}^{l} n_h\right)}{N} - \sum_{h=1}^{S-1} t_h\right]^2}. \quad (9)$$

For the same given SW design, $Var\left(\hat{\theta}_{R-SW}\right)$ is lower with randomization of units into the shifting strata (compared to fixed effects NR-SW) by the ratio of (9)/(8) being $\frac{OTD-IDP}{OTD} < 1$.

Finally, the solution to $Var\left(\hat{\theta}\right)$ for the random effects NR-SW from (7) is difficult as the ρ_s elements in the $V_s's$ lead to numerically complicated inverses and thus is not presented here.

Both $Var\left(\hat{\theta}_{NR-SW}\right)$ and $Var\left(\hat{\theta}_{R-SW}\right)$ are invariant to t_0 and t_S conditional on the sum $t_0 + t_S = E$ (where *E* denotes the number of time points where all units are homogeneous with respect to intervention assignment which occurs on the front or back "edges" of the SW) for an otherwise identical design. Because of the orthogonal coding, $Z_{ij}^* = 0$ for $j < j_1$ and $j \geq j_S$, meaning that observations falling in these periods contribute equally to IDP and OTD and doing so only by dampening V_0^{-1}. For example, in a stepped-wedge design with *S*=3 and $(t_0, t_1, t_2, t_3) = (1, 1, 1, 2)$, $Var\left(\hat{\theta}\right)$ is invariant to t_0 (or t_3) given the sum of the two is some fixed value *E* where $E = t_0 + t_3 = 3$. Thus, $(t_0, t_3) = (1, 2)$ and $(t_0, t_3) = (2, 1)$ (with $t_1 = t_2 = 1$ remaining the same) achieve the same $Var\left(\hat{\theta}\right)$ since 1+2 = 2+1 = 3.

3.2. Variance and Power Estimation

Hussey and Hughes [12] determined power for R-SW designs by using a Wald test for intervention effect. Similarly, we consider two hypotheses for intervention effect with *S* steps and *T* total visits: H_0: $\theta = 0$; H_1: $\theta = \pm\theta_1$. Here θ_1 is the minimum detectable difference (or effect size δ_1 expressed as a multiple of σ, i.e., $\theta_1 = \delta_1\sigma$) for a stepped wedge design given *α, β, N*. For practical repeated measure designs, the sample sizes are often large enough to permit normal approximation of the non-central *t* distribution when *df* > 30 [31]. With α and β being Type I and Type II errors, these are met for a given θ_1 if $Var(\hat{\theta})$is such that

$$\theta_1 = (z_{1-\frac{\alpha}{2}} + z_{1-\beta})\sqrt{Var(\hat{\theta})}. \quad (10)$$

For the fixed effects NR-SW and the R-SW, respectively, $Var(\hat{\theta})$ is obtained from the GLS variance estimates in (8) and (9). Thus if $\emptyset$ is the cumulative distribution function for standard normal $N(0,1)$, rearranging (10) gives the power for conducting a two-sided test of size α as:

$$Power = \emptyset\left(\frac{\theta_1}{\sqrt{Var(\hat{\theta})}} - z_{1-\frac{\alpha}{2}}\right). \quad (11)$$

4. Balanced SW Designs and Optimality Properties

This section focuses on a specific design we define as "balanced" stepped-wedge, which numerically simplifies formulas and thus enables derivation of optimality properties. The SW design is balanced if the same number of units ($n_1 = \cdots = n_S = n$)

are switched per step with equal step sizes ($t_0 = t_1 = \cdots = t_S = t$). For the general balanced design given *t* and *n*, starting at time *t+1*, *n* units switch to intervention and then *t* measures are taken before the next switch. This continues until all units have switched and after *t* more measures the study ends; thus $T = (1 + S) * t$ and $N = n * S$. Such balanced designs could occur in practice if say it took *t* time points to ramp up application to n new units for each new step. Balanced designs can be either non-randomized (BNR-SW) or randomized (BR-SW). This section first presents and compares simplified formulas for variance of the fixed effects BNR-SW and BR-SW, then investigates the optimal design to find the optimal values of *S* (or equivalently *t*) to achieve greatest power for a balanced design with fixed *T* and *N*. Finally, efficiency of the optimal fixed effects BNR-SW design is compared to that of the NR-DD for any given *T* and ρ.

4.1. Power and Sample Size estimation for Balanced SW Designs

For a balanced fixed effects BNR-SW design, variance in (8) simplifies to:

$$Var\left(\hat{\theta}_{BNR-SW}\right) = \frac{12S(S+1)}{NT(S-1)(S+2)}(1-\rho)\sigma^2. \quad (12)$$

Again here and elsewhere $\hat{\theta}_{BNR-SW}$ denotes the estimated intervention effect for the fixed effects balanced non-randomized model as we did not derive closed form variance estimates for random effects non-randomized models.

For a balanced BR-SW design, variance in (9) simplifies to:

$$Var\left(\hat{\theta}_{BR-SW}\right) = \frac{6S[1+(T-1)\rho](1-\rho)}{NT(S-1)[1+(T-1-\frac{St}{2})\rho]}\sigma^2. \quad (13)$$

Note that (13) is the same as formulas in Woertman et. al. [25] and Hussey & Hughes [12] after the previously noted conversion $\rho = \frac{\tilde{\rho}}{\tilde{\rho} + \frac{1-\tilde{\rho}}{m}}$ (also illustrated in Appendix 1), where $\tilde{\rho}$ denotes the intra-class correlation in their cluster-randomized SW papers.

As Appendix 4 shows (with $\sigma^2_{FE-NR}, \sigma^2_R, \rho_{FE-NR}, \rho_R$ defined in that Appendix), compared to fitting any R-SW, the NR-SW with fixed strata effects (FE) under the same setting will result in a lower model within population measurement variance on Y_{ij} ($\sigma^2_{FE-NR} < \sigma^2_R$) together with a smaller within-unit correlation of Y_{ij} andY_{ij}, ($\rho_{FE-NR} < \rho_R$) due to elimination of variance about the α_s from the total variance about a common α_0. However, from (8) and (12) the $Var(\hat{\theta}_{NR-SW})$ only depends on $\sigma^2$and ρ through the product $(1-\rho)\sigma^2$. To that end, this product is unchanged by application of the fixed effects NR-SW design in that $(1-\rho_{FE-NR})\sigma^2_{FE-NR} = (1-\rho_R)\sigma^2_R = \sigma^2_e$. This invariance property means that the "randomized study design" effect parameters $\sigma^2$and ρ can be directly used in (8) and (12) for estimation of the variance of the intervention effect estimate no matter the impact of the fixed effects NR-SW on the final $\sigma^2$and ρ.

4.2. Optimal t for Balanced SW Designs

A balanced SW design may have a fixed total number of longitudinal times *T* because of budget and/or time constraints. For example, a study may be funded for *T = 6* monthly measures on each unit. Finding the optimal balanced SW design with regards to the step size *t* from all possible integer step sizes (*t= 1, 2* and *3*) that can maximize power (or minimize the sample size needed to obtain a given power) would be important. We start with fixed effects balanced non-randomized designs (BNR-SW), which corresponds to

finding the optimal t^*(or equivalently S^*) that maximizes the power (by minimizing the variance in (8)), given *T*, *N* and ρ.

$$S^* = arg \underbrace{max}_{S} Power = arg \underbrace{min}_{S} Var\left(\hat{\theta}_{BNR-SW}\right);$$

$$t^* = \frac{T}{1+S^*}.$$

The derivative of $\log\left(Var\left(\hat{\theta}_{BNR-SW}\right)\right)$ in (8) with respect to *S* is $\left(\frac{1}{S} - \frac{1}{S-1}\right) + \left(\frac{1}{1+S} - \frac{1}{2+S}\right)$ and is negative for $S \geq 1$ meaning $Var\left(\hat{\theta}_{BNR-SW}\right)$ monotonically decreases as *S* increases. The optimal S^* should be as large as possible, and the corresponding optimal t^*should be as small as possible. Accordingly, $t^* = 1$ maximizes power. Likewise for randomized balanced BR-SW design, the derivative of $\log\left(Var\left(\hat{\theta}_{BR-SW}\right)\right)$ in (13) with respect to *S*, is $\left(\frac{1}{S} - \frac{1}{S-1}\right) - \frac{T\rho}{2(1+S)^2}$, which is also negative for $S \geq 1$ and thus is optimized by $t^* = 1$ as has been previously observed or surmised [12, 25-27].

