UNIVERSITY OF CALIFORNIA

Los Angeles

Stepped Wedge Designs: Extensions to Studies with Multiple Interventions and Multistate Outcomes

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Phillip Taylor Sundin

2022

ABSTRACT OF THE DISSERTATION

Stepped Wedge Designs: Extensions to Studies with Multiple Interventions and Multistate Outcomes

by

Phillip Taylor Sundin
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2022
Professor Catherine M Crespi, Chair

Stepped wedge design (SWD) trials are cluster randomized trials that feature staggered, unidirectional cross-over between treatment conditions. Existing literature for SWDs focuses primarily on designs with two conditions, typically a control and an intervention condition, and a continuous outcome. The work for this dissertation is motivated by the NORVAX study, a SWD trial implemented at clinics in a safety-net health system to estimate the effectiveness of two interventions for promoting HPV vaccination among adolescents. The outcome for the NORVAX study is patient vaccination status, which is a multistate outcome (no doses, one dose, or two doses). This dissertation has two parts that make contributions to the literature regarding two salient features of the NORVAX study: the multiple interventions in a SWD and the multistate vaccination status outcome.

The first part of this dissertation develops methods for conducting power calculations for SWDs with multiple treatment conditions and a continuous outcome. We present a linear mixed model for such designs and derive standard errors of the intervention effect coefficients. Power for detecting intervention effects is calculated analytically assuming a normally distributed Wald test statistic under an alternative hypothesis. We apply the proposed method to both repeated cross-sectional and cohort designs. Design features, with a focus on treatment sequencing across periods, are examined to determine their impact on

power. Simulations are used to verify results.

The second part of this dissertation focuses on the vaccination status outcome and quantifying intervention effects for this outcome within the context of a SWD. A goal of the NORVAX study is to estimate intervention effects as changes in study population-level vaccination initiation and completion percentages, clinically meaningful outcomes. We propose a semi-Markov multistate cure model in which the number of doses of a vaccine received by the patient are the states. Sojourn times are assumed to be Weibull distributed. To account for individuals who will never receive their next required dose, we include cure proportions in the multistate model. Using the multistate cure model framework, population-level initiation and completion percentages are obtained by converting transition intensity estimates into transition probabilities. Intervention effects are quantified as changes in initiation and completion percentages attributable to interventions. We apply the model to both simulated and real-world data and highlight challenges of this modeling technique.

The dissertation of Phillip Taylor Sundin is approved.

Gang Li

Hilary Aralis

Beth Glenn

Catherine M Crespi, Committee Chair

University of California, Los Angeles

2022

To my parents, Mel and Kristi Sundin, whose constant encouragement and humor have been so supportive throughout this entire journey

Table of Contents

List of Figures . . . . . **viii**

List of Tables . . . . . **ix**

Acknowledgments . . . . . **x**

Curriculum Vitae . . . . . **xi**

1 Introduction **1**

1.1 Stepped Wedge Designs 1

1.2 Motivating Example 4

1.3 Modeling Objectives and Challenges 7

2 Power Analysis for Stepped Wedge Trials with Multiple Interventions . **9**

2.1 Introduction 9

2.2 Model Description 11

2.3 Power Analysis 14

2.4 Examples 17

2.4.1 Two Separate Single-Intervention SWDs vs a Concurrent SWD . . . 17

2.4.2 Factorial Designs with Additive Treatment Effects 20

2.4.3 Factorial Designs with Interaction Effect 21

2.4.4 Four-Arm Design 23

2.5 Simulation 25

2.6 Discussion 27

3 Vaccination Outcome via Multistate Modeling in a Stepped Wedge Design 31

3.1 Introduction . . . 31

3.2 Data . . . 34

3.3 Multistate Models . . . 35

3.3.1 Cure Proportions . . . 39

3.3.2 Likelihood and Bayesian Formulation . . . 40

3.4 Population Level Percentage Estimates . . . 43

3.4.1 Solving for Transition Probabilities . . . 43

3.4.2 Intervention Effect on Study Population-Level Percentage Outcomes . 45

3.5 Simulation . . . 46

3.6 Application to NORVAX Data . . . 51

3.7 Discussion . . . 55

4 Discussion . . . 58

Appendix A - Derivation of Standard Errors . . . 60

Appendix B - Simulation Procedure . . . 71

Bibliography . . . 73

List of Figures

1.1 Example of Stepped Wedge Design 2

1.2 NORVAX Study Design 5

2.1 Examples of Two Single-Intervention SWDs versus Concurrent SWDs with Two Interventions 18

2.2 Comparison of Power for Either Intervention Effect for a Single-intervention SWD, 12-cluster Concurrent SWD and 10-cluster Concurrent SWD 18

2.3 Power for Comparison of Two Interventions in Concurrent SWDs 19

2.4 Stepped Wedge Factorial Design Examples 20

2.5 Comparison of Power for Main Effects 21

2.6 Variations of Stepped Wedge Factorial Design 23

2.7 Comparison of Power for Detecting Main and Interaction Effects 24

2.8 Comparison of power for each of three interventions for multi-arm trials. Average power is calculated as the mean over all interventions. 25

3.1 HPV Dosing Example with Study Design 35

3.2 Likelihood Example. 42

List of Tables

2.1 Simulation Results: Single-Intervention and Concurrent Designs 26

2.2 Simulation Results: Concurrent Designs and Factorial Design 26

2.3 Simulation Results: Factorial Designs 27

2.4 Simulation Results: Multi-Arm Designs 28

3.1 Simulation Results: Multistate Model with Minimal Censoring 48

3.2 Simulation Results: Multistate Model 49

3.3 Simulation Results: Study-Population Percentages 51

3.4 Simulation Results: Study-Population Intervention Effects 52

3.5 NORVAX Study: Multistate Cure Model Results 53

3.6 NORVAX Study: Observed versus Model-Based Study-Population Level Percentages 54

3.7 Estimates of Intervention Effects on Initiation and Completion Percentages . . . 55

Acknowledgments

Material from Chapter 2 has been published in Statistics in Medicine, please see our work at (Sundin and Crespi, 2022). All work in this dissertation was made possible by PCORI grant PCS-201C1-6482 awarded to Dr. Roshan Bastani.

I would first and foremost like to thank my incredible advisor Dr. Kate Crespi, who has spent many hours reviewing, editing and meeting with me over the years. She has always pushed me to think critically, improve my writing, and challenge me in ways to become a better biostatistician. I would also like to thank Drs. Roshan Bastani, Beth Glenn, and Alison Hermann to help me refine many other skills outside of traditional biostatistics, especially communication and applying statistics to real world applications. I'd like to thank Adriana Diaz and Lina Tieu for the years of working together on the HPV vaccination project. Finally, I would also like to thank Drs. Gang Li, Hilary Aralis, and Beth Glenn for serving on my doctoral committee. They've been exceptional in providing guidance and asking thought-provoking and insightful questions along the way.

To all my classmates, especially those that I started with way back in 2016, I thank you for your humor and positivity during late night study sessions and exam preparation. I also thank my siblings and friends for their support and encouragement along the way. And finally, thank you to Teresa, who always makes me laugh.

Curriculum Vitae

2016–2022	Teaching Assistant, Biostatistics Department, UCLA. Courses: Introduction to Biostatistics, Statistical Power and Sample Size Methods for Health Research
2016–2022	Research Assistant, Biostatistics Department, UCLA.
2018	M.S. (Biostatistics), UCLA, Los Angeles, California.
2014–2016	Deloitte Consulting - Strategy and Operations Management Consultant
2014	B.S. Industrial Engineering Pennsylvania State University, University Park, PA.

Publications

Sundin, P. and Crespi, C. M. (2022). Power analysis for stepped wedge trials with multiple interventions. Statistics in Medicine.

Nguyen, M. , Amoon, A. , Lee, L., Chiang, V., Nham, K., Sun, A., Sundin, P., Flores, Y. (2021). Health Literacy, Knowledge, and Risk Factors for Fatty Liver Disease among Asian American and Pacific Islanders and Latinos in Los Angeles. Asian Pacific Journal of Cancer Prevention: APJCP, 22(6), 1737.

Hunt, X., Laurenzi, C., Skeen, S., Swartz, L., Sundin, P., Weiss, R. E., & Tomlinson, M. (2021). Family disability, poverty and parenting stress: Analysis of a cross-sectional study in Kenya. African Journal of Disability (Online), 10, 1-8.

Laurenzi, C. A., Hunt, X., Skeen, S., Sundin, P., Weiss, R. E., Kosi, V., ... & Tomlinson, M. (2021). Associations between caregiver mental health and young children's behaviour in a rural Kenyan sample. Global Health Action, 14(1), 1861909.

Maxwell, A. E., Sundin, P., & Crespi, C. M. (2020). Disparities in cancer mortality in Los Angeles County, 1999–2013: An analysis comparing trends in under-resourced and affluent regions. Cancer Causes & Control, 31(12), 1093-1103.

Xiong, D., Zhang, L., Watson, G. L., Sundin, P., Bufford, T., Zoller, J. A., ... & Ramirez, C. M. (2020). Pseudo-likelihood based logistic regression for estimating COVID-19 infection and case fatality rates by gender, race, and age in California. Epidemics, 33, 100418.

Laurenzi, C. A., Skeen, S., Sundin, P., Hunt, X., Weiss, R. E., Rotheram-Borus, M. J., & Tomlinson, M. (2020). Associations between young children's exposure to household violence and behavioural problems: Evidence from a rural Kenyan sample. Global Public Health, 15(2), 173-184.

Sundin, P., Callan, J., & Mehta, K. (2016). Why do entrepreneurial mHealth ventures in the developing world fail to scale?. Journal of Medical Engineering & Technology, 40(7-8), 444-457.

Callan, J., Sundin, P., Suffian, S., & Mehta, K. (2014). Designing sustainable revenue models for CHW-centric entrepreneurial ventures. In IEEE Global Humanitarian Technology Conference (GHTC 2014) (pp. 687-693). IEEE.

CHAPTER 1

Introduction

Cluster randomized trials (CRTs) are clinical trials in which entire groups of individuals, called clusters, are randomized to treatment arms (Donner and Klar, 2010). Examples of clusters include schools, hospitals, or health clinics. CRTs are often employed to evaluate interventions delivered at the cluster level, such as a new education program or work-flow modification. A common CRT design is the parallel design in which each cluster receives either the control or treatment condition and outcomes are measured at one time point. CRTs can also utilize a crossover design in which clusters receive either the treatment or control condition for a fixed amount of time and then switch to the other condition (Hooper and Bourke, 2015). The focus of this dissertation is on the stepped wedge design (SWD), an evolving class of CRT designs that incorporate features of both parallel and crossover designs.

1.1 Stepped Wedge Designs

The key element of a SWD is unidirectional crossover, with clusters transitioning from a control condition to a treatment condition at staggered time points throughout the study. Clusters are randomized to different sequences that transition from control to treatment at pre-determined times. All clusters are typically in the treatment condition by the end of the study (Hemming et al., 2017). An example of a SWD is shown in Figure 1.1. Because of the staggered crossover, a SWD introduces new considerations for power calculations and analysis compared to other CRT designs.

SWDs have several potential advantages over parallel and crossover CRT designs. SWDs

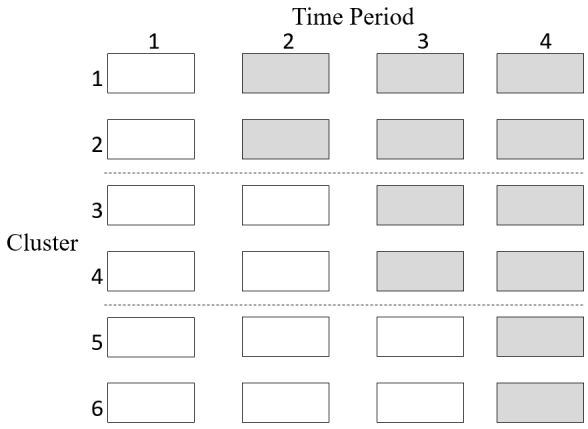


Figure 1.1: Example of a stepped wedge design with six clusters, four time periods, and one treatment. A white cell indicates the control condition and a gray cell represents the treatment condition.

allow comparisons both within cluster and across cluster, potentially resulting in efficiency gains (Woertman et al., 2013; Hemming and Taljaard, 2016). It may be less costly and logistically easier to roll out the intervention over time instead of all at once, as would occur in a typical parallel CRT (Grayling et al., 2017). Additionally, clusters may be more willing to participate if treatment is guaranteed for all clusters. Finally, SWDs can alleviate some ethical concerns because all clusters eventually receive the treatment.

One of the first statistical models for analyzing a SWD is a linear mixed model for a repeated cross-sectional outcome assessment (Hussey and Hughes, 2007). Hussey and Hughes derive a closed form solution for the variance of the estimated treatment effect in this model. This closed form solution allows power calculations to be conducted analytically based on a normally distributed Wald test statistic. For the single treatment SWD with repeated cross-sectional assessments, Woertman et al. derived a design effect for the specific case where each cluster has its own transition step (Woertman et al., 2013). Based on this design effect, closed form solutions for sample size calculations for a repeated cross- sectional SWD have been calculated (Hemming and Taljaard, 2016).

Hemming et al. propose a number of extensions to the Hussey and Hughes linear mixed model (Hemming et al., 2017). These extensions include varying time trends across clusters

using both fixed and random effects, varying treatment effects across cluster using both fixed and random effects, and a treatment interaction with time. Teerenstra et al. discuss SWDs with multiple layers of clustering (Teerenstra et al., 2019). Hughes et al. consider heterogeneous treatment effects across clusters (Hughes et al., 2015).

Sample size calculations have also been determined for both open and closed cohort designs (Hooper et al., 2016). There is additional literature for determining sample sizes of a closed cohort SWD with a random effect for individual and a binary outcome using simulation (Baio et al., 2015). There also exist sample size calculations for generalized estimating equations (GEEs) analysis (Li et al., 2018). Random effect misspecification has also been explored in SWDs (Voldal et al., 2022). Power and sample size considerations for a one treatment SWD with unequal cluster sizes has been explored by Martin et al. and Girling in separate papers (Martin et al., 2019; Girling, 2018).

Taljaard et al. consider the risks, both statistical and clinical, of having few clusters in a SWD (Taljaard et al., 2016a). Girling and Hemming detail optimal design considerations for SWD designs and consider efficiency calculations for different treatment allocations (Girling and Hemming, 2015). Kasza et al. discuss both treatment heterogeneity across clusters and how time periods where no data are collected impact power analysis (Kasza et al., 2019). Analyzing SWDs as a "difference-in-differences" has also been explored when treatment effects are heterogeneuous with respect to time (Lindner and McConnell, 2021).

The literature for stepped wedge designs is a growing body of diverse work as statisticians explore many of the statistical models that can be fit to such designs. Existing literature also features a diverse set of applications of SWDs for many kinds of settings and interventions. However, much of the literature, both in application and methodology, focus on a stepped wedge design with only a single intervention and either a continuous or binary outcome. Stepped wedge designs with more than one treatment have been seen in practice but have not been studied extensively (Whittingham et al., 2014; Reuther et al., 2014; van der Geest et al., 2019). SWDs with multiple treatments and a condition receiving multiple treatments simultaneously have been proposed in literature but without methodological development

(Lyons et al., 2017). Studies may also wish to define outcomes that are neither continuous nor binary. The motivating example for this dissertation requires exploring extensions to both of these modeling features.

1.2 Motivating Example

The University of California, Los Angeles (UCLA) is currently partnering with Northeast Valley Health Corporation (NEVHC) to conduct the Northeast Valley HPV Vaccination Study (NORVAX) study. NEVHC is a large multi-site federally-qualified health center serving a primarily uninsured or publicly insured, low-income, Latino population in Los Angeles County. The goal of the NORVAX study is to evaluate the effectiveness of two interventions to increase the uptake of human papillomavirus virus (HPV) vaccination among adolescents (Bastani, 2017). HPV vaccines were first introduced in the U.S. in 2006 (Jit, 2021); however, the vaccination rate among adolescents remains low, with national rates of HPV vaccine completion close to 50% as of 2018 (Walker et al., 2019). At the beginning of the NORVAX study in November 2017, approximately 35% of NEVHC's patients between the ages of 12 and 17 had completed their HPV vaccine regimen.

The currently utilized HPV vaccine requires multiple doses (CDC, 2019). Individuals who receive their first dose before age 15 require two doses that are at least 6 months apart. Individuals who receive their first dose after their 15th birthday require three doses. (There are very few such individuals in the NORVAX study, and we neglect this possibility.) Individuals who have received at least one dose are considered "initiated"; those who have received all required doses are considered "completed".

The NORVAX study uses a SWD to estimate the effectiveness of two interventions for increasing HPV vaccination initiation and completion. One intervention is a text message reminder sent to parents of adolescents due for an HPV dose. The other intervention is a multi-component clinic-based program comprised of provider and staff education, audit and feedback, establishment of clinic-level policies and protocols, implementation of work-

flow modifications to minimize missed opportunities for vaccination, and patient activation. These interventions are termed the "reminder intervention" and the "clinic-based intervention" for the remainder of the dissertation. The study also features a condition in which clinics implement both interventions simultaneously, termed "the combined condition".

The study randomized seven clinics serving pediatric patients to different sequences of conditions. The SWD of the study is displayed in Figure 1.2. All seven clinics begin in the usual care condition in Period 1. In Period 2, two clinics begin implementing the reminder condition; these clinics transition to the combined condition in Period 5. Three clinics remain in usual care until Period 3, when they begin the clinic-based intervention; these clinics remain in this condition until the end of the study. Two clinics remain in usual care until Period 4, when they begin implementing the combined intervention. The study is currently ongoing as of April 2022, and data have been collected through the end of Period 4. The study is scheduled to conclude data collection in November 2023.

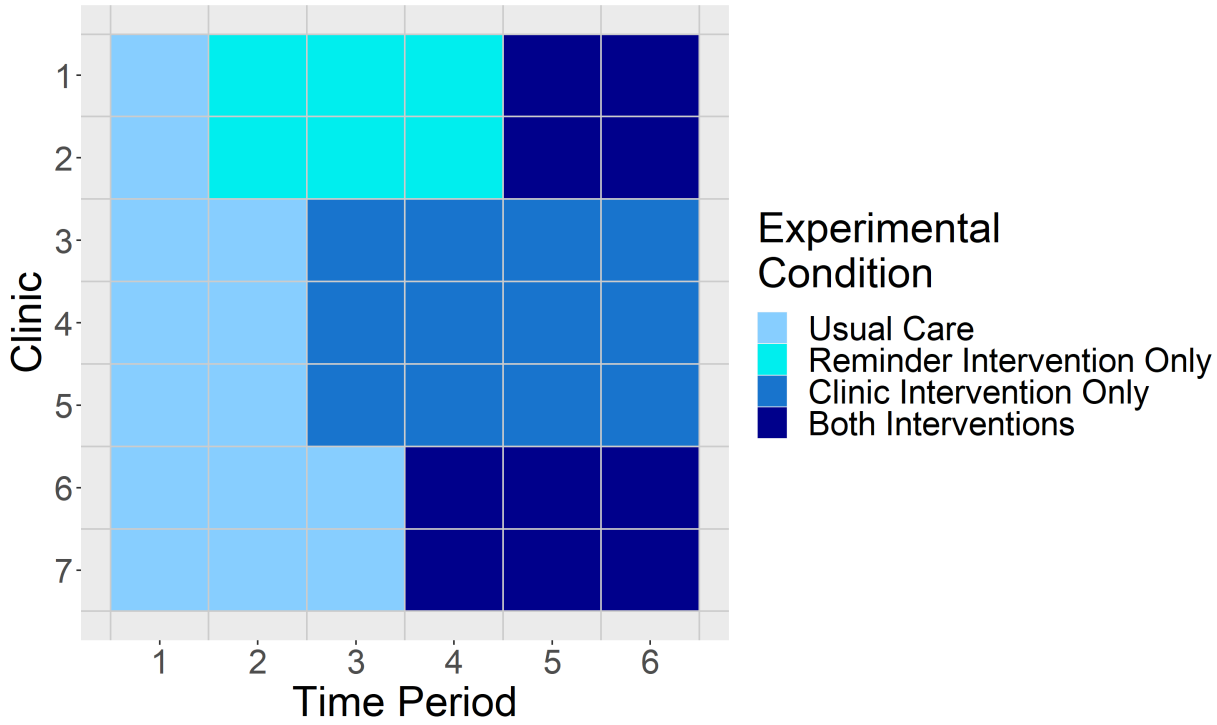


Figure 1.2: NORVAX Study Design

The study population of interest is the adolescent population at NEHVC between the ages of 12 and 17 years (i.e., before the 18^{th} birthday). Although the HPV vaccine can be given as young as 11 years old per CDC recommendations, 12 years was selected as the younger age limit for the study in order to focus on adolescents who were not yet vaccinated after a year of being eligible. When combined with the stepped wedge design, this creates a dynamic, open cohort study population; as the trial progresses, individuals can "age in" or "age out" of the study. Individuals also have to be active patients to be included in the study; an active patient is defined as having a visit to any NEHVC clinic in the past two years. This also contributes to an evolving study population over the course of the study.

There have been several other intervention studies with HPV vaccination completion as the primary outcome. Some of these studies used parallel designs with fixed cohorts (Fu et al., 2016; Borg et al., 2018; Hurley et al., 2019). Such studies were able to assess vaccination outcomes and estimate intervention effectiveness using pre-post comparison of the same individuals. In a SWD with an open cohort, this approach is not possible. A few studies utilize a SWD to test interventions to increase HPV vaccination rates. The DOSE HPV trial utilized a stepped wedge design with the primary outcome measured at the patient visit level, rather than the individual or clinic level (Perkins et al., 2020). The outcome was whether or not a dose-eligible patient received a dose at the visit. A second study, which has a published study protocol but has not yet published an outcome analysis, plans to define the dose-eligible population at the beginning of each step of the SWD and determine vaccination status at the end of the period (Rutten et al., 2018). This framework also uses a binary outcome. A third study in Oregon, also with a published study protocol but no outcome analysis yet published, plans to analyze clinic-level completion and missed opportunity rates measured quarterly using generalized linear mixed models (Carney et al., 2019). No further details were provided on the statistical analysis plan. These three examples illustrate that there is no standard statistical approach for a SWD with a HPV vaccination outcome.

1.3 Modeling Objectives and Challenges

The first notable modeling challenge for the NORVAX study is the implementation of multiple treatments within a stepped wedge design. Modeling data from stepped wedge design trials requires careful consideration given the staggered implementation of the interventions, potential confounding with time trends, and the hierarchical nature of the data (Hussey and Hughes, 2007). Much of the existing SWD literature focuses on designs with only a single treatment. Designs with multiple treatment conditions, including conditions with multiple treatments implemented simultaneously, can be analyzed both as factorial designs and multiarm trials. In the NORVAX study, an individual can be exposed to multiple different study conditions over the course of the study. Two of the seven clinics will experience two different intervention conditions throughout the study. An individual in one of these clinics could potentially experience up to three different conditions including the control condition. Any modeling approach will have to account for such individuals. We noted a lack of literature for these trial designs, particularly in designing and calculating power for detecting intervention effects.

A second major challenge is appropriately modeling the HPV vaccination outcome and quantifying intervention effectiveness in a clinically and policy relevant manner. The study investigators have expressed that defining intervention effectiveness in terms of percentage point differences in completion and initiation between conditions is a clinically meaningful approach. We also want to be able to examine potential moderators of the interventions, some of which are measured at the individual level such as gender. Designing a model that interprets intervention effectiveness in population-level percentage point changes while also making use of individual-level covariates poses significant challenges for statistical modeling. Directly modeling clinic-level initiation and completion percentages could be considered. However, this would hinder the goal of examining individual-level characteristics as potential moderators of intervention effectiveness. We would like to be able to include both individual-level and clinic-level covariates in analyses. Thus the most fruitful approach is likely to involve modeling the individual-level data.

The interpretation of intervention effects is also further complicated by the dynamic, open cohort aspect of the NORVAX study. Individuals can age in or out of the study population during the trial. Individuals can also move between being active and inactive patients depending on how often they have encounters at the health clinics. Thus the study population changes over the course of the study. We would like to be able to conduct analyses that consider patients to be exposed to a condition if they were dose-eligible and an active patient at a clinic while it was assigned to that condition. This will yield a practical estimate of how effective the interventions are at increasing vaccination coverage in the target population. We would like to be able to answer the question, if we apply this intervention to this target population, by how many percentage points can we expect HPV vaccination initiation or completion to increase?

We ideally also want a single model that captures both vaccination initiation and completion. Each individual in the NORVAX study contributes information on both initiation and completion; by handling both within the same model, we more properly account for the correlation of these outcomes. Furthermore, we would be able to determine whether the effects of interventions or other covariates are the different for initiation and completion.

In this dissertation, we present methodological work based on the NORVAX trial which can be categorized into two areas. We first develop power calculation methods for SWDs with multiple treatments and a continuous outcome. Although the NORVAX study has a non-continuous outcome, it is an important contribution to addressing gaps in the stepped wedge design literature for multiple interventions. In the second area of our work, we focus on the vaccination outcome and develop a multistate cure model for the vaccination outcome within the context of a SWD.

The dissertation is organized as follows. Chapter 2 presents our work on conducting power analyses for stepped wedge designs with multiple treatments and a continuous outcome. Chapter 3 then presents a multistate cure model for modeling the vaccination outcome of the NORVAX. We conclude with a discussion in Chapter 4.

CHAPTER 2

Power Analysis for Stepped Wedge Trials with Multiple Interventions

This section is adapted from (Sundin and Crespi, 2022).

2.1 Introduction

Most research on the design and analysis of stepped wedge trials has focused on SWDs with one intervention condition contrasted with a control condition. There is a small but growing body of literature on SWDs with more than one intervention condition. Some work (Grayling et al., 2019) has focused on studies in which there is a nested natural order of D interventions such that intervention d consists of intervention $d-1$ plus some additional factor. The authors discuss the optimization of treatment sequence allocations and focus on optimal design for such trials. The variance of treatment effect estimates in SWDs with nested interventions has also been studied (Zhang et al., 2020). However, SWDs with multiple interventions that are not nested within one another have not been well studied, and interaction effects also have not received much attention (Lyons et al., 2017).

SWDs with multiple treatment arms are being conducted despite a scarcity of methodological literature. There are several examples of stepped wedge design trials that feature two interventions implemented alone and in combination, as in a 2×2 factorial design. These examples include a trial of the comparative effectiveness of two interventions to promote human papillomavirus vaccination among adolescents (Bastani, 2017), a study examining two interventions for reducing hyperbilirubinaemia in infants (van der Geest et al., 2019), and a

study compared two interventions for addressing behavioral problems in children with cerebral palsy (Whittingham et al., 2014). In these studies, clusters were assigned to sequences that could include periods spent in usual care, the two single intervention conditions, and/or a combined condition.

There are also examples in the literature of several related single-intervention stepped wedge trials conducted simultaneously. The FallDem study used two stepped wedge trials to examine two interventions for improving the lives of dementia patients (Lyons et al., 2017; Reuther et al., 2014), and Durovni et al. conducted two separate SWD trials for tuberculosis screening (Durovni et al., 2013, 2014). In some cases, it might be advantageous to combine two separate trials with stepped wedge designs into one trial with multiple treatment conditions, akin to a multiarm trial.

In this chapter, we consider stepped wedge design trials with more than one intervention, including both multi-arm designs, which involve a control and two or more treatment conditions, and factorial designs, in which interventions are implemented alone and in combination. Multi-arm trials have several advantages, such as allowing for direct comparison of alternative treatments (comparative effectiveness) and resource savings due to "reusing" the same control condition to compare to several interventions. Factorial designs also have potentially increased efficiency and can allow for the estimation of interaction effects (Oelhert, 2010). Thus extending SWDs to incorporate multi-arm and factorial design features could be quite beneficial. We develop power analysis methods for such trials and examine factors that influence power for stepped wedge designs with a normally distributed outcome variable.

The chapter is organized as follows. Section 2.2 introduces the models for the SWD with multiple treatment conditions. Section 2.3 develops power analysis methods. Section 2.4 uses examples to examine the influence of different design features on power. Section 2.6 presents results from a simulation study. Section 2.6 discusses the implications of our work, possible extensions, limitations and future work.

2.2 Model Description

We first present a model for a stepped wedge design with a single intervention and then consider designs with any number of interventions. We focus on designs with only two interventions and an interaction effect as designs with more than two interventions have not yet been seen in practice. The section concludes with an overview of the derivation of the standard errors of the estimated treatment effect coefficients, with details in Appendix A.