In the Supplementary Appendix, we investigate designs that are equivalent in terms of *N*, *T* and *S* with ($t_1 = \cdots = t_{S-1} = t$) but t_0 and t_S unconstrained, which we denote as "internally balanced". For these designs, variance is often minimized with t_0 and t_S being less than *t*, which for *t=1* is at $t_0 = t_S = 0$. However, if many pre-existing baseline ($t_0 \gg 1$) or/and post full implementation ($t_S \gg 1$) measures will be available, the Supplementary Appendix shows that reducing the number of steps (i.e. *S*) may increase power.

4.3. Variance Ratio of Balanced Fixed Effects BNR-SW to BR-SW Designs

For the balanced setting with the same *N*, *T*, *S* and *t*, the ratio of $\frac{Var(\widehat{\theta}_{BR-SW})}{Var(\widehat{\theta}_{BNR-SW})}$ in (9)/(8) reduces to (13)/(12) which is a function of *T* and ρ that falls between $\frac{1}{2}$ and 1 as shown below.

$$\frac{1}{2} < \frac{Var(\widehat{\theta}_{BR-SW})}{Var(\widehat{\theta}_{BNR-SW})} = \frac{(T+t)[1+(T-1)\rho]}{2T[1+(\frac{T+t}{2}-1)\rho]} < 1. \quad (14)$$

In particular, with optimal choice $t^* = 1$, the ratio in (14) reduces to:

$$\frac{1}{2} < \frac{Var(\widehat{\theta}_{BR-SW})}{Var(\widehat{\theta}_{BNR-SW})} | \{t^* = 1\} = \frac{(T+1)[1+(T-1)\rho]}{2T[1+(\frac{T-1}{2})\rho]} < 1. \quad (15)$$

The ratio plots in Figure 2 summarize the variance comparisons for optimal balanced BR-SW versus fixed effects BNR-SW designs (*t=1)* as functions of *T* and ρ. The ratio of variances in (14) converges to 1 as ρ increases for any *T*, but at a slower rate as *T* becomes larger.

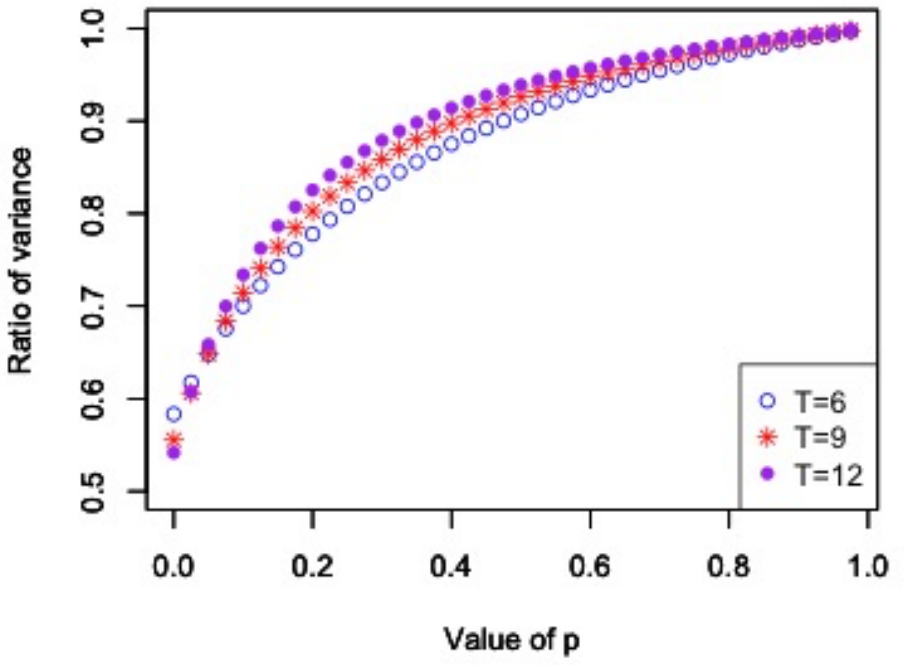

Figure 2: Ratio of variance in optimal BR-SW versus BNR-SW for *T*=6, 9, 12 ($t^* = 1$)

5. Comparing Balanced Fixed Effects BNR-SW to NR-DD Designs

It is also of interest to assess the relative efficiency of NR-SW compared to the more traditional NR-DD study with the same *N* and *T* as there may be settings where an investigator is not able to randomize and has to choose between SW and DD designs. The optimal NR-DD design (Fixed Effects) which minimizes $Var(\hat{\theta}_{NR-DD})$ switches n^*units to intervention after b^* time points [14] where for N even, $n^* = \frac{N}{2}, b^* = \begin{cases} \frac{T}{2}, & if\ T\ is\ even \\ \frac{T-1}{2}\ or\ \frac{T+1}{2} & if\ T\ is\ odd \end{cases}$.

The $Var(\hat{\theta}_{NR-DD})$ in this optimal NR-DD design is $\frac{4T}{Nb^*(T-b^*)}(1-\rho)\sigma^2$.

Taking the ratio of the variance of fixed effects balanced NR-SW versus optimal NR-DD gives (16).

$$\frac{Var(\hat{\theta}_{BNR-SW})}{Var(\hat{\theta}_{NR-DD})|\{optimal\ b^*\}} = \frac{3S(S+1)b^*(T-b^*)}{T^2(S-1)(S+2)} = \begin{cases} \frac{3(S+1)S}{4(S-1)(S+2)}, & if\ T\ is\ even \\ \frac{3(S+1)S}{4(S-1)(S+2)}\frac{(T-1)(T+1)}{T^2}, & if\ T\ is\ odd \end{cases} \quad (16)$$

In particular, for optimal balanced NR-SW with $t^* = 1$, the variance in (12) simplifies to $Var(\hat{\theta}) = \frac{12(T-1)(1-\rho)}{N(T-2)(T+1)}\sigma^2$ and the ratio in (16) reduces to (17).

$$\frac{Var(\hat{\theta}_{BNR-SW})|\{optimal\ t^*=1\}}{Var(\hat{\theta}_{NR-DD})|\{optimal\ b^*\}} = \frac{3(T-1)b^*(T-b^*)}{T(T-2)(T+1)} = \begin{cases} \frac{3(T-1)T}{4(T-2)(T+1)}, & if\ T\ is\ even \\ \frac{3(T-1)^2}{4T(T-2)}, & if\ T\ is\ odd \end{cases} \quad (17)$$

The ratio in (16) and (17) depends only on *T* and *S* (or equivalently *T* and *t* based on $t = \frac{T}{1+S}$). Figure 3 uses (16) to illustrate the ratio of variance for balanced fixed effects NR-SW versus optimal NR-DD for different values of *t*. Each symbol stands for one specific value of *t*. Note that to fit a balanced SW, *T* must be divisible by *t*. Under fixed effects for the same *T (T>3)* and *N*, the ratio in (17) for optimal balanced NR-SW and NR-DD indicates $\frac{Var(\hat{\theta}_{BNR-SW})}{Var(\hat{\theta}_{NR-DD})} < 1$. Now the two designs produce the same variance when *T=3*. But the ratio in (17) increases roughly as *T* increases, and converges to 0.75 when *T*>15. However, for a given *T* as *t* increases, the variance reduction of $Var(\hat{\theta})$ from the fixed effects NR-SW versus the optimal NR-DD, reduces and can reverse, but again for any fixed *t*, as *T* increases, the ratio decreases and converges to 0.75.