We begin with the classic stepped wedge design model with a single binary treatment factor (Hussey and Hughes, 2007). For a design with I clusters observed at T times, and N different individuals per time per cluster, let Y_{ijk} be a continuous outcome for individual k in cluster i at time j. The model for Y_{ijk} is

$$Y_{ijk} = \mu + \alpha_i + \psi_{ik} + \nu_{ij} + \beta_j + X_{ij}\theta_1 + e_{ijk} \tag{2.1}$$

where μ is an intercept, $\alpha_i \sim N(0, \sigma^2_\alpha)$ is a random intercept for cluster i, $\psi_{ik} \sim N(0, \sigma^2_\psi)$ is a random intercept for individual k in cluster i, $\nu_{ij} \sim N(0, \sigma^2_\nu)$ is a random intercept for cluster i in time j, β_j is a fixed effect for time j, X_{ij} is a $\{0,1\}$ indicator for whether cluster i at time j receives treatment, θ_1 is the treatment effect, and $e_{ijk} \sim N(0, \sigma^2_e)$. The total variance of an individual level outcome is $\sigma^2_y = \sigma^2_\alpha + \sigma^2_\psi + \sigma^2_\nu + \sigma^2_e$.

It is straightforward to expand this model to include multiple binary treatment factors (Lyons et al., 2017). Assuming additive treatment effects, the model with R treatment factors is

$$Y_{ijk} = \mu + \alpha_i + \psi_{ik} + \nu_{ij} + \beta_j + \sum_{r=1}^{R} X_{ijr}\theta_r + e_{ijk}, \tag{2.2}$$

where X_{ijr} is a $\{0,1\}$ indicator of whether cluster i at time j receives treatment r and θ_r is the treatment effect for treatment r. For the remainder of this section, we take $R = 2$ for simplicity, with results generalizable to $R > 2$. Adding an interaction effect θ_3, the model becomes

$$Y_{ijk} = \mu + \alpha_i + \psi_{ik} + \nu_{ij} + \beta_j + X_{ij1}\theta_1 + X_{ij2}\theta_2 + X_{ij1}X_{ij2}\theta_3 + e_{ijk}. \tag{2.3}$$

Individual auto-correlation (IAC) is defined as the proportion of the individual-level variance (which in this model is $\sigma_\psi^2 + \sigma_e^2$) that is time-invariant. In model (2.3), the IAC is $\pi = \sigma_\psi^2/(\sigma_\psi^2 + \sigma_e^2)$. Setting $\pi = 0$ yields a repeated cross-sectional design. We can also define the cluster auto-correlation (CAC) as the proportion of cluster level variance that is time-invariant. In this model, the cluster-level variance is $\sigma_\nu^2+\sigma_\alpha^2$ and CAC $= \sigma_\alpha^2/(\sigma_\nu^2+\sigma_\alpha^2) = \rho_a/\rho_w$ (Teerenstra et al., 2012; Feldman and McKinlay, 1994). We also define two intraclass correlation (ICC) values. The within-period ICC, $\text{Corr}(y_{ijk}, y_{ijk'})$ is now $\rho_w = (\sigma_\nu^2 + \sigma_\alpha^2)/\sigma_y^2$ and the across-period ICC, $\text{Corr}(y_{ijk}, y_{ij'k'})$, is $\rho_a = \sigma_\alpha^2/\sigma_y^2$.

Standard errors are needed to compute power. To derive standard errors, it is convenient to work with cluster-level outcomes. Let $\overline{Y}_{ij\cdot} = \frac{1}{N}\sum_{k=1}^{N} Y_{ijk}$ be the mean outcome of cluster i at time j across N individuals. The model for cluster-period means with two treatments and an interaction term is

$$\overline{Y}_{ij\cdot} = \mu + \alpha_i + \psi_i + \nu_{ij} + \beta_j + X_{ij1}\theta_1 + X_{ij2}\theta_2 + X_{ij1}X_{ij2}\theta_3 + e_{ij\cdot}, \tag{2.4}$$

where $e_{ij\cdot} = \frac{1}{N}\sum_{k=1}^{N} e_{ijk} \sim N(0, \sigma_c^2 = \frac{\sigma_e^2}{N})$ and $\psi_i = \frac{1}{N}\sum_{k=1}^{N} \psi_{ik} \sim N(0, \sigma_\zeta^2 = \frac{\sigma_\psi^2}{N})$. In this model, the variance of a cluster-period mean is $\text{Var}(\overline{Y}_{ij\cdot}) = \sigma_c^2 + \sigma_\alpha^2 + \sigma_\zeta^2 + \sigma_\nu^2$, and $\text{Cov}(\overline{Y}_{ij\cdot}, \overline{Y}_{ij'\cdot}) = \sigma_\alpha^2 + \sigma_\zeta^2$.

Define the outcome vector $\boldsymbol{Y} = (\overline{Y}_{11\cdot}, \ldots, \overline{Y}_{iT\cdot}, \ldots, \overline{Y}_{I1\cdot}, \ldots, \overline{Y}_{IT\cdot})'$. Assuming clusters are independent, the variance-covariance matrix of $\boldsymbol{Y}$ is a $IT \times IT$ matrix of the form

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V_1} & 0 & 0 & 0 \\ 0 & \boldsymbol{V_2} & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \ldots & \boldsymbol{V_I} \end{bmatrix},$$

with each $T \times T$ matrix $\boldsymbol{V_i}$ having structure

$$\boldsymbol{V_i} = \begin{bmatrix} \sigma_c^2 + \sigma_\alpha^2 + \sigma_\nu^2 + \sigma_\zeta^2 & \sigma_\alpha^2 + \sigma_\zeta^2 & \dots & \sigma_\alpha^2 + \sigma_\zeta^2 \\ \sigma_\alpha^2 + \sigma_\zeta^2 & \sigma_c^2 + \sigma_\alpha^2 + \sigma_\nu^2 + \sigma_\zeta^2 & \dots & \sigma_\alpha^2 + \sigma_\zeta^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 + \sigma_\zeta^2 & \sigma_\alpha^2 + \sigma_\zeta^2 & \dots & \sigma_c^2 + \sigma_\alpha^2 + \sigma_\nu^2 + \sigma_\zeta^2 \end{bmatrix}.$$

Some practitioners find that standardization of the model can be convenient for power calculations. To standardize the model in (2.3), one divides through by σ_y. The cluster random intercept α_i now has standardized variance ρ_a, the cluster-by-time random intercept ν_{ij} has variance $\rho_w - \rho_a$ for $\rho_w > \rho_a$, the individual-level random intercept ψ_{ik} has variance $\pi(1-\rho_w)$ and the error term e_{ijk} has variance $(1-\pi)(1-\rho_w)$. Thus the variances can be specified in terms of the parameters ρ_w, ρ_a and π. The matrix $\boldsymbol{V_i}$ will have diagonal elements $\rho_w + \frac{(1-\rho_w)}{N}$ and off-diagonal elements $\rho_a + \frac{\pi(1-\rho_w)}{N}$.

Now we turn to the design matrix of the fixed effects. While β_1 rather than β_T is often set equal to zero when the model is fit to data, we follow (Hussey and Hughes, 2007) and set $\beta_T = 0$ for identifiability. The choice is immaterial for power calculations. The $(T+3) \times 1$ regression coefficient vector for the fixed effects is

$$\boldsymbol{\eta} = \begin{bmatrix} \mu & \beta_1 & \dots & \beta_{T-1} & \theta_1 & \theta_2 & \theta_3 \end{bmatrix}'.$$

The full $IT \times (T+3)$ design matrix $\boldsymbol{Z}$ becomes

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z_1} \\ \boldsymbol{Z_2} \\ \vdots \\ \boldsymbol{Z_I} \end{bmatrix}$$

where each matrix $\boldsymbol{Z_i}$ has dimension T $\times$ (T+3) and takes the form

$$\boldsymbol{Z_i} = \begin{bmatrix} \boldsymbol{1_T} & \boldsymbol{I_{T-1}} & \boldsymbol{X_{i1}} & \boldsymbol{X_{i2}} & (\boldsymbol{X_1 X_2})_i \\ & \boldsymbol{0'_{T-1}} & & & \end{bmatrix}.$$

The elements of the vector $\boldsymbol{X_{i1}} = (X_{i11}, X_{i21}, \ldots, X_{iT1})'$ are indicators of whether cluster i at time j receives treatment 1, the elements of $\boldsymbol{X_{i2}} = (X_{i12}, X_{i22}, \ldots, X_{iT2})'$ are indicators of receipt of treatment 2, and $(\boldsymbol{X_1 X_2})_{\boldsymbol{i}}$ is the Hadamard product of $\boldsymbol{X_{i1}}$ and $\boldsymbol{X_{i2}}$, with a value of 1 if cluster i receives both treatments at time j and 0 otherwise. The matrix $\boldsymbol{I_{T-1}}$ contains indicators for each time j from $1, \ldots, (T-1)$. The vector $\boldsymbol{0'_{T-1}}$ corresponds to time T. For designs with $R > 2$, $\boldsymbol{Z_i}$ can be expanded to include the additional indicators.

2.3 Power Analysis

Inference for fixed effects in linear mixed models can be conducted using Wald tests or likelihood ratio tests. We focus on Wald tests. For hypotheses of the form $H_0 : \eta = 0$, where η is a fixed effects coefficient, the Wald test statistic takes the form $\hat{\eta}/\sqrt{\text{Var}(\hat{\eta})}$, where $\hat{\eta}$ is the estimated coefficient, and has an approximate standard normal distribution when the null hypothesis is true (Verbeke and Molenberghs, 2009). The power to reject H_0 for a specific true value of η, denoted η_a, with type I error rate α and a two-sided test, is approximately

$$P\left(\left| \frac{\eta_a}{\sqrt{\text{Var}(\hat{\eta})}} \right| \geq z_{1-\frac{\alpha}{2}} \,\middle|\, \eta = \eta_a \right)$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})^{th}$ percentile of the standard normal distribution.

To calculate power, we need an expression for $\text{Var}(\hat{\eta})$. We derive expressions for $\text{Var}(\hat{\eta})$ using the cluster-period mean models in (2.4). We focus on models with $R = 2$ treatments and an interaction term assuming a factorial design; the results are generalizable to $R > 2$ and multi-arm trials as discussed in Chapter 4. Given the linear mixed model formulation, the variance-covariance matrix of the estimated fixed effect coefficients has the form $\boldsymbol{C} = (\boldsymbol{Z'V^{-1}Z})^{-1}$, where $\boldsymbol{Z}$ is the fixed effects design matrix and $\boldsymbol{V}$ is the variance-covariance matrix of the outcome vector. Our approach is to find expressions for the variances and covariances of treatment effect coefficient estimates, $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$. We do so by calculating $\boldsymbol{Z'V^{-1}Z}$ then invert it to get the elements of $(\boldsymbol{Z'V^{-1}Z})^{-1}$, corresponding to the variances and covariances of the treatment effect coefficients.

Let $\boldsymbol{Z}$ be the $IT \times (T+3)$ design matrix and $\boldsymbol{V}$ be the $IT \times IT$ variance-covariance matrix of the cluster-level outcomes. Let $\sigma_{diag} = \sigma_c^2 + \sigma_\nu^2$ and $\sigma_{off} = \sigma_\alpha^2 + \sigma_\zeta^2$, and for standardized models, $\sigma_{diag} = \frac{(1-\pi)(1-\rho_w)}{N} + \rho_w - \rho_a$ and $\sigma_{off} = \rho_a + \frac{\pi(1-\rho_w)}{N}$. Assuming clusters are independent, $\boldsymbol{V}$ has block diagonal structure with elements $\boldsymbol{V_i} = \sigma_{diag}^2 \boldsymbol{I_T} + \sigma_{off}^2 \boldsymbol{1_T 1'_T}$, where $\boldsymbol{I_T}$ is a $T \times T$ identity matrix and $\boldsymbol{1_T}$ is a $T \times 1$ vector of 1's. Using the Sherman-Morrison formula (Sherman and Morrison, 1949; Bartlett, 1951), we can obtain its inverse as

$$\boldsymbol{V_i}^{-1} = \frac{1}{\sigma_{diag}^2(\sigma_{diag}^2 + T\sigma_{off}^2)} \left[(\sigma_{diag}^2 + T\sigma_{off}^2)\boldsymbol{I_T} - \sigma_{off}^2 \boldsymbol{1_T 1_T}'\right].$$

This matrix has off-diagonal elements

$$\frac{-\sigma_{off}^2}{\sigma_{diag}^2(T\sigma_{off}^2 + \sigma_{diag}^2)}$$

and diagonal elements

$$\frac{(T-1)\sigma_{off}^2 + \sigma_{diag}^2}{\sigma_{diag}^2(T\sigma_{off}^2 + \sigma_{diag}^2)}.$$

Due to the block diagonal structure of $\boldsymbol{V}$, we have

$$\boldsymbol{Z'V}^{-1}\boldsymbol{Z} = \sum_{i=1}^{I} \boldsymbol{Z_i}'\boldsymbol{V_i}^{-1}\boldsymbol{Z_i},$$

where $\boldsymbol{Z_i}$ is the $T \times (T+3)$ part of the design matrix corresponding to cluster i. We can then rewrite

$$\boldsymbol{Z_i}'\boldsymbol{V_i}^{-1}\boldsymbol{Z_i} = \frac{1}{\sigma_{diag}^2(\sigma_{diag}^2 + T\sigma_{off}^2)} \left[(\sigma_{diag}^2 + T\sigma_{off}^2)\boldsymbol{Z_i}'\boldsymbol{Z_i} - \sigma_{off}^2 \boldsymbol{Z_i}'\boldsymbol{1_T 1_T}'\boldsymbol{Z_i}\right]. \tag{2.5}$$

We then use block matrix inversion techniques to solve for the submatrix corresponding to the coefficients of interest. A full derivation is provided in Appendix A.

In the case of a design with $R = 2$ interventions and no interaction term, a closed form solution for the variance of the estimated intervention effects for intervention 1 and 2 can be

calculated using the inverse of a 2×2 matrix, with variances written as $Var(\hat{\theta}_1) =$

$$\frac{l_2 - z_2 - \frac{y_2^2}{fT} - \frac{1}{f+gT}\left(w_2 - \frac{l_2^2}{T}\right)}{\left(l_2 - z_2 - \frac{y_2^2}{fT} - \frac{1}{f+gT}\left(w_2 - \frac{l_2^2}{T}\right)\right)\left(l_1 - z_1 - \frac{y_1^2}{fT} - \frac{1}{f+gT}\left(w_1 - \frac{l_1^2}{T}\right)\right) - \left(q_1 - \frac{y_1 y_2}{fT} - \frac{1}{f+gT}\left(w_{XW} - \frac{l_1 l_2}{T}\right)\right)^2},$$

and $Var(\hat{\theta}_2) =$

$$\frac{l_1 - z_1 - \frac{y_1^2}{fT} - \frac{1}{f+gT}\left(w_1 - \frac{l_1^2}{T}\right)}{\left(l_2 - z_2 - \frac{y_2^2}{fT} - \frac{1}{f+gT}\left(w_2 - \frac{l_2^2}{T}\right)\right)\left(l_1 - z_1 - \frac{y_1^2}{fT} - \frac{1}{f+gT}\left(w_1 - \frac{l_1^2}{T}\right)\right) - \left(q_1 - \frac{y_1 y_2}{fT} - \frac{1}{f+gT}\left(w_{XW} - \frac{l_1 l_2}{T}\right)\right)^2},$$

with all terms defined in Appendix A. Standard errors are calculated by taking the square root of these variances. Closed form solutions for the model with the interaction effect are found in Appendix A.