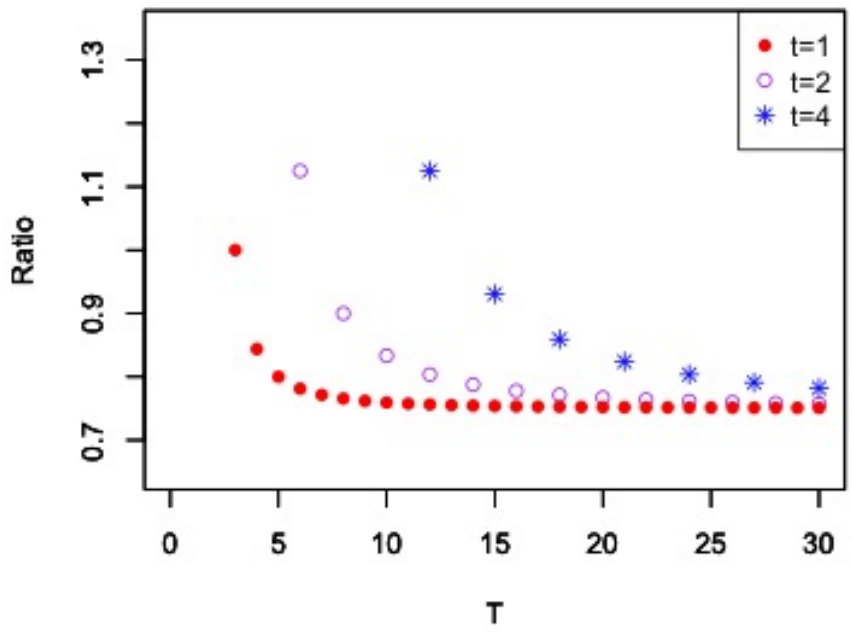

Figure 3: Ratio of variance for BNR-SW versus optimal NR-DD (both are fixed effects designs)

6. Examples from New Jersey Long Term Care Facilities

Both urinary incontinence and ulcers (bedsores) are common chronic conditions for residents that are improved by better treatment at long term care facilities (LTCF). Five Star Quality Data [32] over 7 quarters from Spring 2012 through Fall 2013 reported the average percentages of all long-stay residents that had incontinence and ulcers in each of 270 New Jersey Long Term facilities. The overall quarter averaged unit average of binary outcomes with incontinence was 32% (or 0.32) with $\sigma = 0.14$ and correlation of repeated measures in the same unit was $\rho = 0.85$. For ulcers the unit average was 9.5% (or 0.095) with $\sigma = 0.0475$ and $\rho = 0.50$. We used this normative data to guide estimation of minimal detectable effect sizes δ_1 for BNR-SW designs and compare these

to NR-DD and BR-SW designs for a two sided $\alpha = 0.05, \beta = 0.20$ on intervention trials that would be conducted at *N*=30 long term care facilities lasting from 1.25 (*T=6*) years. While ρ ranged from 0.50 to 0.85, for outcomes in New Jersey LTCF we added $\rho =$ 0.00 and 0.30 to provide insight into outcomes with lower ρ.

Suppose it is impossible to plan a randomized study and one must choose between a NR-DD and a BNR-SW. Table 1 presents the minimal effect size $\delta_1$that can be detected with fixed effects analysis of BNR-SW designs for *t = 1, 2* and *3* from (10) and (12) and optimal NR-DD for $b^* = \frac{T}{2}$. We do not consider random effects model here as it will be directionally biased for NR-DD with only two strata that are unbalanced with respect to proportions treated.

Table 1: Minimal detectable Effect Size (δ_1) in a Study of 30 LTCF (*T*=6) for Non-Randomized designs

T=6	BNR-SW (fixed effects)		NR-DD[1]
	t=1 (S=5)	*t=2 (S=2)*	$b^* = \frac{T}{2}$
$\rho = 0.85$[2]	0.290	0.343	0.323
$\rho = 0.50$[3]	0.529	0.626	*0.590*
$\rho = 0.30$	0.626	0.741	0.699
$\rho = 0.00$	0.749	0.886	0.835

1. Fixed Effects by Default as a Random Effects Model on a Difference in Difference Design is structurally biased
2. This ρ was observed for incontinence over 7 quarters at New Jersey LTCF
3. This ρ was observed for ulcers over 7 quarters at New Jersey LTCF

Thus using ulcers ($\rho = 0.50$), with *T=6* and *t=1*, the minimal detectable effect size from the BNR-SW fixed effects design is δ_1=0.529 or 0.095 ± 0.529*0.0475 which is ≤0.07 or ≥0.12. By contrast, the minimal detectable δ_1=0.590 or 0.095 ± 0.590*0.0475 which is ≤0.067 or ≥0.123 for an NR-DD. While the BNR-SW is preferable from this standpoint, one would have to consider if this benefit were enough if the SW design was more complicated to implement. For *t=2* the BNR-SW is less efficient than the NR-DD when *T=6*.

We caution the reader on one point for interpreting Table 1. Often a large number of baselines with $t_0 > t$ measures is available from historical data where all units were untreated over a long longitudinal monitoring period meaning the t_s's can only be balanced in the future with $t_0 > t_1 = \cdots = t_s$. While the full details are beyond this chapter, the Supplementary Appendix suggests that if this is the case the NR-DD approach becomes more favorable relative to the best possible NR-SW than what is seen in Table 1.

Now suppose a balanced SW design will be used, but the investigator wants to determine if randomization is worth the extra effort. Table 2 compares minimal detectable δ_1 for BR-SW (from (13)), BNR-SW designs analyzed as fixed effects (from (12)) and as random effects (calculated on computer using $\Lambda = (X'V^{-1}X)^{-1}\sigma^2$ with V based on (7)). We let σ and ρ be the same for the BNR-SW and BR-SW random/fixed effects designs since as Section 4.1 and Appendix 4 show any changes on σ and ρ from non-randomization in the fixed effects NR-SW formulation cancel out. We assume ρ_s is proportional to ρ as it seems reasonable that the level of differentiation between the non-randomized shifting strata intercepts will be proportional to the differentiation between

persons. We choose $\rho_s = 0.1\rho$ for a small, $\rho_s = 0.25\rho$ for a noticeable, and $\rho_s = \rho$ as an extreme value for intra-stratum correlation. Note the result for $\rho_s = 0$ (no strata effects) is mathematically the same as BR-SW.

Table 2: Minimal detectable Effect Size (δ_1) in a Study of 30 LTCF (T=6) for Balanced Stepped-Wedge designs

T=6 *N=30*	t	Non-randomized Stepped-wedge				Randomized Stepped-wedge
		Fixed Effects	Random Effects			
			$\rho_s = \rho$	$\rho_s = 0.25\rho$	$\rho_s = 0.1\rho$	
$\rho = 0.85$[1]	$t^* = 1$	0.290	0.289	0.289	0.288	0.287
	$t = 2$	0.343	0.343	0.343	0.342	0.341
$\rho = 0.50$[2]	$t^* = 1$	0.529	0.524	0.517	0.511	0.504
	$t = 2$	0.626	0.625	0.621	0.616	0.605
$\rho = 0.30$	$t^* = 1$	0.626	0.613	0.596	0.584	0.572
	$t = 2$	0.741	0.727	0.727	0.717	0.694
$\rho = 0$	$t^* = 1$	0.749	NA	NA	NA	*0.572*
	$t = 2$	0.886	NA	NA	NA	0.723

Thus, for example, with *T=6,* $\rho = 0.85$, *t=1,* the randomization benefit is barely noticeable with the minimal detectable δ_1 only dropping from 0.290 in a fixed effects BNR-SW down to 0.287 in a R-SW. However, for $\rho = 0$ when *T=6, t=1,* the minimal detectable δ_1 drops from 0.749 in a fixed effects BNR-SW down to 0.572 in the BR-SW. The BR-SW performs better than both fixed and random effects BNR-SW; and the random effects BNR-SW is more powerful than fixed effects BNR-SW in terms of minimum detectable effect size particularly as ρ_s decreases. However, for larger values

of ρ ($\rho \geq 0.50$), as was seen in New Jersey LTCF outcomes, the differences between minimal detectable $\delta_1$between even BR-SW and fixed effects BNR-SW are very small. This suggests that if $\rho \geq 0.50$, the penalty for doing BNR-SW instead of BR-SW on the variance of estimated intervention effect may be ignorable and also that a BNR-SW design should be analyzed as fixed rather than random effects to avoid bias. For smaller ρ ($\rho \leq 0.30$), well below the range of what we saw for outcomes in New Jersey LTCF, the range between minimal detectable δ_1 from BR-SW and random/fixed effects NR-SW are larger. In these settings, it may be more important to fit BR-SW or use random effects analysis on BNR-SW to preserve power. However, other designs such as randomized parallel may be preferable to the SW if $\rho \leq 0.30$ [26].

7. Conclusion

This chapter presents generalized stepped-wedge designs expanded to non-randomized settings. An orthogonalized framework for estimating GLS variance and corresponding power for intervention effects is developed assuming compound symmetry for intra-unit correlation of repeated measures. With the above orthogonal coding for intervention effect, we showed the following properties. First, for any given SW design, randomized unit allocation achieves greater power than does non-randomized allocation analyzed using fixed effects due to the added IDP penalty term in the denominator of the NR-SW variance estimate (i.e. (8) versus (9)). Second, for any otherwise equivalent R-SW and NR-SW fixed effects design, the GLS power estimate is invariant to t_0 $(or\ t_S)$ conditional on the sum $t_0 + t_S (= E)$ because observations falling in those two edges contribute equally to the GLS variance.

To further investigate the optimal design in terms of power, we focused on the balanced SW design that simplifies variance formulas to (12) and (13). For both BR-SW and BNR-SW using fixed effects for a given *T*, the power increases as step size (*t*) decreases and thus the optimal design for both BR-SW and BNR-SW is with $t^* = 1$. In a more comprehensive investigation of optimality for the R-SW, Lawrie et al [27] also observed optimization at $t^* = 1$, but as their analysis allowed n_s to vary (we did not), they observed having $\{n_1 = n_S\} > \{n_2 = \cdots = n_{S-1}\}$ maximized efficiency.

For power approximation in balanced SW designs, the advantage of random allocation decreases as ρ increases and becomes ignorable when $\rho \geq 0.50$. Therefore, for ρ in this range, as was the case for our illustrative example of New Jersey LTCF, we believe a BNR-SW design even analyzed using fixed effects may achieve very similar power as does the comparable BR-SW design. However, potential biases from differential secular trends in non-randomized designs need to be considered [33]. For small values of ρ (i.e. $\rho < 0.30$), we suggest researchers should be cautious to use NR-SW instead of R-SW and perhaps not use any SW design at all. However, in this range of ρ, use of random effects rather than fixed effects analyses to model the non-randomized strata effects considerably improves power in the BNR-SW design. Thus, further research into whether the strata effects could be modeled as random effects in a NR-SW design may be warranted.

In non-randomized settings with fixed *T* and *N*, we discovered that the optimal BNR-SW ($t^* = 1$) is always better than optimal NR-DD in terms of lower variance for the intervention effect estimate. More specifically, for all *t* the relative efficiency of BNR-SW to NR-DD increases as *T* increases, but the ratio of variance eventually converges to

0.75 as *T* gets larger. For any fixed *T*, as the step size t increases, the advantage of BNR-SW to NR-DD gets smaller and eventually reverses to favor NR-DD.

Several limitations should be mentioned. We assumed a constant intervention effect across unit and time, which could be extended by modeling an interaction term of intervention and time and/or including intervention heterogeneity into the covariance structure. While compound symmetry might be a usable approximation if the intra-unit repeated measure correlation does not change or decays slowly over time [31], there may be cases where time decay is too large for CS to be reasonable. In such cases, extension to Toeplitz decay covariance structures may yield a better power estimation. Analytical based methods get much more complicated in random effects NR-SW models due to the additional level of intra-stratum correlation; it is unclear if simple variance estimates can be achieved for this setting. As is often done for stepped wedge studies, we assumed normality or that sample size was large enough for the central limit theorem do hold. Future work on simulation-based methods may be promising to address all of the above issues because they provide flexible alternatives in power and sample size calculation that can deal with specific features of given studies at hand [34].

In conclusion, researchers have recognized the usefulness of stepped-wedge designs in recent years [2, 12, 25, 35]. While considerable development has been made into deriving variance estimates of the intervention effect, this process is still at the beginning for randomized designs and to our knowledge has not been explored for the non-randomized setting. We developed an orthogonalized least squares framework for both R-SW and NR-SW in which number of steps (S), length of step sizes (t_s) and number of units switched at each step (n_s) can be varied. We then focused on balanced settings to

obtain insights on optimal designs in terms of power. While BR-SW always achieved lower variance for the intervention effect estimate than did BNR-SW designs, the differences are small for $\rho \geq 0.50$. Compared to the traditional NR-DD design, optimal BNR-SW ($t^* = 1$) is more efficient. As the length of step size t increases in BNR-SW, the advantage over NR-DD gets smaller and eventually reverses. Further, as the Supplementary Appendix implies, a large number of historical untreated baseline time points may further shift the advantage to the NR-DD approach. Future work perhaps including structured simulations may help to clarify numerous unresolved issues.

References

1. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodology* 2006;6:54.

2. Mdege N, Man M, Brown C and Torgersen D. Systematic review of stepped wedge cluster randomised trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol* 2011;64:936–48.

3. Viera AJ, Garrett JM. Preliminary study of a school-based program to improve hypertension awareness in the community. *Family Medicine* 2008; 40:264e70.

4. Liddy C, Hogg W, Russell G, et al. Improved Delivery of Cardiovascular Care (IDOCC) through Outreach Facilitation: study protocol and implementation details of a cluster randomized controlled trial in primary care. *Implement Science* 2011, 6:110.

5. Gambia Hepatitis Study Group. The Gambia Hepatitis Intervention Study. *Cancer Research* 1987;47:5782e7.

6. Husaini BA, Reece MC, Emerson JS, Scales S, Hull PC and Levine RS. A church-based program on prostate cancer screening for African American men: reducing health disparities. *Ethnicity & Disease* 2008;18(2 Suppl. 2):179e84.

7. Hughes J, Goldenberg RL, Wilfert CM, Valentine M, Mwinga KS and Stringer JSA. Design of the HIV prevention trials network (HPTN) protocol 054: a cluster randomized crossover trial to evaluate combined access to nevirapine in developing countries. *UW Biostatistics Working Paper Series* 2003;195.

8. Somerville M, Basham M, Foy C, et al. From local concern to randomised trial: the Watcombe Housing Project. *Health Expect* 2002;5:127e35.

9. Ciliberto MA, Sandige H, Ndekha MJ, et al. Comparison of home-based therapy with ready-to-use therapeutic food with standard therapy in the treatment of malnourished Malawian children: a controlled, clinical effectiveness trial. *Am J Clin Nutr* 2005;81:864e70.

10. Winani S, Wood S, Coffey P, et al. Use of a clean delivery kit and factors associated with cord infection and puerperal sepsis in Mwanza, Tanzania. *J Midwifery Womens Health* 2007;52(1):37e43.

11. Allegri M, Pokhrel S, Becher H, et al. Step-wedge cluster-randomised community-based trials: an application to the study of the impact of community health insurance. *Health Res Policy Syst* 2008;6:10.

12. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;28(2):182–191.

13. Margaret A. Handley, Dean Schillinger and Stephen Shiboski. Quasi-Experimental Designs in Practice-based Research Settings: Design and Implementation Considerations. *JABFM* 2011; Vol. 24, No. 5.

14. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: The difference-in-differences approach. *JAMA guide to statistics and methods* 2014, Volume 312, Number 22.

15. Athey S, Imbens GW. Identification and inference in nonlinear difference-in-differences models. *Econometrica* 2006; Vol.74, No.2, 431-497.
16. Moher D, Schultz KF, Altman DG, The CONSORT Group: The CONSORT Statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet. 2001, 357: 1191-1194.

17. Pearson D, Torgerson D, McDougall C and Bowles R. Parable of two agencies, one of which randomizes. *Ann Am Acad Pol Soc Sci* 2010;628(1):11e29.

18. Meldrum ML. A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. *Hematol Oncol Clin North Am* 2000, 14 (4): 745–60.

19. West SG, Duan N, Pequegnat W, et al. Alternatives to the randomized controlled trial. *Am J Public Health* 2008; 98(8): 1359-1366.

20. Abadie A. Semiparametric difference-in-differences estimators. *Review of Economic Studies* 2005, 72 (1): 1–19. doi:10.1111/0034-6527.00321.

21. Brand SL, Musgrove A, Jeffcoate WJ and Lincoln NB. Evaluation of the effect of nurse education on patient-reported foot checks and foot care behaviour of people with diabetes receiving haemodialysis. *Diabet Medicine* 2015 Jun 4. doi: 10.1111/dme.12831.

22. Davey C, Boulay M and Hargreaves JR. Strengthening nonrandomized studies of health communication strategies for HIV prevention. *J Acquir Immune Defic Syndr* 2014; 66 (suppl 3): S271–S277.

23. Murray DM, Varnell SP, and Blitstein JL. Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments. *American Journal of Public Health*. 2004, Vol. 94, No. 3, pp. 423-432.
24. Sedgwick P. Randomised controlled trials: balance in baseline characteristics. *BMJ* 2014;349:g5721.

25. Woertman W, de Hoop E, Moerbeek M, et al. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol* 2013, 66(7):752–758.

26. Hemming K, Taljaard MJ. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *Clin Epidemiol*. Epub ahead of print 2015 Sep 5. DOI: 10.1016/j.jclinepi.2015.08.015.