The standard errors thus derived enable power calculations for hypothesis testing. In factorial and multi-arm design trials, there will typically be multiple hypotheses of interest. When multiple hypotheses are tested simultaneously, multiplicity adjustments should be taken into account in power analysis to control experimentwise Type I error. If a single-step method such as Bonferroni is used, the power calculations can be adjusted by adjusting the significance level for each test. Accounting for the use of other multiplicity adjustment procedures, such as the Hochberg or fixed sequence procedures, can be more complex (Grayling and Wason, 2020).

The calculations also enable the testing of linear contrasts. For example, a comparative effectiveness hypothesis comparing two active treatments may involve the hypothesis $H_0 \colon \theta_1 - \theta_2 = 0$, which can be tested using the Wald statistic $(\hat{\theta}_1 - \hat{\theta}_2)/\sqrt{\text{Var}(\hat{\theta}_1 - \hat{\theta}_2)}$, where $\text{Var}(\hat{\theta}_1 - \hat{\theta}_2) = \text{Var}(\hat{\theta}_1) + \text{Var}(\hat{\theta}_2) - 2\text{Cov}(\hat{\theta}_1, \hat{\theta}_2)$ and the variance and covariances can be obtained as described.

The power method described makes use of a normality-based z-test, which may not hold up well for a small number of clusters. However, for the examples we present in the next section, there was no evidence of small-sample bias, suggesting that this is not always an

issue. The topic of small-sample bias corrections for stepped wedge designs has been explored elsewhere (Ford and Westgate, 2020; Li, 2019).

2.4 Examples

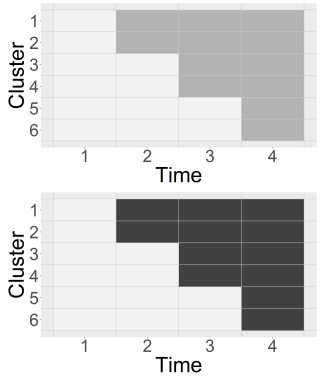
Since the formulas are complex, we present examples to illustrate how power is affected by design features of SWDs, focusing on the impact of sequencing of treatment conditions within clusters. The examples in Sections 2.4.1 and 2.4.2 use standardized effect sizes and realistic but arbitrary values of standardized variance parameters. The examples in Sections 2.4.3 and 2.4.4 use simple effect sizes (in original units) and variance parameter values derived from a real study. For all examples, we set the experimentwise type I error rate to 0.05 and use a Bonferroni correction when conducting multiple simultaneous tests within the same design. Calculations were performed in R version 3.6.1 (R Core Team, 2019) with code available at https://github.com/phillipsundin/SWFD.

2.4.1 Two Separate Single-Intervention SWDs vs a Concurrent SWD

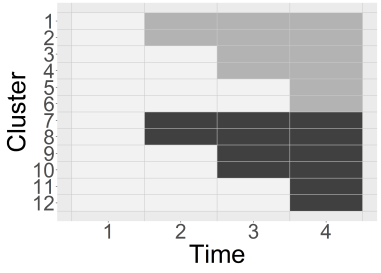
Several studies have conducted two related but separate single-intervention SWD trials (Reuther et al., 2014; Durovni et al., 2013). We explore potential advantages of combining two single-intervention trials into one trial with two interventions, including efficiency gains and comparative effectiveness.

Consider the two single-intervention SWD trials, each with six clusters and four time periods, in Figure 2.1a. Figure 2.1b stacks the two designs into a single 12-cluster trial; such a design has been called a concurrent design (Lyons et al., 2017). Figure 2.1c shows a concurrent design with only 10 clusters. Let δ_1 and δ_2 denote the standardized effect sizes for Interventions 1 and 2 compared to the control condition. We set $\delta_1 = \delta_2 = 0.4$, representing medium effect sizes (Cohen, 1988). Within each design, power for the two intervention effects is the same due to symmetry. We specify $N = 15$ individuals per cluster-period in a repeated cross-sectional design. Power for detecting an intervention effect in one of the

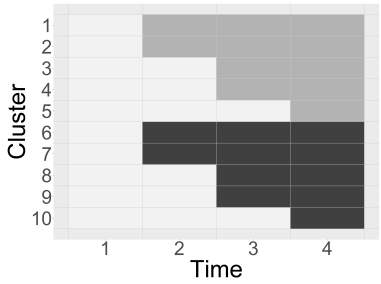
single-intervention SWDs was calculated using model (2.1); for the concurrent SWDs, power was calculated using model (2.2). Type I error was set to 0.05 for hypothesis tests in the single-intervention SWDs and 0.025 for the concurrent designs. In these examples, we fix $\rho_w = \rho_a$, equivalent to setting $\sigma^2_\nu = 0$, and examine power under two different values of π.



(a) Two separate single intervention SWDs



(b) Concurrent SWD with two interventions, 12 clusters



(c) Concurrent SWD with two interventions, 10 clusters

Figure 2.1: Examples of two single-intervention SWDs versus concurrent SWDs with two interventions. White cells indicate cluster-periods in the control condition. Light and dark gray cells indicate treatment conditions for Interventions 1 and 2, respectively.

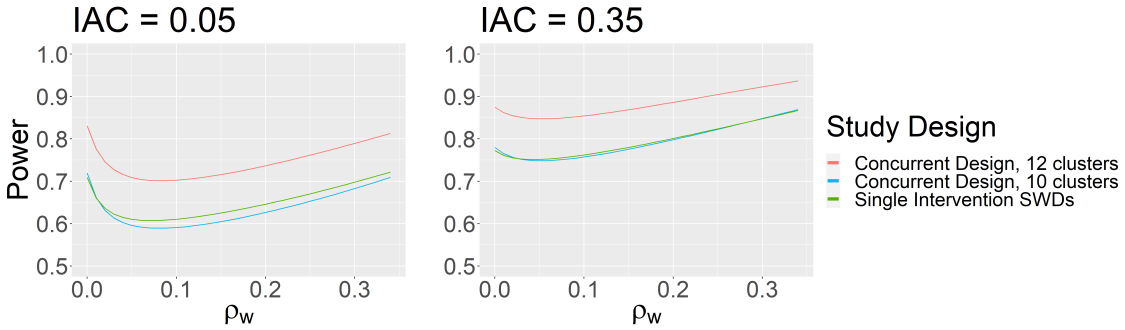


Figure 2.2: Comparison of Power for Either Intervention Effect for a Single-intervention SWD, 12-cluster Concurrent SWD and 10-cluster Concurrent SWD

Figure 2.2 displays power for either intervention effect for the three designs as a function of ρ_w. The convex shapes of the power curves are similar to those observed for SWDs with only one treatment (Woertman et al., 2013; Hemming and Taljaard, 2016; Baio et al., 2015). For both values of π, the 12-cluster concurrent design, which maintains the same total number of clusters as the two separate single intervention SWDs, has power gains ranging

from 0.10 to 0.13 compared to the other designs for the values of ρ_w considered even with a reduced type I error rate. The 10-cluster concurrent design, which reduces the total sample size by about 17% compared to the designs with 12 clusters, has power comparable to that of a single-intervention SWD when $\pi = 0.35$; for $\pi = 0.05$, power for the 10-cluster concurrent design is at most 0.02 lower.

Another advantage of including two interventions in one study is the ability to directly compare them. This can be accomplished using tests of linear contrast, which can be powered using our methods. Suppose we assume standardized effect sizes of 0.30 and 0.70 for the two interventions compared to control, entailing a difference of 0.4 between them (a difference this large may be unrealistic for some studies, but helps to illustrate the principle). We set $\rho_w = \rho_a$ and $\pi = 0.05$. Type I error was set to $0.05/3 = 0.0167$ for each of three tests: the two intervention-to-control comparisons and comparison between the two interventions. Figure 2.3 displays power for the linear contrast as a function ρ_w. The relationship between power and ρ_w for the comparative effectiveness contrast is similar to that for the intervention-to-control hypothesis tests. We note that in a concurrent design, the interventions are conducted in parallel and thus the intervention-to-intervention contrast is less susceptible to confounding by time than the intervention-to-control comparisons.

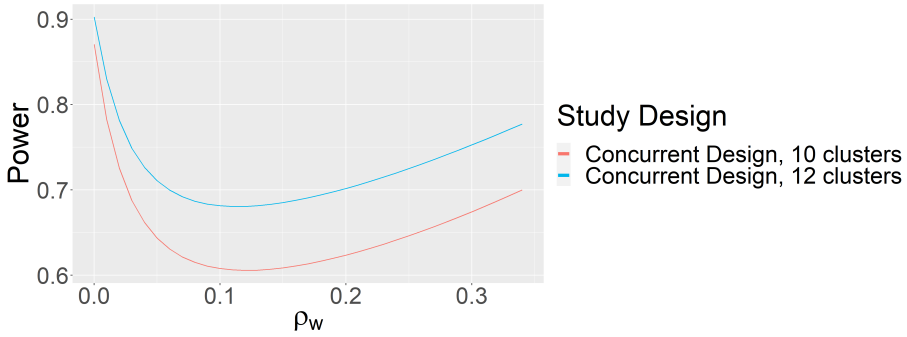
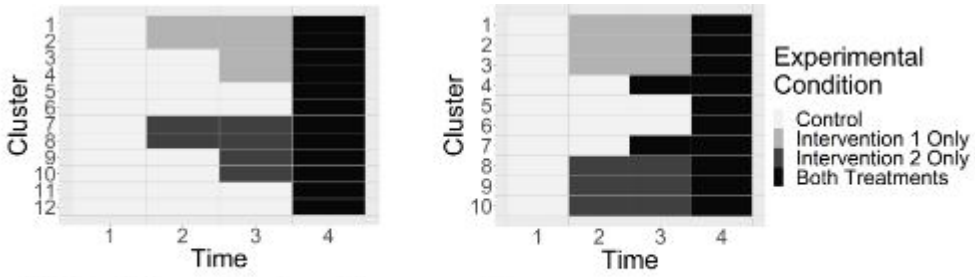


Figure 2.3: Power for Comparison of Two Interventions in Concurrent SWDs

2.4.2 Factorial Designs with Additive Treatment Effects

Our methods enable power analysis for factorial designs, which can be highly efficient when effects are additive, i.e., when there is no interaction effect. To investigate stepped wedge factorial designs, we begin by comparing the 12-cluster, 4-period concurrent design in Figure 2.1b with designs that assign some cluster-periods to a combined condition. Figure 2.4a shows a 12-cluster "late" factorial design in which all clusters transition to the combined condition in the last period. Figure 2.4b shows an "earlier" factorial design with only ten clusters that introduces the combined condition earlier. Additive intervention effects are assumed for these examples.

Both designs feature six cluster-periods in each single intervention condition and twelve cluster-periods in the combined condition. We assume repeated cross-sectional observations, moderate effect sizes ($\delta_1 = \delta_2 = 0.4$) and $N = 15$ individuals per cluster-period.



(a) "Late" factorial design with 12 clusters

(b) "Earlier" factorial design with 10 clusters

Figure 2.4: Stepped Wedge Factorial Design Examples.

Figure 2.5 compares power for the main effect of each intervention for the three designs (Figures 2.1b, 2.4a and 2.4b) as a function of ρ_w for two values of the IAC. Power for the two main effects is identical due to symmetry. The 12-cluster "late" factorial design has the lowest power for all values of ρ_w and π. For $\rho_w < 0.02$, the 12-cluster concurrent and 10-cluster "earlier" factorial designs have similar power; for $\rho_w > 0.02$, the 10-cluster "earlier" design has highest power while still maintaining a 17% reduction in sample size compared to the concurrent design.

This example illustrates several points. First, as expected for a factorial design, when

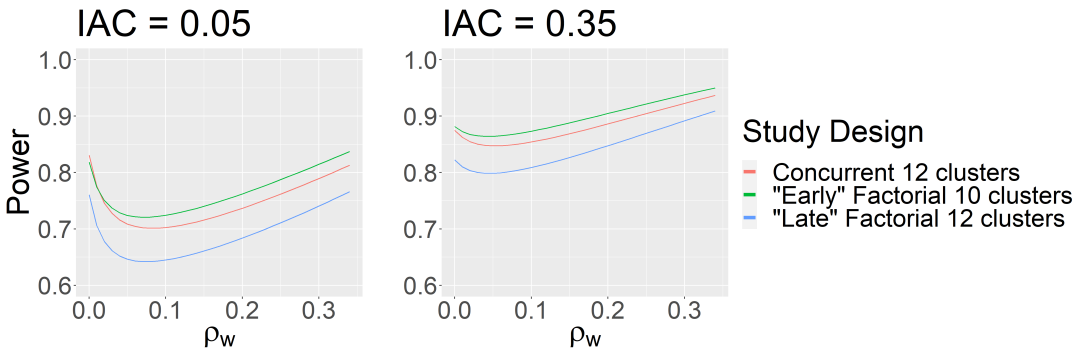


Figure 2.5: Comparison of Power for Main Effects

intervention effects are additive, including a combined condition can increase efficiency and reduce the sample size requirement. However, the timing of transitions to the combined condition matters. Simply assigning all clusters to the combined condition for the last period reduced power compared to a concurrent design. Rather, the combined condition needs to be introduced earlier to realize efficiency gains. Further, if an interaction effect is present, the design in Figure 2.4a suffers from identifiability issues, as the interaction effect would be perfectly collinear with the last time period.

2.4.3 Factorial Designs with Interaction Effect

To study power for detecting an interaction, we consider hypothetical SWDs for evaluating two school-based interventions to reduce obesity among children. The primary outcome will be age- and sex-standardized BMI z-score. Variance parameter values were estimated using data from a previous study that measured BMI z-scores at three time points over 13 month among 286 children at 9 schools (unpublished data). A linear mixed model based on Equation 2.1 was fit to the these data to obtain the estimates $\sigma_e^2 = 1.11$, $\sigma_\nu^2 = 0.14$, $\sigma_\psi^2 = 3.54$, and $\sigma_\alpha^2 = 0.24$ with total variance $\sigma_y^2 = 5.29$. On a standardized scale, these values correspond to $\rho_w = (0.24{+}0.14)/5.29 = 0.07$, $\rho_a = 0.24/5.29 = 0.05$, $\pi = 3.54/(3.54{+}1.11) = 0.76$ and CAC $= 0.24/(0.24{+}0.14) = 0.63$. The study is to be powered on detecting effect sizes of 1 on the z-score scale for each intervention and an interaction effect of 0.5 (also on the z-score scale), corresponding to standardized effect sizes of $1/\sqrt{5.29} = 0.44$ for main effects

and 0.22 for the interaction effect. We note that the z-score outcomes are based on the 2000 Centers for Disease Control and Prevention (CDC) growth charts (Kuczmarski et al., 2002) and not our sample, which had substantially higher variance. The planned study will involve 8 schools with 90 children at each school, and will have five 6-month periods. For simplicity, we assume no dropout.