27. Lawrie J, Carlin JB, Forbes AB. Optimal stepped wedge designs. *Statistics and probability letters* 2015; 210-214.

28. Hausman JA. Specification Tests in Econometrics. *Econometrica* 1978; 46 (6): 1251–1271.

29. Snijders TAB, Fixed and random effect. *Encyclopedia of Statistics in Behavioral Science* 2005; Volume 2, 664-665.

30. Aitken AC. On Least-squares and Linear Combinations of Observations. *Proceedings of the Royal Society of Edinburgh* 1934; 55: 42–48.

31. Fisher, RA. Applications of "Student's" distribution. *Metron* 1925; 5: 90–104.

32. Centers for Medicare and Medicaid Services Five Star Quality Rating System, https://www.cms.gov/medicare/provider-enrollment-and-certification/certificationandcomplianc /fsqrs.html

33. Frison L, Pocock SJ. Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Stat Med* 1992; 11: 1685–704.

34. Baio G, Copas A, Ambler G, et al. Sample size calculation for a stepped wedge trial. *Trials* 2015; 16:354.

35. de Hoop EO (2015) Efficient designs for cluster randomized trials with small numbers of clusters. PhD Thesis http://repository.ubn.ru.nl/bitstream/handle/2066/134179/134179.pdf.

Appendix 1: Conversion of Cluster-Randomized Designs to our Setting

We should caution those readers who are familiar with cluster-randomized trials (CRT) that; while our design has level *i*=unit (denoted as cluster in those papers) and *j*=time, it does not have a level *k* nested within *i* and *j* as our examples do not have data down to such a level. This differs from cluster-randomized SW designs [12, 25-27], which do have a level *k* (person-visit measure) from *m* randomly chosen independent patients of cluster *i* at time *j*. In such cluster-randomized designs, Y_{ijk} denotes the outcome of the k^{th}person of cluster *i* from time *j*; however $\bar{Y}_{ij}$ the average outcome of all *m* persons at *i, j* is the functional outcome used in those papers and corresponds to Y_{ij} used here.

Further our notation for ρ is for within *i* intra-class correlation of Y_{ij} and $Y_{ij'}$, which would correspond to the intra-class correlation of $\bar{Y}_{ij}$, $\bar{Y}_{ij'}$. This differs from the intra-class correlation used in cluster-randomized stepped wedge designs taken down to the nested level *k* which are for the correlation of repeated measures Y_{ijk} and $Y_{ij'k'}$ within the same *i*, because our response here is a single measure (Y_{ij}) as opposed to the average of *m* independent observations for a cluster *i* at time *j* ($\bar{Y}_{ij}$). However, if $\tilde{\rho}$ denotes the "ρ" used in those "cluster-randomization notation" papers for intra-class correlation of Y_{ijk} and $Y_{ij'k'}$, then $\frac{\tilde{\rho}}{\tilde{\rho}+\frac{1-\tilde{\rho}}{m}}$ will be the intra-class correlation of $\bar{Y}_{ij}$, $\bar{Y}_{ij'}$ from those designs. Thus to apply our formulas to such cluster-randomized trials, just substitute for ρ in our given formula $\frac{\tilde{\rho}}{\tilde{\rho}+\frac{1-\tilde{\rho}}{m}}$ where $\tilde{\rho}$ is the "ρ" in those cluster-randomized design papers.

Appendix 2: Orthogonal Coding for Intervention Effect

With the orthogonal coding for intervention effect shown below, we are able to simplify the GLS component and calculate the variance of $\hat{\theta}$.

Table A: Summary of orthogonal coding of Z_s^*(*s=1, …, S*) at each step

Transition stratum	$(1, j_1)$ contains t_0 timepoints	(j_1+1, j_2) contains t_1 timepoints	(j_2+1, j_3) contains t_2 timepoints	…	$(j_{S-1}+1, j_S)$ contains t_{S-1} timepoints	(j_S+1, T) contains t_S timepoints
n_S units in S^{th} stratum/step	0	$-\frac{n_1}{N}$	$-\frac{\sum_{s=1}^{2} n_s}{N}$	…	$-\frac{\sum_{s=1}^{S-1} n_s}{N}$	0
n_{S-1} units in $S-1^{\text{th}}$ stratum/step	0	$-\frac{n_1}{N}$	$-\frac{\sum_{s=1}^{2} n_s}{N}$	…	$\frac{N-\sum_{s=1}^{S-1} n_s}{N}$	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n_2 units in 2^{nd} stratum/step	0	$-\frac{n_1}{N}$	$\frac{N-\sum_{s=1}^{2} n_s}{N}$	…	$\frac{N-\sum_{s=1}^{S-1} n_s}{N}$	0
n_1 units in 1^{st} stratum/step	0	$\frac{N-n_1}{N}$	$\frac{N-\sum_{s=1}^{2} n_s}{N}$	…	$\frac{N-\sum_{s=1}^{S-1} n_s}{N}$	0

$$Z_1^* = (\underbrace{0, \dots, 0}_{t_0}, \underbrace{\frac{N-n_1}{N}, \dots, \frac{N-n_1}{N}}_{t_1}, \underbrace{\frac{N-\sum_{s=1}^{2} n_s}{N}, \dots, \frac{N-\sum_{s=1}^{2} n_s}{N}}_{t_2}, \dots, \underbrace{\frac{N-\sum_{s=1}^{S-1} n_s}{N}, \dots, \frac{N-\sum_{s=1}^{S-1} n_s}{N}}_{t_S});$$

$$Z_2^* = (\underbrace{0, \dots, 0}_{t_0}, \underbrace{-\frac{n_1}{N}, \dots, -\frac{n_1}{N}}_{t_1}, \underbrace{\frac{N-\sum_{s=1}^{2} n_s}{N}, \dots, \frac{N-\sum_{s=1}^{2} n_s}{N}}_{t_2}, \dots, \underbrace{\frac{N-\sum_{s=1}^{S-1} n_s}{N}, \dots, \frac{N-\sum_{s=1}^{S-1} n_s}{N}}_{t_S});$$

…

$$Z_S^* = (\underbrace{0, \dots, 0}_{t_0}, \underbrace{-\frac{n_1}{N}, \dots, -\frac{n_1}{N}}_{t_1}, \underbrace{-\frac{\sum_{s=1}^{2} n_s}{N}, \dots, -\frac{\sum_{s=1}^{2} n_s}{N}}_{t_2}, \dots, \underbrace{-\frac{\sum_{s=1}^{S-1} n_s}{N}, \dots, -\frac{\sum_{s=1}^{S-1} n_s}{N}}_{t_S});$$

$$Z^* = (\underbrace{Z_1^*, \ldots, Z_1^*}_{n_1}, \underbrace{Z_2^*, \ldots, Z_2^*}_{n_2}, \ldots \underbrace{Z_S^*, \ldots, Z_S^*}_{n_S}).$$

The advantage of this orthogonal coding is to set cross-products that involve the time periods and the intervention in $X'V^{-1}X$ to 0, which makes it easier to invert the cross-product matrix for the purpose of finding the variance of the intervention effect estimate as is done in Appendix 3.

Appendix 3: Derivation of GLS Variance Estimate

The goal is to find the last element of $(X'V^{-1}X)^{-1}\sigma^2$ which is $Var(\hat{\theta})$. Because of the orthogonal coding, we are able to compute $(X'V^{-1}X)^{-1}$ by applying the inverse of partitioned matrix twice.

$$X'V^{-1}X = \begin{bmatrix} \frac{n_1 T}{[1+(T-1)\rho]} & \cdots & 0 & 0 & \cdots & 0 & I'_{\{i\in A_1\}}V^{-1}Z^* \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \frac{n_S T}{[1+(T-1)\rho]} & 0 & \cdots & 0 & I'_{\{i\in A_S\}}V^{-1}Z^* \\ 0 & \cdots & 0 & \frac{2N}{(1-\rho)} & \cdots & \frac{N}{(1-\rho)} & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \frac{N}{(1-\rho)} & \cdots & \frac{2N}{(1-\rho)} & 0 \\ I'_{\{i\in A_1\}}V^{-1}Z^* & \cdots & I'_{\{i\in A_S\}}V^{-1}Z^* & 0 & \cdots & 0 & \text{OTD} \end{bmatrix}.$$