We consider the designs displayed in Figure 2.6. Each of these designs has seven cluster-periods in Intervention 1 only, seven in Intervention 2 only, and ten in the combined condition. Design #1 is a concurrent design with the combined condition as another "stack". Design #2 is similar to a two-intervention concurrent design but has most clusters further transition to the combined condition. Designs #3 and #4 combine elements of Designs #1 and #2; they are distinguished by Design #4 having earlier introduction of the combined condition and featuring some clusters that never transition to the combined condition. Designs #1, #3, and #4 are symmetric in Interventions 1 and 2 and thus have equal power for these two effects. Design #2 is close to symmetric, but symmetry can be impossible to achieve with a small number of clusters. Type I error was set to $0.05/3 = 0.0167$ for three hypothesis tests.

The investigators considered the IAC of $\pi = 0.76$ in the prior study to be relatively high and thought that it might be lower in the planned study. To explore the impact of IAC on power, Figure 2.7a displays power for main effects for a range of plausible values of π. As shown by others (Hooper et al., 2016), power is an increasing function of π. Design #2 has the highest power for main effects for all values of π. In this design, power for Intervention 1 is slightly higher than that for Intervention 2 due to its more balanced sequencing over time (2 clusters receiving intervention in periods 2, 3 and 4 rather than 1, 2, then 3 clusters). For all values of π, Design #1 has lowest power. Focusing on power for the interaction, displayed in Figure 2.7b, Design #2 has by far the highest power; power for the three other designs is similar. Overall, power to detect the interaction is low.

Design #2 is clearly superior for detecting both main and interaction effects. In Design #2, clusters transition between conditions more than any other design. When there are more transitions, within-cluster comparisons are increased, and thus power to detect effects

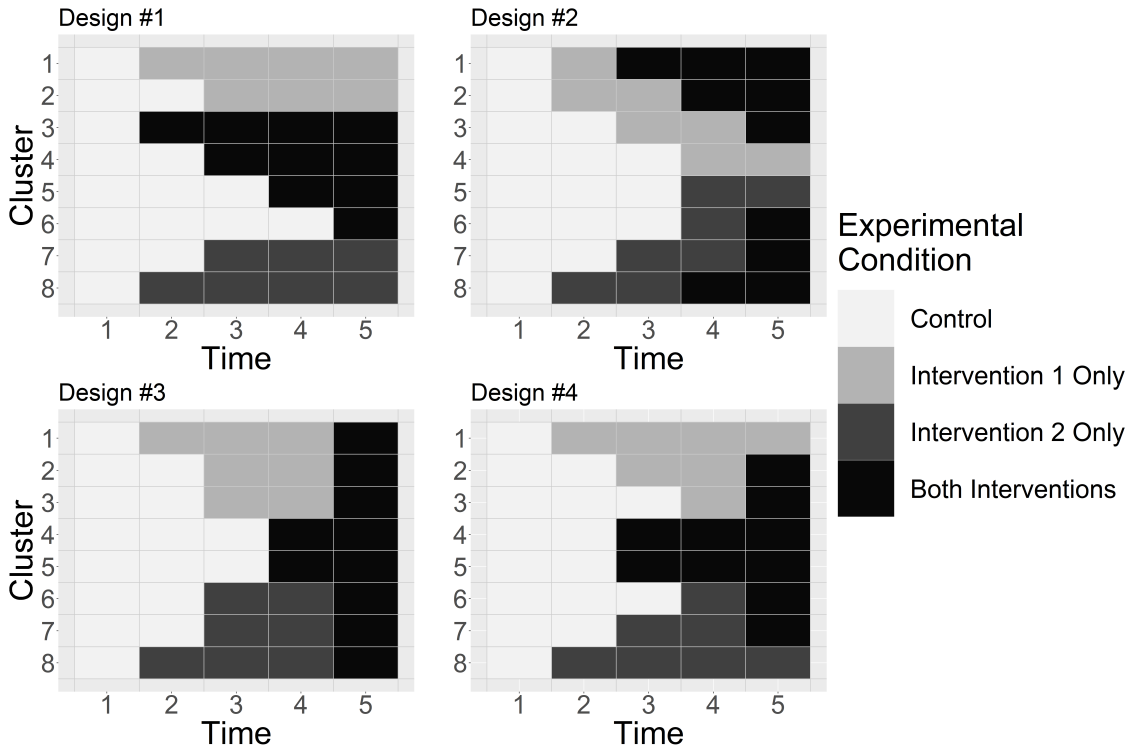


Figure 2.6: Varying sequences in a stepped wedge factorial design.

is increased. In Design #1, cluster transition only once, and thus this design has the lowest power for main effects. Beyond power, other drawbacks of the designs should be considered. For example, in Design #3, time in the combined condition occurs almost entirely during the last period, risking confounding with time. This example also illustrates that to power on the interaction term, designs should include two features: clusters that experience the control, single intervention and combined conditions, and relatively early introduction of the combined condition.

2.4.4 Four-Arm Design

To study multi-arm trials, we continue with SWDs for child obesity interventions using BMI z-score as the outcome variable and the variance parameter estimates from the previous similar study described in Section 2.4.3. We study the designs in Figure 2.6 but regard the combined condition as a third intervention (Intervention 3), and assume the goal is to

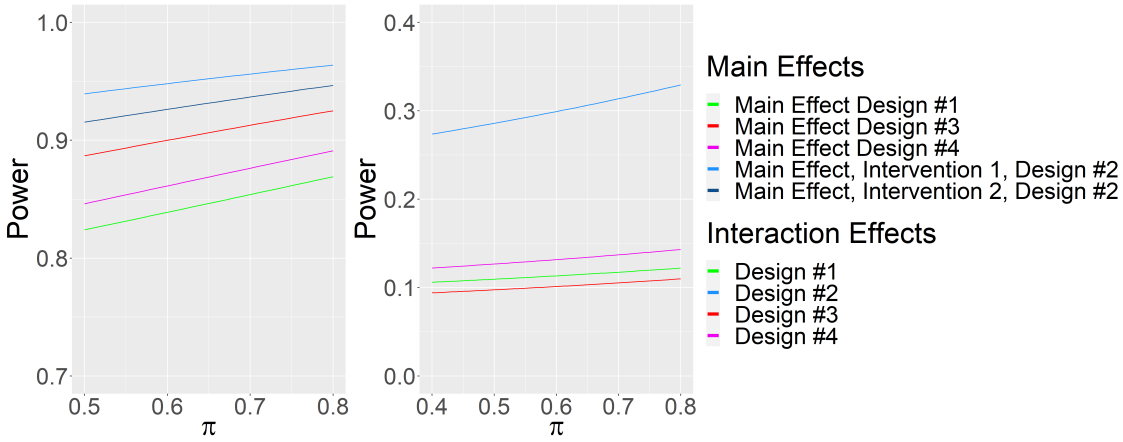


Figure 2.7: Comparison of power for detecting main and interaction effects

compare each of the three interventions to the control condition. Other hypotheses could include direct comparisons of interventions. We assume a simple effect size of 0.92 for each of the intervention arm, corresponding to a standardized effect size of 0.4. Each cluster-period has $N = 90$ individuals. We use the same variance parameters as above, and allow the individual auto-correlation π to vary. Type I error is set to $0.05/3 = 0.0167$ for each of 3 tests. We compute power for each intervention as well as average power.

Power to detect all three interventions individually and averaged is shown in Figure 2.8 as a function of π. This example shows that unlike factorial designs, power in multi-arm trials is less dependent on clusters transitioning to multiple intervention conditions and more dependent on when interventions are first introduced in the study. For Interventions 1 and 2, Design #2 yields the highest power across all values of π, as it introduces the intervention early in the study across multiple clusters. However, for Intervention 3, we see that Design #1 yields the highest power, as this design introduces Intervention 3 earlier in the study than any other design. We also note that for Intervention 3, Design #3 yields the lowest power, as only two cluster-periods are in this condition prior to the final time period, resulting in a significant amount of confounding between time and an intervention effect.

Design #2 has higher power than Design #3 for all interventions. However, Design #2 only outperforms Designs #1 and #4 for Interventions 1 and 2 and has lower power for Intervention 3. This can be attributed to the fact that Design #2 primarily features

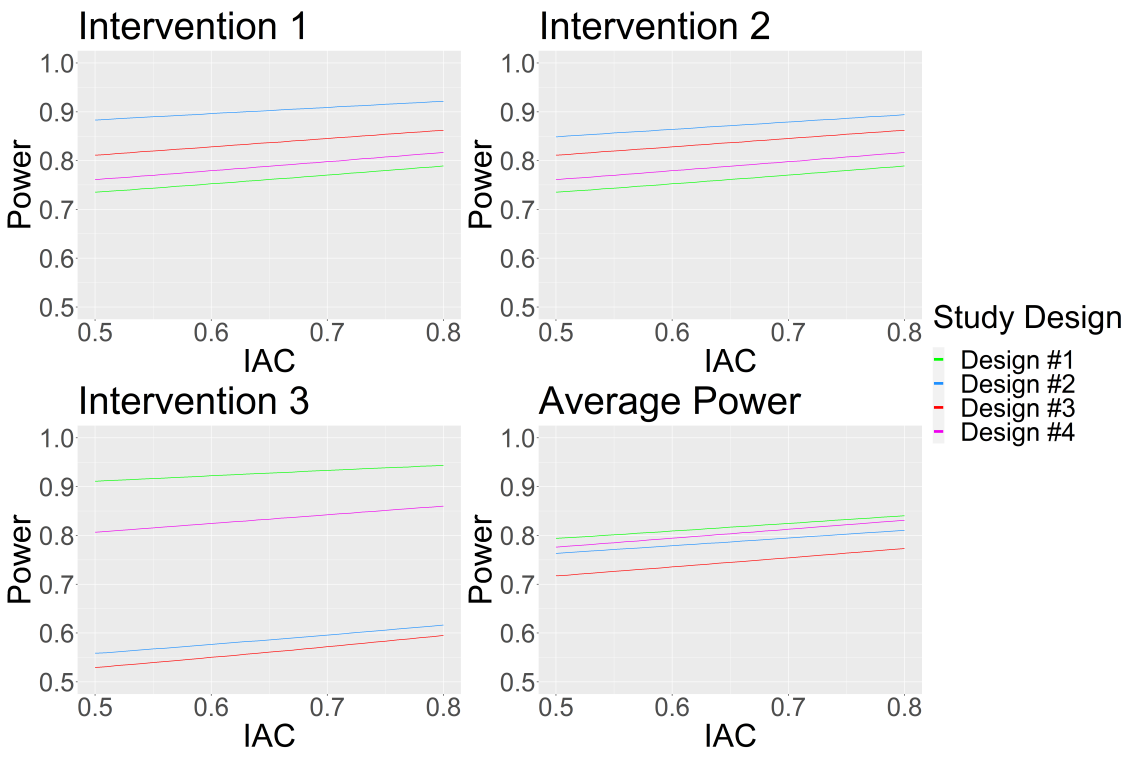


Figure 2.8: Comparison of Power for Each of Three Interventions for Multi-arm Trials

Intervention 3 in time periods 4 and 5. Looking at average power for all three interventions, Design #1 has the highest average power, but for higher π, has about average equal power with Design #4. Design #2 has about 0.03 lower average power compared with Design #1 for all π values shown, and Design #3 has about 0.07 lower average power than Design #1.

2.5 Simulation

We used simulation to verify the power calculations and Type I error rates for all examples in Section 2.4. We simulated 1000 data sets under the alternate hypothesis using representative values of the variance parameters that were allowed to vary to verify power for each example. Linear mixed models were fit to each simulated data set using restricted maximum likelihood as used in other stepped wedge simulation studies (Hooper et al., 2016), using the lme4 package in R (Bates et al., 2015). No small sample size corrections were made. Power was calculated as the percentage of simulations in which the null hypothesis

was rejected using a Wald test at the Bonferroni-corrected type I error level. Type I error rates were estimated by simulating data under the null hypothesis, i.e., setting all treatment effects to 0, and calculating the percentage of simulations in which the null hypothesis was falsely rejected using an experimentwise Type I error rate of 0.05 and Bonferroni corrections as described in the examples.

Tables 2.1 - 2.4 compare power calculated using our method to power estimated by simulation for each set of examples. For all examples, power calculated numerically using our method and simulated power were similar, with no indication of systematic under- or over-estimation of power. Similarly, Type I error rates from the simulations were reasonably close to the nominal level, and did not appear to be systematically over- or under-estimated.

	$\pi = 0.05$		$\pi = 0.35$	
	$\rho_w = 0.05$	$\rho_w = 0.30$	$\rho_w = 0.05$	$\rho_w = 0.30$
Design	$\delta_1 = 0.4$	$\delta_1 = 0.4$	$\delta_1 = 0.4$	$\delta_1 = 0.4$
Single Intervention	.61 (.61)	.71 (.70)	.78 (.75)	.86 (.85)
12-Cluster Concurrent	.73 (.71)	.80 (.79)	.83 (.85)	.93 (.92)
10-Cluster Concurrent	.60 (.60)	.68 (.68)	.77 (.75)	.88 (.85)
Type I error (nominal error = 0.025)				
Single Intervention	.026	.027	.020	.023
12-Cluster Concurrent	.022	.019	.039	.026
10-Cluster Concurrent	.027	.024	.024	.032

Table 2.1: Comparison of Power Based on Simulation and Proposed Method (in parentheses). Single-Intervention and Concurrent Designs in Section 2.4.1.

	$\pi = 0.05$		$\pi = 0.35$	
	$\rho_w = 0.05$	$\rho_w = 0.30$	$\rho_w = 0.05$	$\rho_w = 0.30$
Design	$\delta_1 = 0.4$	$\delta_1 = 0.4$	$\delta_1 = 0.4$	$\delta_1 = 0.4$
12-Cluster Concurrent	.73 (.71)	.80 (.79)	.83 (.85)	.93 (.92)
12-Cluster Late Design	.66 (.65)	.76 (.75)	.79 (.80)	.89 (.89)
10-Cluster Early Design	.71 (.72)	.83 (.79)	.88 (.86)	.94 (.94)
Type I error (nominal error = 0.025)				
12-Cluster Concurrent	.019	.018	.036	.022
12-Cluster Late Design	.018	.024	.028	.016
10-Cluster Early Design	.026	.026	.023	.020

Table 2.2: Comparison of Power Based on Simulation and Proposed Method (in parentheses). Concurrent Designs and Factorial Design in Section 2.4.2.

	$\rho_w = 0.07$, $\rho_a = 0.05$, $\pi = 0.5$			$\rho_w = 0.07$, $\rho_a = 0.05$, $\pi = 0.7$		
	Intvn 1	Intvn 2	Intxn	Intvn 1	Intvn 2	Intxn
Design	$\delta_1 = 0.44$	$\delta_1 = 0.44$	$\delta_3 = 0.22$	$\delta_1 = 0.44$	$\delta_2 = 0.44$	$\delta_3 = 0.22$
1	.83 (.82)	.85 (.82)	.15 (.11)	.85 (.85)	.89 (.85)	.15 (.11)
2	.94 (.94)	.91 (.92)	.28 (.29)	.95 (.96)	.94 (.94)	.31 (.31)
3	.88 (.89)	.89 (.89)	.13 (.10)	.91 (.91)	.92 (.91)	.14 (.11)
4	.87 (.85)	.86 (.85)	.15 (.13)	.90 (.88)	.89 (.88)	.16 (.14)
Type I error (nominal = 0.0167)						
1	.016	.019	.014	.014	.018	.013
2	.011	.013	.015	.009	.014	.014
3	.016	.021	.017	.015	.021	.011
4	.011	.017	.011	.012	.016	.008

Table 2.3: Comparison of Power Based on Simulation and Proposed Method (in parentheses). Factorial Designs in Section 2.4.3.

2.6 Discussion

Stepped wedge designs with more than one intervention are being used in practice despite a paucity of literature on their statistical design and analysis. We have presented power calculation methods for stepped wedge design trials that have multiple interventions, both as multi-arm and factorial designs. We focus on studies that include a relatively small number of clusters, which is common for stepped wedge trials (Taljaard et al., 2016b). In our examples, it was not feasible to explore all possible design options. However, the examples demonstrate several principles. We found that a concurrent design, in which two one-treatment stepped wedge trials are conducted as a single study, is more efficient than two separate one-treatment studies, which is supported by Lyons et al. (Lyons et al., 2017). Our methods enable power calculations for such studies. In concurrent designs, cluster-periods in the control condition perform "double duty" by serving as controls for both treatment conditions. Such trials are essentially three-arm trials in which two interventions are each compared to a control condition.

Our results also illustrate that stepped wedge factorial designs that include cluster-periods in a combined condition can increase power substantially compared to concurrent designs when treatment effects are additive. However, since the presence of an interaction generally

	$\rho_w = 0.07$, $\rho_a = 0.05$, $\pi = 0.5$			$\rho_w = 0.07$, $\rho_a = 0.05$, $\pi = 0.7$		
	Intvn 1	Intvn 2	Intvn 3	Intvn 1	Intvn 2	Intvn 3
Design	$\delta_1 = 0.5$	$\delta_1 = 0.4$	$\delta_3 = 0.4$	$\delta_1 = 0.4$	$\delta_2 = 0.4$	$\delta_3 = 0.4$
1	.75 (.74)	.79 (.74)	.91 (.91)	.79 (.77)	.81 (.77)	.93 (.93)
2	.89 (.88)	.85 (.85)	.58 (.56)	.91 (.91)	.88 (.88)	.62 (.60)
3	.81 (.81)	.81 (.81)	.52 (.53)	.84 (.85)	.85 (.85)	.57 (.57)
4	.80 (.76)	.79 (.76)	.80 (.81)	.83 (.80)	.82 (.80)	.84 (.84)
Type I error (nominal = 0.0167)						
1	.016	.019	.016	.014	.018	.016
2	.011	.013	.010	.009	.014	.009
3	.016	.021	.024	.015	.021	.025
4	.011	.017	.010	.012	.016	.011

Table 2.4: Comparison of Power Based on Simulation and Proposed Method (in parentheses). Multi-arm Designs in Section 2.4.4

decreases power for detecting main effects in factorial designs (Green et al., 2002), power may end up being inadequate if a potential interaction was not taken into account in power calculations. One approach for guarding against this eventuality is to conduct sensitivity analyses that assume some interaction between interventions when designing the study. Our power calculation methods can be used for this purpose.

In some studies, detecting an interaction effect may be of interest. Our work shows that in a stepped wedge factorial design where the aims include detecting an interaction effect, treatment sequencing is critical. We found that in general, designs in which clusters transition from control to single treatment to a combined treatment will be more powerful than designs in which clusters make only one transition, from control to a single treatment or control to combined condition. Such multiple-transition designs allow for more within-cluster comparisons, which are a driving factor in power for stepped wedge trials in general (Hemming et al., 2020).

A common method of handling interactions in factorial designs is to test for an interaction and drop it if it is not significant. This approach has been shown to lead to biased results (Kahan, 2013). We follow Kahan in recommending reporting results both as a factorial design and as a multi-arm analysis, where a condition with multiple treatments is considered as a

separate treatment condition altogether.

Our examples included both repeated cross-sectional and cohort designs. Power for repeated cross-sectional versus cohort designs has been addressed by others (Feldman and McKinlay, 1994); in general, cohort designs have higher power than cross-sectional designs (Hooper et al., 2016; Feldman and McKinlay, 1992). However, there is often a lack of information about parameter values to support power analysis for cohort designs. ICCs are typically reported as the within-time, within-cluster correlation, ρ_w; the across-time, within-cluster correlation ρ_a and individual auto-correlation π are often not reported. Given this lack of information, it may be sensible to make the simplifying assumption that $\rho_w = \rho_a$, which corresponds to the repeated cross-sectional design.

When conducting multiple hypothesis tests in stepped wedge trials with multiple interventions, investigators should consider the need to control the experimentwise type I error rate. We note that when multiple treatment groups are each compared to a common control group, Dunnett's method may be used for experimentwise type I error rate control (Dunnett, 1964). For other multi-arm or factorial designs, there are several possible methods to control for familywise error rate (Soulakova, 2011). In our examples, we used a Bonferroni correction. As the number of hypotheses increases, the familywise error rate may be better addressed using other techniques.

In this dissertation, we focus on SWDs with 2 or 3 treatment conditions. However, our results are generalizable to designs with more interventions. Consider a model with M main effects and B two-way interaction terms, where $M \geq 2$ and $0 \leq B \leq \frac{M(M-1)}{2}$. The variance-covariance matrix of the regression coefficients would be a $(M+B)\times(M+B)$ matrix. The elements of this matrix would have the same form as the elements of the 3×3 matrix in Appendix A for diagonal and off-diagonal elements for both main and interaction effects. Solving for $[(M+B)\times(M+B)]^{-1}$ would yield the variance-covariance matrix for the estimated coefficients. Note that this approach holds for two-way interactions only; higher-order interactions are not considered.

There are several limitations to our work. We use standardized effect sizes. Standardized

effect sizes may be misleading if underlying distributions are skewed (Botta-Dukat, 2016). A more extensive discussion of advantages and disadvantages of simple versus standardized effect sizes is found elsewhere (Botta-Dukat, 2009). We consider continuous outcomes only; further development is needed for non-continuous outcomes, including binary, survival, categorical and count outcomes. In the model we present, the cluster autocorrelation is constrained to be the same for cluster means across time periods, regardless of the length of time between observing cluster level outcomes. This may not be an accurate assumption, as cluster means observed closer in time may be more correlated than those that are farther apart (Hemming et al., 2017). There are models for one treatment SWDs that allow the correlation between cluster means to decay over time (Li, 2019; Grantham et al., 2018; Li et al., 2020). For linear mixed models with a decaying correlation structure, the covariance matrix is a Toeplitz matrix and requires the use of the Trench algorithm to numerically invert (Grantham et al., 2018). We did not include this feature in our work here as we focused on the derivation of closed form variances and covariances of treatment and interaction effects. We only consider complete SWDs. Incomplete designs, in which data are not collected from some clusters in some periods, have been addressed for stepped wedge trials with a single treatment (Hemming et al., 2014; Kasza et al., 2019). Another topic that could be explored further would be determining the minimal designs to yield a certain level of power. By fixing certain parameters, investigators may be interested in knowing the minimum number of clusters, individuals per clusters, or design sequences to obtain a level of power. Finally, we have assumed that treatment effects are instantaneous and do not consider delays in treatment effects, which have been considered for SWDs with a single treatment (Hussey and Hughes, 2007; Li et al., 2020; Hughes et al., 2015). Future work could explore how delays in one or both treatment effects may impact power of main and interaction effects.

CHAPTER 3

Vaccination Outcome via Multistate Modeling in a Stepped Wedge Design

In this chapter, we focus on estimating intervention effects for stepped wedge trials with a vaccination outcome. We first discuss some existing methods that might be applied to analyzing the NORVAX data, and provide rationale for the need for a new modeling approach. We propose a continuous time multistate cure model. Key features of our proposed model are the use of individual-level data, modeling the multiple-treatment design and estimation of treatment effects in terms of completion and initiation percentage point differences.

3.1 Introduction

There are a few other studies that share the goal of the NORVAX study of estimating the effectiveness of interventions to promote HPV vaccination, and some use a stepped wedge design. In a recently proposed SWD trial for increasing HPV vaccination, analysis will be conducted by following fixed cohorts of individuals throughout each step and using random effects logistic regression for analysis (Rutten et al., 2018). However, the NORVAX study has an open cohort design, and we wish to utilize all available information if possible. Thus this approach is not a good fit. Another trial with a stepped wedge design and HPV vaccination outcome conducts statistical analysis using clinic-level outcomes (Carney et al., 2019). We have fit such models, and discuss their strengths and limitations below. Finally, there is a study that uses data at the patient-visit level (Perkins et al., 2020), with the outcome of whether or not a patient who was eligible for a vaccine dose actually received a

dose at the visit. The main outcomes of the NORVAX study are initiation percentage, the percentage of population that has received at least one dose, and completion percentage, the percentage of the population that has received all required doses. This approach does not estimate population-level completion and initiation percentages. Thus, none of these studies use methods that will achieve the goals of the NORVAX study.

One possible approach is to use clinic-level completion or initiation percentages at specific time points (e.g., quarterly) as the observations. The data could then be modeled using linear mixed models. Random effects can be included to account for dependencies of repeated observations from the same clinic. The use of linear models allows for flexibility in specifying the covariance matrix of the error terms. For example, we could allow each clinic to have a unique error variance to account for different trends by clinic. Another possibility would be to incorporating clinic size and allow observations to have size-weighted contributions to the likelihood function.

Percentages are bounded by 0 and 100 and thus could violate the assumption of normality of errors of linear models, particularly if these percentages approach either the upper or lower bound. An alternative approach is to use a beta distribution, which restricts outcomes to take values between a and b ($a < b$). Ferrari and Cribari-Neto propose a modeling technique for beta-distributed outcomes (Ferrari and Cribari-Neto, 2004).

Models with a clinic-level percentage outcome variable have several advantages. Regression coefficients from these models are easily interpreted and provide intervention effects in terms of initiation and completion percentage differences. Another advantage is the abundance of existing software for linear mixed models and beta regression (Cribari-Neto and Zeileis, 2020; Magnusson et al., 2020). Even the more complicated linear mixed models with weightings can be fit using existing software (Pinheiro et al., 2020). Finally, given the small size of the data set, they are computationally inexpensive.

However, the clinic-level modeling approach has significant limitations. Because clinic is the unit of observation, only clinic-level covariates can be included. Thus we cannot adjust for individual-level covariates such as age, gender, and socioeconomic status nor test them

as moderators. In the NORVAX study, we only have seven clinics, and using clinic as the unit of observation means a rather small sample size. Additionally, we would need to choose the frequency for examining the cluster-level outcomes. Should it be annually, quarterly, monthly, or even daily? This is a somewhat arbitrary decision and the impact of different sampling frequencies would need to be explored.

Another possible method for analysis is using individual-level initiation and completion status as the outcome. Mixed effects logistic regression models have been used in parallel study designs with HPV vaccination outcomes (Borg et al., 2018; Fu et al., 2016; Hurley et al., 2019). A possible approach is as follows: create a repeated binary outcome for each individual at specific time points (e.g., quarterly or monthly); fit a mixed effects logistic regression model, with random effects for individuals nested within clinics. Indicators for each time point could be included to indicate intervention condition and to control for secular trends. The individual-level logistic regression framework suffers from the same frequency issue as the clinic-level analysis, as we need to arbitrarily select the time points to examine each individual's outcome. Also, it would be difficult to model time-varying covariates that do not align with selected time points. However, the major issue with using repeated measures of individual-level initiation or completion as the outcome is that these outcomes are not probabilistic; once an individual is initiated, they remain initiated, and once they complete the vaccine series, they remain completed. So by repeating observations of an individual who has already completed or initiated their vaccination, we would be recording observations that cannot change. This violates basic model assumptions. Despite this concern, we attempted to fit such models, and found that models with random intercepts for individuals did not converge.

Due to these limitations, we pursued an alternative modeling framework for the NORVAX study. Our aim was to use individual-level data on vaccination status in a modeling framework that would enable us to estimate intervention effects as changes in initiation and completion percentages. To accomplish the modeling objectives, we propose a multistate cure model.

3.2 Data

Data for the study are extracted from the electronic health records of individual patients. To be considered part of the study-eligible population at any point in time, an individual must be aged 12-17 years and an "active" patient, defined as having had an encounter at any of the seven clinics in the past two years. These criteria create a dynamic cohort, with individuals entering and exiting the study over time as they age in or out and change from being active to inactive patients or vice versa. The dynamic cohort combined with the stepped wedge factorial design means that an individual may contribute data over multiple time periods and intervention conditions.

Another salient feature of the data is that the exact dates of vaccination are known, because they are recorded in the electronic health records. Thus the time points at which patients transition from one state to another (e.g., unvaccinated to one dose) are known.

An example of data from an individual in the study is shown in Figure 3.1. In this figure, the top graphic displays the individual's times of receiving doses in days, and the middle graphic shows the sequence of treatment conditions at the individual's clinic. The individual enters the study with 0 doses at the beginning of Period 1, when his/her clinic is in the usual care condition. The individual continues to have 0 doses when the clinic transitions to the reminder condition at the start of Period 2. At day 700, while the clinic is still in the reminder condition, the individual receives his/her first dose. After the first dose, there are 152 days of ineligibility before they are due for their next dose. The individual remains due for their second dose until they receive it on day 1600, by which time the clinic has transitioned to the combined condition (Period 5).