For the first inverse of portioned matrix, $(X'V^{-1}X)^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$

where $A_{11} = \begin{bmatrix} \frac{n_1 T}{[1+(T-1)\rho]} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{n_S T}{[1+(T-1)\rho]} \end{bmatrix}$,

$$A_{12} = A'_{21} = \begin{bmatrix} 0 & \dots & 0 & I'_{\{i\in A_1\}}V^{-1}Z^* \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & I'_{\{i\in A_S\}}V^{-1}Z^* \end{bmatrix}, \text{ and } A_{22} = \begin{bmatrix} \frac{2N}{(1-\rho)} & \cdots & \frac{N}{(1-\rho)} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \frac{N}{(1-\rho)} & \cdots & \frac{2N}{(1-\rho)} & 0 \\ 0 & \dots & 0 & \text{OTD} \end{bmatrix}.$$

We then apply the inverse of the partitioned matrix to B_{22} as below. $B_{22} = (A_{22} -$

$$A_{21}{A_{11}}^{-1}A_{12})^{-1} = \begin{bmatrix} \frac{2N}{(1-\rho)} & \cdots & \frac{N}{(1-\rho)} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \frac{N}{(1-\rho)} & \cdots & \frac{2N}{(1-\rho)} & 0 \\ 0 & \dots & 0 & \text{OTD} - \text{IDP} \end{bmatrix}^{-1}.$$

For the "nested" inverse of portioned matrix, $B_{22} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}^{-1} = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}$

where $C_{11} = \begin{bmatrix} \frac{2N}{(1-\rho)} & \cdots & \frac{N}{(1-\rho)} \\ \vdots & \ddots & \vdots \\ \frac{N}{(1-\rho)} & \cdots & \frac{2N}{(1-\rho)} \end{bmatrix}$, $C_{12} = C'_{21} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$, and $C_{22} = [OTD - IDP]$. It turns out

that $D_{12} = D'_{21} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$. Therefore, $Var(\hat{\theta}) = D_{22}\sigma^2 = (C_{22} - C_{21}{C_{11}}^{-1}C_{12})^{-1}\sigma^2 =$

$\frac{1}{OTD-IDP}\sigma^2$.

Where the full expansions of OTD, IDP and OTD-IDP are:

$OTD = \frac{1}{(1-\rho)}\sum_{l=1}^{S-1}\frac{t_l\left(\sum_{h=1}^{l}n_h\right)\left(N-\sum_{h=1}^{l}n_h\right)}{N} - \frac{\rho}{[1+(T-1)\rho](1-\rho)}\sum_{s=1}^{S}n_s[\sum_{l=1}^{S-1}t_l\frac{N-\sum_{h=1}^{l}n_h}{N} -$

$\sum_{l=1}^{s-1}t_l I_{\{s>1\}}]^2$;

$IDP = \sum_{s=1}^{S}\frac{[1+(T-1)\rho]}{n_s T}({X_{\alpha_s}}'V^{-1}X_{Tx})^2 = \frac{1}{[1+(T-1)\rho]T}\sum_{s=1}^{S}n_s[\sum_{l=1}^{S-1}t_l\frac{N-\sum_{h=1}^{l}n_h}{N} -$

$\sum_{l=1}^{s-1}t_l I_{\{s>1\}}]^2$;

$$OTD - IDP = \frac{1}{(1-\rho)}\{\sum_{l=1}^{S-1}\frac{t_l\left(\sum_{h=1}^{l} n_h\right)\left(N-\sum_{h=1}^{l} n_h\right)}{N} - \frac{1}{T}\sum_{s=1}^{S} n_s\,[\sum_{l=1}^{S-1}\frac{t_l\left(N-\sum_{h=1}^{l} n_h\right)}{N} - \sum_{h=1}^{s-1} t_h]^2\}.$$

Note that in the R-SW design from (3), IDP=0 and

$$X'V^{-1}X = \begin{bmatrix} \frac{NT}{[1+(T-1)\rho]} & 0 & \cdots & 0 & 0 \\ 0 & \frac{2N}{(1-\rho)} & \cdots & \frac{N}{(1-\rho)} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \frac{N}{(1-\rho)} & \cdots & \frac{2N}{(1-\rho)} & 0 \\ 0 & 0 & \cdots & 0 & \text{OTD} \end{bmatrix}. \text{ Therefore, } Var\left(\hat{\theta}_{R-SW}\right) = \frac{1}{OTD}\sigma^2.$$

Appendix 4: Variance in Fixed Effects NR-SW is Invariant to Absorption from Strata Dispersion

For the randomized model in (3),

$$Y_{ij} = \mu_i + \beta_j^* + \theta Z_{ij}^* + \varepsilon_{ij},$$

where the mean displacement of unit i is $\mu_i \sim N(\alpha_0, \tau^2)$, $\varepsilon_{ij} \sim N(0, \sigma_e^2)$. In this R-SW design the total variance decomposition $\sigma_R^2 = \tau^2 + \sigma_e^2$, and the correlation of repeated measures from the same unit is $\rho_R = \frac{\tau^2}{\tau^2+\sigma_e^2}$, thus $1 - \rho_R = \frac{\sigma_e^2}{\tau^2+\sigma_e^2}$.

A non-randomized stepped wedge (NR-SW) to the same setting can be presented in (4) as

$$Y_{ij} = \alpha_s + (\mu_i - \alpha_s) + \beta_j^* + \theta Z_{ij}^* + \varepsilon_{ij},$$

where $\alpha_s \sim N(\alpha_0, \sigma_s^2)$, $(\mu_i - \alpha_s) \sim N(0, \tau^2 - \sigma_s^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_e^2)$; α_s, μ_i are independent. The variance with $\mu_i$treated as random effects (RE) decomposes into $\sigma_{RE-NR}^2 = \sigma_s^2 + (\tau^2 - \sigma_s^2) + \sigma_e^2 = \tau^2 + \sigma_e^2$ (which is the same total variance as σ_R^2). Now the within-unit factor $\mu_i$has been decomposed into "stratum mean" and "unit mean

about stratum mean" random factors $\alpha_s + (\mu_i - \alpha_s)$. When fitting a random effects model to the NR-SW design, the within-unit repeated measure correlation is $\rho_{RE-NR} = \frac{\tau^2}{\tau^2+\sigma_e^2}$ (the same value as ρ_R), but there is also a within strata correlation of measures of different units in the same stratum of $\rho_s = \frac{\sigma_s^2}{\tau^2+\sigma_e^2}$.

However, for the NR-SW fixed effects (FE) model, α_s the strata effects are treated as fixed eliminating σ_s^2 from both the overall variance and the within-unit correlation. The within-unit factor now is $(\mu_i - \alpha_s)$. Thus for the NR-SW fixed effects model applied to this setting, the overall variance of an observation is $\sigma_{FE-NR}^2 = (\tau^2 - \sigma_s^2) + \sigma_e^2$ and the within-unit mean variance correlation is $\rho_{FE-NR} = \frac{\tau^2-\sigma_s^2}{\sigma_{NR}^2} = \frac{\tau^2-\sigma_s^2}{(\tau^2-\sigma_s^2)+\sigma_e^2}$, and $(1 - \rho_{FE-NR}) = \frac{\sigma_e^2}{(\tau^2-\sigma_s^2)+\sigma_e^2}$.

This leads to a potential awkwardness in Tables 1 and Table 2 where different NR-SW fixed effects designs are compared to each other and to the R-SW and NR-SW random effects designs in that both the variance and the intra-class correlations are changed from $\sigma^2 = \sigma_R^2$ and $\rho = \rho_R$ to $\sigma^2 = \sigma_{RE-NR}^2$ and $\rho = \rho_{RE-NR}$ by the absorption of variance into the stratum specific intercepts in the NR-SW designs. However, $Var(\hat{\theta}_{NR-SW})$ in (8) and $Var(\hat{\theta}_{BNR-SW})$ in (12) for fixed effects NR-SW only depend on $\sigma^2$and ρ through the product$(1 - \rho)\sigma^2$. To that end, this product is unchanged by application of the fixed effects NR-SW design in that $(1 - \rho_{FE-NR})\sigma_{FE-NR}^2 = (1 - \rho_R)\sigma_R^2 = \sigma_e^2$. Thus the parameters for $\sigma^2$and ρ from the "pre-nonrandomized" study design population can be used in (8) and (12) no matter what the impact of the fixed effects NR-SW on the final $\sigma^2$and ρ.

A special case of this principle is when $n_s = 1$ for all s. In this case μ_i and α_s are unidentifiable and the maximization process sets $\mu_i = \alpha_s$ which results in $\sigma_s^2 = \tau^2$ or the estimated target $\sigma_{FE-NR}^2 = \sigma_e^2$; $\rho_{FE-NR} = 0$. But again, $(1 - \rho_{FE-NR})\sigma_{FE-NR}^2 = (1 - \rho_R)\sigma_R^2 = \sigma_e^2$ so (12) can be applied with normative parameters from the randomized design.