To address this combination of time-varying exposure and time-varying vaccination state, we break an individual's contribution to the dataset into multiple observations, with a new observation starting whenever the intervention exposure or the individual's state changes. The bottom graphic shows the distinct observations that this individual would contribute to the dataset. Given the nature of these observations, we develop a continuous time multistate

model for the data.

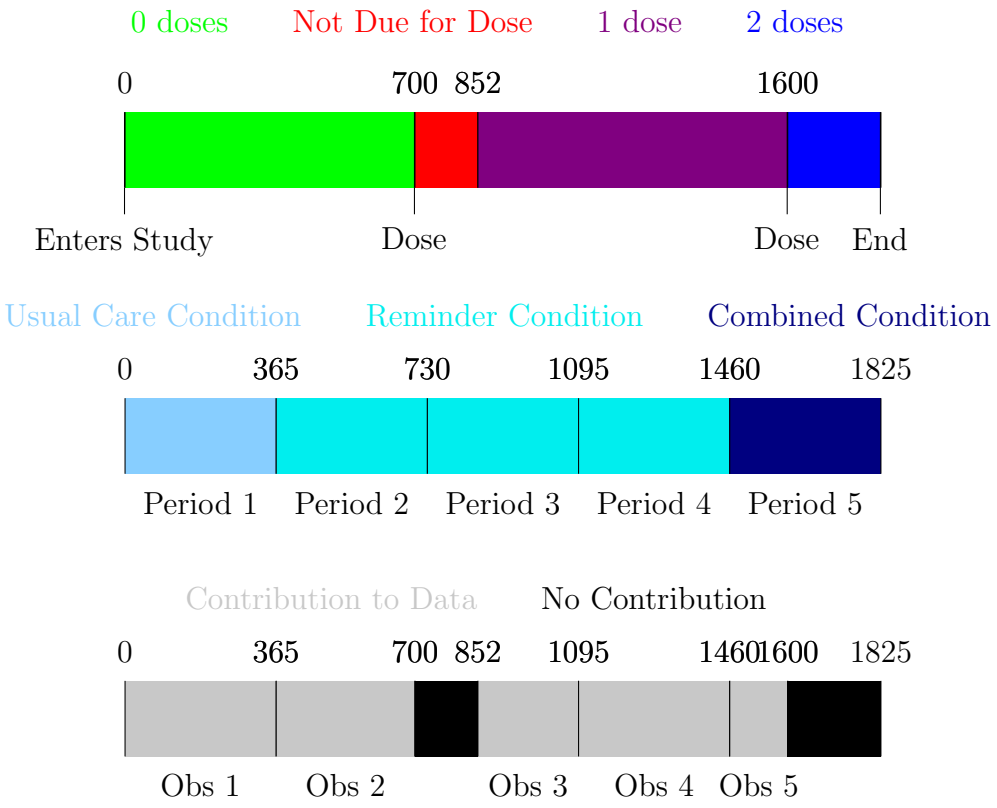


Figure 3.1: HPV dosing example. Time measured in days

3.3 Multistate Models

In multistate models, individuals transition among a finite number of states, and the model parametrizes the rate or intensity of transitions among states. These models have been used in a wide variety of health-related applications, such as illness-death models (von Cube et al., 2017), tumor progression (Wu et al., 2008), and psoriatic arthritis development (O'Keeffe et al., 2017).

One use of such models is to evaluate the effects of interventions. When the model parameterizes transition intensities between states, intervention effects for multistate models can be estimated as hazard ratios. For example, in a multistate model for stroke therapy, states represent increasing levels of disability for recovering stroke patients, and intervention

effects were modeled as reductions in the hazard rates for transitioning to states corresponding to worsened disability (Cassarly et al., 2017). These are other examples of this approach in the literature (Le-Rademacher et al., 2018; Fintzi et al., 2021; Gran et al., 2015). However, hazard ratios can be difficult to interpret in clinically meaningful terms. Estimation of intervention effects on population-level percentages, such as the percentage of individuals in a given state, may be of interest to clinicians and statisticians alike. For example, it would be of interest to quantify the change in the percent of the population that has received at least one dose of the HPV vaccine due to exposure to an intervention.

We propose a Bayesian continuous time multistate cure model for the HPV vaccination outcome using the number of doses as different states within the stepped wedge design. We include a cure model component because not all individuals will receive the next required dose of their HPV vaccine for a variety of reasons (Dilleya et al., 2020). The multistate cure model (also known as mover-stayer model (Yiu et al., 2017)) is an extension to the continuous time multistate model in which a percentage of individuals never transition out of each states (Beesley and Taylor, 2019). We then use parameter estimates from fitting the multistate cure model to estimate intervention effects on population-level HPV vaccination percentages. This model overcomes several of the challenges posed by existing models for both HPV vaccination outcomes and stepped wedge designs. This multistate cure modeling framework 1) allows the model to adequately account for a dynamic cohort 2) incorporates cure proportions to estimate the percentage of individuals who will not receive their next required dose and 3) makes use of individual-level data to model population-level vaccination percentages.

Continuous time multistate models can be parameterized by transition intensities $\lambda_{cd}(t, \mathscr{F}(t))$, the instantaneous probability of transition from state c to state d at time t with filtration $\mathscr{F}(t)$. The transition intensity is defined as

$$\lambda_{cd}(t, \mathscr{F}(t)) = \lim_{\Delta t \to 0} \frac{q_{cd}(t, t+\Delta t)}{\Delta t},$$

where $q_{cd}(t_1, t_2)$ is the transition probability between states c and d during times $t_1 < t_2$. In

this dissertation, we assume semi-Markov processes. Semi-Markov processes feature transition intensities that are independent of the event history contained in $\mathscr{F}(t)$ but dependent on the time spent in the current state. The transition intensity for a semi-Markov process between states c and d is $\lambda_{cd}(t, t-T_c)$ for $t < T_c$, the time of most recent entry into state c. A continuous time multistate process also can be fully defined by the transition probability matrix $\boldsymbol{Q}(t_1, t_2)$, with entries $q_{cd}(t_1, t_2)$, $t_1 < t_2$. Let $\boldsymbol{S}$ be a state space with s states and the state process $Y(t)$ take a value from the set $c, d \in \{0, \ldots, s-1\}$ at time t. For a semi-Markov process, the entries in the transition probability matrix are

$$q_{cd}(t_1, t_2) = P\left(Y(t_2) = d \mid Y(t_1) = c; t_1 - T_c\right) \text{ for } t_1 < t_2.$$

For the HPV vaccination trial, the states of the multistate model are the current number of doses received, with $s = 3$ states corresponding to 0, 1, or 2 doses. We ignore the possibility of third doses because our data include very few individuals who become due for or receive a third dose. We write the transition intensity matrix for the 3-state model as

$$\boldsymbol{\Lambda(t)} = \begin{bmatrix} -\lambda_{01}(t) & \lambda_{01}(t) & 0 \\ 0 & -\lambda_{12}(t) & \lambda_{12}(t) \\ 0 & 0 & 0 \end{bmatrix}.$$

Individuals may enter the study in any of the three states. In this model, the transition intensity λ_{01} corresponds to the rate at which individuals without any doses receive the first dose. Because individuals must receive 1 dose before receiving 2 doses, an individual must go through State 1 before reaching State 2 i.e. $\lambda_{02} = 0$. Individuals in each state only have one possible transition, making our model a progressive multistate model (Hsieh et al., 2002). When an individual receives their first dose, they cannot return to having received 0 doses, i.e. $\lambda_{10} = 0$. Finally, State 2, in which an individual has received 2 doses, is an absorbing state, and all transition intensities in the final row of $\boldsymbol{\Lambda(t)}$ are 0. A feature of transition intensity matrices is that its rows sum to 0; thus we have $\lambda_{00}(t) = -\lambda_{01}(t)$ and

$\lambda_{11}(t) = -\lambda_{12}(t)$, i.e., the distribution of sojourn times for States 0 and 1 are equivalent to the distribution of time it takes to transition to States 1 and 2, respectively. The progressive nature of our multistate model also leads to $\lambda_{01}(t)$ and $\lambda_{12}(t)$ having the form of parametric hazard functions in our model.

For multistate models in which exact transition times are unknown, likelihood formulations are constructed using transition probabilities rather than transition intensities (Kalbfleisch and Lawless, 1985) and rely on a Markov assumption for mathematically tractability (Kay, 1986). In our application, we observe exact transition times between states and do not rely on the Markov assumption. Instead, we can directly model the sojourn times, i.e., the time spent in a state, and construct the likelihood using a time-to-event framework. We assume Weibull distributions for sojourn times and therefore transition times as well. A Weibull distributed random variable T for the time spent in state c before transitioning to state d has probability density function (pdf)

$$f(t;\ \alpha_{cd}, \gamma_{cd}) = \frac{\alpha_{cd}}{\gamma_{cd}} \left(\frac{t}{\gamma_{cd}}\right)^{\alpha_{cd}-1} \exp\left(-\left(\frac{t}{\gamma_{cd}}\right)^{\alpha_{cd}}\right), \quad \alpha_{cd} > 0, \gamma_{cd} > 0, t > 0,$$

where γ_{cd} is a scale parameter and α_{cd} is a shape parameter. Transition intensities λ_{01} and λ_{12} now have the hazard and survival functions for the Weibull distribution as

$$h(t;\ \alpha_{cd}, \gamma_{cd}) = \frac{\alpha_{cd}}{\gamma_{cd}} \left(\frac{t}{\gamma_{cd}}\right)^{\alpha_{cd}-1} \qquad S(t;\ \alpha_{cd}, \gamma_{cd}) = \exp\left(-\left(\frac{t}{\gamma_{cd}}\right)^{\alpha_{cd}}\right).$$

We reparametrize the scale parameter γ_{cd} using a proportional hazards framework to incorporate covariates, defining

$$\gamma_{cd} = \exp\left(-\frac{\gamma_{cd0} + \boldsymbol{X}\boldsymbol{\beta}_{cd} + \boldsymbol{u}}{\alpha_{cd}}\right), \tag{3.1}$$

where γ_{cd0} is an intercept, $\boldsymbol{X}$ is a design matrix and $\boldsymbol{\beta_{cd}}$ is the vector of regression coefficients corresponding to the transition between states c and d. In our model, we estimate two sets of transition intensities, λ_{01} and λ_{12}. We assume the same set of predictors $\boldsymbol{X}$ across both

transitions but allow regression coefficients to differ for each transition intensity. The model also includes a clinic-level random intercept for unobserved heterogeneity where u_i is an instance of $\boldsymbol{u} = \{u_1, \ldots, u_I\} \sim N(0, \sigma_i^2)$ and $\boldsymbol{I_I}$ is an $I \times I$ identity matrix for I number of clinics. This clinic-level effect is assumed to be the same across transitions.

Models for stepped wedge designs typically include indicators for each time period to account for secular trends (Hussey and Hughes, 2007). In our model, every regression coefficient vector $\boldsymbol{\beta}_{cd}$ contains $(j-1)$ binary {0,1} indicators periods $2, \ldots j$, with Period 1 used as reference. The two columns in the design matrix corresponding to indicators for the two interventions are denoted as $\boldsymbol{X}_{tx}$. The interventions are modeled as time-varying {0,1} indicators, per the time intervals in the study design in Figure 1.2. Clinics only switch intervention conditions at the start of a time period. If individual k in clinic i enters into a time interval in which an intervention is implemented at the clinic, that individual's value of the intervention indicator becomes 1 for all time spent in that clinic-interval.

3.3.1 Cure Proportions

It is expected that some patients will never receive their next dose of HPV vaccine and thus will remain indefinitely in State 0 (no doses) or State 1 (one dose). To address this, we introduce cure proportions into our model. Being "cured" means not receiving a dose and thus not experiencing a transition to the next state.

Let V_{ck} be a latent indicator variable with $V_{ck} = 1$ indicating that individual k is "cured" and will stay in state c and $V_{ck} = 0$ indicating that individual k is "non-cured" and will eventually leave state c. Let $\pi_{ck}(\boldsymbol{Z_k}) = P(V_{ck} = 1 \mid \boldsymbol{b_c}, \boldsymbol{Z_k})$ be the probability of individual k being cured in state c, which depends on the regression coefficient vector $\boldsymbol{b_c}$ and covariate values $\boldsymbol{Z_k}$. We use a logistic link for π_{ck} to incorporate covariates,

$$\pi_{ck} = \frac{\exp(b_{c0} + \boldsymbol{b_c'Z_k})}{1 + \exp(b_{c0} + \boldsymbol{b_c'Z_k})}, \tag{3.2}$$

where b_{c0} corresponds to an intercept for the cure proportion in state c. We assume that

covariates in $\boldsymbol{Z_k}$ are measured at the individual level, and allow for different effects of these covariates depending on state, i.e., for states $c \neq d$, the vectors $\boldsymbol{b_c}$ and $\boldsymbol{b_d}$ may not necessarily be equal. The survival function for each transition is only defined when $V_{ck} = 0$ for an individual in state c. Using a mixture cure model framework, the conditional survival function $S_{ck}(t)$ for individual k in state c at time t becomes

$$S_{ck}(t) = \pi_{ck} + (1 - \pi_{ck})\, S_{ck0}(t \mid V_{ck} = 0),$$

where $S_{ck0}(t \mid V_{ck} = 0)$ is a proper Weibull survival function and $S_{ck}(t)$ is no longer proper as $\lim_{t \to \infty} S_{ck}(t) = \pi_{ck}$.

3.3.2 Likelihood and Bayesian Formulation

Our population contains individuals with both known and unknown cure status. For an individual who is observed leaving state c, the cure indicator $V_{ck} = 0$. We write the observed likelihood for an individual k with $V_{ck} = 0$ who experiences a transition out of state c into state d at time t as

$$L_{ck} = (1 - \pi_{ck}) \prod_{j=j_1}^{j_r} S_{cjk0}(t_{1jk}, t_{2jk};\ \boldsymbol{X}) \{h_{cjk}(t; \boldsymbol{X})\}^{\delta_{ckj}},$$

where the individual is observed during time intervals j_1 to j_r, S_{cjk0} and h_{cjk} are the proper survival function and hazard function, respectively, of the sojourn time of state c during interval j for individual k who survives from times t_{1jk} to t_{2jk} dependent on set of covariates $\boldsymbol{X}$. The $\{0, 1\}$ censoring indicator δ_{ckj} is equal to 1 if individual k leaves state c during interval j at time t and 0 otherwise. For every individual, the time in state c starts at $t_{1j} = 0$ for $j = j_1$.