Supplementary Appendix: Design Changes in Fixed Effects Models

Most of the paper focused on balanced designs with $t_0 = t_1 = \cdots = t_{S-1} = t_S = t$, and $n_1 = n_2 = \cdots = n_S = n$ as these are easiest to resolve mathematically and often have logistical advantages. We show here that given a fixed *N* and under certain constraints on *S*, *T*, and/or or *E*, other designs can achieve lower variance for $\hat{\theta}$. Note that the formula numbers given here and References cited are in the main chapter. For easier conceptualization, we refer to the time periods on t_0 with no units treated as the "Front Edge" of the wedge and the time periods on t_S with all units treated as the "Back Edge" of the wedge and combined t_0, t_S as the "Edges". The remaining time periods on $t_1, \ldots, t_{S-1}$ are the "Interior".

A. Minimal Fixed effects NR-SW Variance with $S \geq 3, t_0 \geq 1$and $t_S \geq 1$

Often for logistical and/or ethical regions the Front Edge and Back Edge should have at least one time period each and there must be at least one interior period (or $S \geq 3, t_0 \geq 1$, $t_S \geq 1$. Lawrie, Forbes and Carlin [27]) showed that under these common constraints, the optimal R-SW design is almost fully balanced with, $S = T - 1$ and $t = 1$ but not constant *n*; $n_1 = n_S = \frac{(1+2\rho)N}{2(1+\rho(T-1)}$ and $n_2 = \cdots = n_{S-1} = \frac{\rho N}{(1+\rho(T-1)}$. Setting $\rho = 1$ in the previous fractions derives the optimal allocation for the NR-SW design under the same constraints thus $S = T - 1, t = 1, n_1 = n_S = \frac{3N}{2T}$ and $n_2 = \cdots = n_{S-1} = \frac{N}{T}$ with the ratio $\frac{3N}{2T} / \frac{N}{T}$ =1.5. This for example follows from that as ρ goes to 1, the ratio of (8) / (9) goes to 1 meaning that the optimal R-SW design converges to the optimal NR-SW design. However, the optimal NR-SW design is invariant to ρ, which only appears as a multiplier $(1 - \rho)$ in (13). As an illustrative example consider a fully balanced design $T = 4$

months (January, February, March and April), $S = 3$, N=120, $t = 1$ and in the balanced design n=40. As the first row of Table A shows $Var(\hat{\theta}_{BNR-SW})$ for this design from (11) is 0.0300 $(1-\rho)\sigma^2$. Going back to this example in Table A (keeping T, S and N the same under the constraints that t_0 and t_S each > 1), the optimal 1.5: 1 Edge: Interior allocation is $n_1 = n_3 = 45$, $n_2 = 30$, (i.e. ratio of n_1 (or n_3) to n_2 is 1.5). The variance from (8) is reduced very slightly from the balanced design to 0.0296 $(1-\rho)\sigma^2$.

Table A: $Var(\hat{\theta}_{NR-SW})$ for Balanced and Other SW Designs with N=120 subjects, $t_j = 1$ unless specified otherwise

Study Design	Month / Cumulative Number Treated				$Var(\hat{\theta}_{NR-SW})$
	Jan (and before)	Feb	March	April (and after)	
Balanced (S=3, T=4)	0	40 ($n_1 = 40$)	80 ($n_2 = 40$)	120 ($n_3 = 40$)	$0.0300(1-\rho)\sigma^2$
Optimal 1.5:1 Allocation (S=3, T=4)	0	45 ($n_1 = 45$)	75 ($n_2 = 30$)	120 ($n_3 = 45$)	$0.0296(1-\rho)\sigma^2$
Edges Trimmed Internally Balanced $t_0 = t_3 = 0$ (S=5, T=4)	24 (n_1 = 24)	48 ($n_2 = 24$)	72 ($n_3 = 24$)	96[a] ($n_4 = 24$)	$0.0278(1-\rho)\sigma^2$
Edges Expanded to $t_0 = t_3 = 3$[b] Internally Balanced (S=3, T=8)	0[b] $t_0 = 3$	40 ($n_1 = 40$)	80 ($n_2 = 40$)	120[b] ($n_3 = 40$) t_3=3	$0.0232(1-\rho)\sigma^2$
Edges Expanded to $t_0 = t_2 = 3$[b,c] Internal Steps Merged (S=2, T=8)	0[b] $t_0 = 3$	60 ($n_1 = 60$) $t_1 = 2$		120[b] ($n_2 = 60$) $t_2 = 3$	$0.0208(1-\rho)\sigma^2$

a Note $n_5 = 24$ are never treated
b $t_0 = 3$ (November, December, January), $t_S = 3$ (April, May, June)
c Equivalent to a DD design $t_0 = 6$ (August, September, …, February), $t_1 = 2$ (March, April), $t_2 = 0$

B. Trimming the Edges lowers $Var(\hat{\theta})$ when T is constrained

Now allow the edges t_0, t_S to take on different values when the interior times (t_s's) are constant $t_1 = \cdots = t_{S-1} = t$ and $n_1 = n_2 = \cdots = n_S = n$. We call this an internally balanced design in that the times in the middle are the same. Such designs may occur in practice for example if the intervention once started must be ramped up on a standard

time schedule but otherwise measures at varying numbers of times can be available before any units have been treated and/after all units have been treated. Again let $t_0 + t_S = E$, then $T = E + (S-1)t$.

We start by optimizing the ratio of E to t under constrained *S* and *T*. With constant *n*, as defined before $N = n * S$. Therefore, the variance for intervention effect in (8) and (9) can be rewritten for the internal balanced non-randomized stepped wedge (IBNR-SW) as $Var(\hat{\theta}_{IBNR-SW}) = \frac{12T(1-\rho)}{n(S-1)(S+1)t(2T-St)}\sigma^2$. Taking the derivatives of $Var(\hat{\theta}_{IBNR-SW})$with regards to t and setting to 0 (as the second derivative is negative) yields the optimal allocation ratio of $\frac{E}{t}$, i.e., $\frac{\partial Var(\hat{\theta}_{IBNR-SW})^{-1}}{\partial t} = 2T - 2St = 2(E-t)$ which equals 0 when $E^* = t^*$ which means $t^* = \frac{T}{S}$. Thus, the IBNR-SW under constrained T and S is optimized when $(t_0 + t_S)^* = t^*$the number of time periods in each internal step.

A similar result happens for the internally balanced R-SW (IBR-SW) where $Var(\hat{\theta}_{IBR-SW}) = \frac{6[1+(T-1)\rho](1-\rho)}{nt(S-1)[1+(T-1-\frac{St}{2})\rho]}\sigma^2$. Similarly, $\frac{\partial Var(\hat{\theta}_{IBR-SW})^{-1}}{\partial t} = 1 + (T-1-St)\rho = 1 + (E-t-1)\rho = 0$ when $E^* = t^* + 1 - \frac{1}{\rho}$ Thus, for the IBR-SW under constrained *T* and *S*, the optimal allocation is given by $E^* = max\ (t^* + 1 - \frac{1}{\rho}, 0)$, which means $T = (S-1)t^* + max\ (t^* + 1 - \frac{1}{\rho}, 0)$. Therefore, $(t_0 + t_S)^* \leq t^*$. Thus if $\frac{S-1}{T+S-1} < \rho$, then $E^* = 0$ meaning no "Edge" time periods.

Moreover, as was seen in this chapter, *t=1* is often optimal. With *t=1*, $T = E + (S-1)t = E + (S-1), E \geq 0$. After plugging in the above design parameters, the variance for intervention effect in (8) simplifies for IBNR-SW with *t*=1 to:

$$Var\left(\hat{\theta}_{IBNR-SW}\right) = \frac{12T(1-\rho)}{n(S-1)(S+1)(2T-S)}\sigma^2 \quad \text{(A1)}$$

Taking the first derivative of with regards to *S* yields to the optimal allocation of *E* and *t* as: $\frac{\partial Var\left(\hat{\theta}_{IBNR-SW}\right)^{-1}}{\partial S} = 4ST - 3S^2 + 1 > 0$ for $S \in (0, T+1)$. Thus the variance is maximized by maximizing *S* by making $E^* = (t_0 + t_S)^* = 0$. Going back to the illustrative example of Table A with *T*=4 and *N*=120, now with $t_0 = t_S = 0$, *S* becomes 5 and *n* = 120/5 = 24. The third row of Table A shows $Var\left(\hat{\theta}_{IBNR-SW}\right)$ for this design from (A1) reduces to $0.0278(1-\rho)\sigma^2$.

It should be noted that a similar result holds for the IBR-SW with *t*=1:

$$Var(\hat{\theta}_{IBR-SW}) = \frac{6[1+(T-1)\rho](1-\rho)}{n(S-1)[1+(T-1-\frac{S}{2})\rho]}\sigma^2, \quad \text{(A2)}$$

so $\frac{\partial Var\left(\hat{\theta}_{IBR-SW}\right)^{-1}}{\partial S} = 1 + \left(T - S - \frac{1}{2}\right)\rho$ for $S \in (0, T+1)$ which again is > 0 for $S \in (0, T+1)$ or again S is maximized when $E^* = (t_0 + t_S)^* = 0$. However, we could not find a simple expansion for the Larwy, Forbes, Carlin optimal design [27] to R-SW or NR-SW settings where t_0 and/or $t_S = 0$ was allowed.

C. Expanding Edges lowers $Var\left(\hat{\theta}\right)$ when Number of Interior Times is Constrained

In some settings number of interior times of a NR-SW may be constrained in that ramp up must occur over a fixed number of consecutive time points possibly using an

interior balanced design with *n* units added to treatment each time. However, there may be flexibility in the edges in that pre ramp up time $t_0 > t$ and post ramp up times $t_S > 1$ could be used. From (8) for the NR-SW such an expansion of *E* lowers the variance through increasing *T* on the order of $(1 - \frac{1}{T})$ and thus has limited benefit for *T* large. (A similar result in (9) dampened by ρ holds for the R-SW when $\rho > 0$). For example with the NR-SW in Table A if t_0 is expanded to 3 (November, December, January) and t_S to 3 (April, May, June, then *T*=8. From (A1) the variance reduces to $0.0232(1 - \rho)\sigma^2$.

But we should note that it the investigator knows $E = t_0 + t_S = 6$ homogenous treatment measures will be available from before/after implementation of the treatment ramp up, a better approach would be to merge the internal steps thus shifting half the sample (60 subjects) to treatment in February and shift the remaining 60 subjects onto treatment in April. This variance of the estimated intervention effect further reduces to $0.0208(1 - \rho)\sigma^2$(row 4 of Table A). While a full exploration is beyond the scope of this chapter, availability of extra observations at either end of the edges may push the optimal design in the interior from being "SW" towards being flat (dampening the internal steps with ~50% of subjects treated throughout). If the extra observations are all available at t_0 (the front edge), this is pushing the design towards DD.

While a full exploration is beyond the scope here, availability of extra observations at either end of the edges may push the optimal design in the interior from being "SW" towards being flat (dampening the internal steps with ~50% of units treated throughout). If the extra observations are all available at t_0 (the front edge), this is pushing the design towards DD.

In summary, we believe that the NR-SW design differences for $Var(\hat{\theta})$ seen in Table A were often small and would be even smaller for larger *T* and *S*. Thus other considerations such as ethical and logistical may be more important for choosing between potential NR-SW (and also R-SW) designs. However, if preexisting baseline ($t_0 \gg 1$) or post full implementation ($t_S \gg 1$) will be available, using all of these measures improves power more; including that elimination or dampening of internal steps (i.e. reducing *S*) could be beneficial.

Overall Conclusions

The three main chapters comprising my dissertation build upon one another to investigate the intervention effect using GLS power estimation framework based on covariance of repeated measures in longitudinal one-way crossover studies. Based on the number of crossover time points when units switched onto intervention, one-way crossover studies have been further classified into difference-in-differences (DD as discussed in Chapter 1 and Chapter 2) and the stepped-wedge designs (SW as illustrated in Chapter 3).

Chapter 1 and Chapter 2 start with the simplest one-way crossover design by studying difference-in-differences designs where all the units that are switched to the intervention are done so at the same time point. By investigating on the simple compound symmetry and more general Toeplitz correlation structures, Chapter 1 developed a unified GLS power estimation framework together with the alternative lower bound approaches for power estimation in the randomized difference-in-differences (R-DD) studies. The theoretical results for optimal pre-post allocation based on CS approximation are presented and compared to the empirical Toeplitz results from the nursing homes and HIV infected patients' examples with $T=b+k=7$. For these examples where $T=b+k=7$ in the R-DD studies, setting the number of pre-intervention measurements $b=1$ produced optimal or close to optimal results to maximize power to detect an intervention effect, but having $b > 1$ often performed nearly as well in terms of power (i.e. variance of the intervention effect estimate).

Although randomization is always preferred as a gold standard in clinical trials, it is not always feasible due to practical constraints. By modeling the non-randomization

effect associated with the central tendencies of each intervention arm using general linear model, Chapter 2 extended the GLS power estimation framework to the non-randomized difference-in-differences (NR-DD) setting. The optimal pre-post allocation for NR-DD studies is given by equal number of pre-and post-intervention measures ($b=k$) for T even and $|b-k|=1$ for T odd. With the advantage of closed form GLS variance formulas for R-DD and NR-DD under CS approximation, the superiority of randomized over non-randomized setting are quantitatively measured. While given the same b and k, randomized designs are superior, non-randomized designs deliver nearly as precise estimates of intervention effect for high within-unit correlation and/or with more baseline than follow-up measurements ($b \gg k$).

As illustrated in Chapter 1 and Chapter 2 where there is uncertainty about the exact Toeplitz structure in Difference-in-Differences studies, CS approaches approximate the "unknown" variance of the estimated intervention effect well when $b=1$ but can greatly underestimate this variance when $b > 1$. To avoid overestimation in power, two conservative approaches are proposed: PCS approximation based on mean summary statistics can serve as a conservative lower bound for GLS power calculation but greatly underestimate the power in two of our examples; an alternative lower bound approach with $T=b+k=2$ longitudinal measures ($b=1$ and $k=1$) can obtain nearly as precise estimates of the intervention effect as did any design with $T=b+k=7$ measures where $b > 1$ in these two cases. However, none of these approximations performed uniformly well.

Chapter 3 presents the general one-way crossover designs known as the stepped-wedge designs, where the intervention is delivered at sequential time points but

eventually all units would receive the intervention. The Orthogonalized Least Squares power estimation framework is developed based on compound symmetry approximation. To investigate the optimal designs in terms of power, balanced SW designs are investigated because they can further simplify implementation of GLS variance formulas. For both BR-SW and BNR-SW designs, the optimal design to maximize power is given by $t_s \equiv 1$ which means that new units are shifted to intervention at each time point in the study. In the examples we used from New Jersey nursing homes, when compared to BR-SW designs, equivalent BNR-SW designs even with intercepts of non-randomly stepped switching strata analyzed using fixed effects sacrifice little efficiency given an intra-unit repeated measure correlation $\rho \geq 0.50$. Compared to traditional NR-DD designs, optimal BNR-SW designs are more efficient with the ratio of variances of these designs converging to 0.75 when T>10. For any fixed T, as the step size t_s increases ($t_s > 1$), the advantage of BNR-SW to NR-DD gets smaller and eventually reverses to favor NR-DD.

Several limitations need to be considered for all of the chapters in this dissertation. In the general linear models for both DD and SW designs, the assumption of constant intervention effect across unit and time may not hold. However, if so, the models presented here can be extended by modeling an interaction term of intervention and time and/or including intervention heterogeneity into the covariance structure. Although the covariance is assumed to be static (a minimum requisite to use historical data for future correlation estimation), it could change over time due to uncontrollable mechanisms in practice. Relaxation of the above assumptions may likely lead to complicated settings that perhaps can only be addressed with simulation. The illustrative examples presented are limited with a fixed total time points ($T = 7$). While more comprehensive analyses

for other values of T in general and other correlation structures is beyond the scope of this dissertation, the correlation structures in the four examples presented here are likely generalizable and that $T \approx 7$ may be reasonable for many settings with repeated measures taken at 3-6 month intervals.

In conclusion, this "three paper" dissertation develops an Orthogonalized Least Squares power estimation framework based on covariance of repeated measures in longitudinal one-way crossover studies. For researchers who are interested in planning a one-way crossover study with longitudinal repeated measurements, our efforts to identify simple and conservative approximations based on compound symmetry and mean summary approaches have mixed success.