The individuals who do not leave state c are a mixture of those who are cured and those who are uncured and have not experienced a transition yet. For an individual k who does

not leave state c, the likelihood contribution is

$$L_{ck} = \pi_{ck} + (1-\pi_{ck}) \prod_{j=j_1}^{j_r} S_{cjk0}(t_{1jk}, t_{2jk};\ \boldsymbol{X}),$$

allowing for the possibility to be either cured or not cured. Time-dependent covariates are represented as a series of left-truncated survival times, which may also be right-censored (Austin, 2012).

An example of likelihood construction is found in Figure 3.2. The individual enters the study in State 0 at calendar day 0 and is censored in State 1 at calendar day 950. This individual has covariate value changes at times 365 and 730 days. The individual receives their first dose after age 15 and therefore must wait at least 28 days until s/he is eligible for the next dose. This wait time is not included in the likelihood. For sets of time-varying covariates $\boldsymbol{X} \neq \boldsymbol{X}^* \neq \boldsymbol{X}^{**}$, this individual's likelihood contributions for each transition would be

$$L_{0k} = (1-\pi_{0k}) \left(\frac{S_{0j_1k0}(0, 365;\ \boldsymbol{X}) S_{0j_2k0}(0, 600;\ \boldsymbol{X}^*)}{S_{0j_2k0}(0, 365;\ \boldsymbol{X}^*)} \right) h_{0j_2k}(600; \boldsymbol{X}^*) \text{ and}$$

$$L_{1k} = \pi_{1k} + (1-\pi_{1k}) \left(\frac{S_{1j_2k0}(0, 102;\ \boldsymbol{X}^*) S_{1j_3k0}(0, 322;\ \boldsymbol{X}^{**})}{S_{1j_3k0}(0, 102;\ \boldsymbol{X}^{**})} \right).$$

Optimizing the likelihood function with the inclusion of both clinic-level random intercept and cure model parameters would prove challenging. We therefore use a Bayesian framework for estimating the parameters of the model. Cure models often have been formulated using the expectation-maximization (EM) algorithm. However, Bayesian techniques are a viable alternative to the EM algorithm for the estimation of cure models (Ma et al., 2020), with several applications seen in practice (Yu and Tiwari, 2012; Wang et al., 2020).

For our 3-state multistate cure model, the parameters of interest include Weibull parameters $\gamma_{010}, \alpha_{01}, \gamma_{120}$ and α_{12}, fixed effect regression coefficients $\boldsymbol{\beta}$, clinic-level hierarchical components $\boldsymbol{u}$ and σ_I, and cure parameters b_{00}, b_{10}, $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$. The joint posterior of

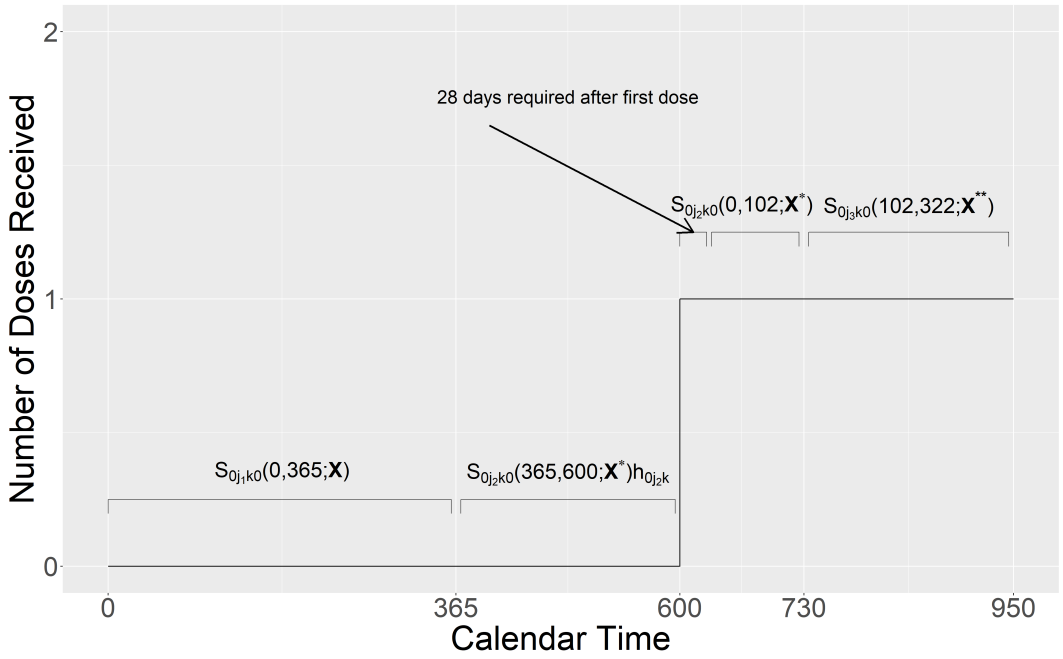


Figure 3.2: Likelihood Example for an individual who receives their first dose after age 15.

these parameters is analytically intractable, and so we conduct posterior estimation using the Monte Carlo Markov Chain (MCMC) technique. In particular, we use the no-U-turn sampler (NUTS), a variant of Hamiltonian Monte Carlo, via STAN to estimate posterior distributions (Stan Development Team, 2022). The target average acceptance probability (adapt_delta parameter in Stan) is set to 0.95, which results in a smaller sampling step size to reduce the frequency of divergent transition after warmup.

We specified diffuse prior distributions for all parameters, as there is not much prior information about these parameters. We expect that our data will provide enough information to estimate each parameter. For covariates in $\boldsymbol{\beta_{cd}}$, we specify a normal prior distributions of $\sim N(0, 2)$. The random effect u_i had a prior distribution of $N(0, 1)$, and for the rest of the covariates, we use default prior distributions in STAN, which are flat, improper uniform distributions from $(-\infty, \infty)$ (Stan Development Team, 2022). Prior distributions for each parameter are assumed independent. To assist convergence, boundaries were placed on cure intercepts $-5 \leq b_{00} \leq 4$ and $-5 \leq b_{01} \leq 4$, corresponding to cure probabilities between 0.6% and 98%. Further constraints were placed on Weibull scale and shape parameters and σ_I to

ensure non-negativity. For each model we present, 8,000 samples were drawn with a burn-in of 4,000 iterations. R-hat values are close to 1, which while not guaranteeing convergence, is an indicator of adequate model fit. Our models also show good effective sample size for all parameters. Posterior analysis was conducted in R version 3.6.1 (R Core Team, 2019) with the RStan package (Stan Development Team, 2020).

3.4 Population Level Percentage Estimates

The model estimates intervention effects as hazard ratios. An objective of the NORVAX study is to estimate intervention effects in terms of differences in study population-level initiation and completion percentages attributable to interventions. This requires solving for the transition probability matrix $\boldsymbol{Q}$.

3.4.1 Solving for Transition Probabilities

For a multistate model, the relationship between the transition probability matrix $\boldsymbol{Q}$ and transition intensity matrix $\boldsymbol{\Lambda}$ between times t_1 and t_2 is defined by the Kolmogorov Forward Equation (KFE). The initial condition is $\boldsymbol{Q}(0,0) = \boldsymbol{I_s}$, where $\boldsymbol{I_s}$ is an $s \times s$ identity matrix. If transition intensities are assumed to be constant over time (i.e. time homogeneity), a solution to the KFE is the matrix exponential $\boldsymbol{Q}(t) = \text{Exp}(t\boldsymbol{\Lambda})$ (Clements, 2019). If the transition intensities in $\boldsymbol{\Lambda}(t)$ are non-homogeneous and a function of t, the KFE must be solved directly, which can be difficult. Solving for transition probabilities has been explored extensively for non-homogeneous Markov models (Titman, 2011) and non-Markov models (Titman, 2015). Obtaining such solutions also has been explored in a frequentist setting using B-splines (Titman, 2011).

Solving for transition probabilities is less challenging for progressive multistate models. Hsieh et al. (2002) provide closed form solutions of transition probabilities for a three-state progressive multistate model, which can also be derived from Cook & Lawless (Cook and Lawless, 2018). For a 3-state model with states denoted as $s = \{0, 1, 2\}$, the transition

probability matrix has entries

$$p_{00}(t_1,t_2) = \exp[-H_0(t_1;t_2)] \qquad p_{01}(t_1,t_2) = \int_{t_1}^{t_2} \exp[-H_0(t_1;t)]h_{01}(t)\exp[-H_1(t;t_2)]dt$$

$$p_{02}(t_1,t_2) = 1-p_{00}(t_1,t_2)-p_{01}(t_1,t_2) \;\; p_{11}(t_1,t_2) = \exp[-H_1(t_1;t_2)]$$

$$p_{12}(t_1,t_2) = 1-p_{11}(t_1,t_2) \tag{3.3}$$

where $H_0(t_1;t_2)$ and $H_1(t_1;t_2)$ are cumulative hazard functions for an individual between the time t_1 that an individual enters the study and some time $t_2 > t_1$, for States 0 and 1, respectively. The term $h_{01}(t)$ is the hazard function associated with the transition intensity between States 0 and 1 evaluated at time t. To evaluate the integral in $p_{01}(t_1,t_2)$, Hsieh et al. (2002) use a trapezoidal approximation; we use a Monte Carlo estimator. Let $f(t) = \exp[-H_0(t_1;t)]h_{01}(t)\exp[-H_1(t;t_2)]$ and $F = \int_{t_1}^{t_2} f(t)dt$. We approximate F with a Monte Carlo estimator

$$\hat{F} = (t_2-t_1)\frac{1}{N}\sum_{i=1}^{N} f(X_i), \quad X_i \sim \text{Unif}(t_1,t_2).$$

There is a specific amount of time after receiving the first dose before an individual is eligible for the second dose. We denote this required time as t^*. For $t_2-t_1 \leq t^*$, we let $p_{00} = \exp[-H_0(t_1,t_2)]$, $p_{01} = 1-p_{00}$, and $p_{02} = 0$. For $t_2-t_1 > t^*$, we let

$$p_{01}(t_1,t_2) = (1-\exp(-H_0(t_2-t^*,t_2)))+\int_{t_1}^{t_2-t^*} \exp[-H_0(t_1,t)]h_{01}(t)\exp[-H_1(t,t_2-t^*)]dt,$$

where p_{01} is now the sum of the probability of either 1) receiving the first dose up until t_2-t^*, in which a patient could theoretically have received both doses during that time or 2) receiving the first dose in the most recent t^* days, in which that individual would not be eligible for the second dose.

The transition probability matrix $\boldsymbol{Q(t_1,t_2)}$ with cure proportions for an individual between times t_1 and t_2 is now

$$\begin{bmatrix} (1-\pi_0)p_{00}(t_1,t_2)+\pi_0 & (1-\pi_0)p_{01} & (1-\pi_0)p_{02}(t_1,t_2) \\ 0 & (1-\pi_1)p_{11}(t_1,t_2)+\pi_1 & (1-\pi_1)p_{12}(t_1,t_2) \\ 0 & 0 & 1 \end{bmatrix},$$

with entries defined as

$$\begin{bmatrix} p^*_{00}(t_1,t_2) & p^*_{01}(t_1,t_2) & p^*_{02}(t_1,t_2) \\ 0 & p^*_{11}(t_1,t_2) & p^*_{12}(t_1,t_2) \\ 0 & 0 & 1 \end{bmatrix}.$$

The transition probabilities in the matrix $\boldsymbol{Q(t_1,t_2)}$ are all functions of the transition intensities in $\boldsymbol{\Lambda}(t)$ in Equation 3.3. To obtain values for the elements of $\boldsymbol{\Lambda}(t)$, we use Bayes estimators from the posterior distributions. Bayes estimators are estimates from a posterior distribution that minimize the posterior expected loss value with respect to a loss function (Samaniego, 2010). Common Bayes estimators include posterior mean, median or modes. Posterior distributions for both simulated and real data did not yield substantial differences in these estimators. We use posterior means for populating the elements of $\boldsymbol{\Lambda}(t)$, corresponding to minimizing the posterior expected value with a mean squared error loss function.

3.4.2 Intervention Effect on Study Population-Level Percentage Outcomes

For the NORVAX study, we seek to estimate study population-level initiation and completion percentages. We denote the study population-level initiation and completion percentages at time t as $\zeta_i(t)$ and $\zeta_c(t)$, respectively. In the multistate framework, initiation corresponds to the percentage of individuals in States 1 or 2, or equivalently, the percentage of individuals not in State 0. Completion corresponds to the percentage of individuals in State 2. Estimates of these outcomes at time t can be formulated as

$$\zeta_i(t) = \frac{\sum_{k=1}^{N_0(t)}(1-p^*_{00}(t_{1k},t))+N_1(t)+N_2(t)}{\sum_{s=0}^{S=2}N_s(t)}$$

$$\zeta_c(t) = \frac{\sum_{k=1}^{N_0(t)}(p^*_{02}(t_{1k}, t)) + \sum_{k=1}^{N_1(t)}(p^*_{12}(t_{1k}, t)) + N_2(t)}{\sum_{s=0}^{S=2} N_s(t)} \tag{3.4}$$

where t_{1k} is the calendar time an individual enters the study and $N_s(t)$ denotes the number of individuals who entered the study in state s and are study-eligible at time t.

We obtain a model-based estimate of the effect of a treatment on study population-level initiation and completion percentage outcomes as follows. We first estimate $\zeta_c(t)$ and $\zeta_i(t)$ by estimating transition probabilities for all individuals at some time t, given the observed intervention conditions. We then set $\boldsymbol{X}_{tx} = 0$, corresponding to the absence of intervention, and estimate $\zeta_c(t)$ and $\zeta_i(t)$ again. One can set either one or both columns of $\boldsymbol{X}_{tx}$ to be $\mathbf{0}$, to estimate the effects of each intervention or the combined effect. In our model, intervention effects are assumed to be additive. Study population-level completion and initiation values with either column of $\boldsymbol{X}_{tx}$ set to $\mathbf{0}$ are denoted $\zeta^*_c(t)$ and $\zeta^*_i(t)$, respectively. The intervention effect on either outcome is then quantified as $\zeta_c(t)$ - $\zeta^*_c(t)$ for completion and $\zeta_i(t)$ - $\zeta^*_i(t)$ for initiation.

3.5 Simulation

We conducted a simulation study to examine the performance of the model in quantifying intervention effects. For the simulation study, we used the patient population of the NORVAX study and retained all patient characteristics, including study entry time, the state a patient entered the study, and demographic information. Clinic membership was also fixed for each patient, which dictates treatment sequencing based on the stepped wedge design presented in Figure 1.2 across $I = 7$ clinics. We simulated cure indicators $V_{ck} \sim Bernoulli(\pi_{ck})$ for all individuals and set the time to next dose to infinity for those individuals who are cured in their current state. The models for the cure proportions π_{ck} include an intercept b_{c0} and a single covariate for gender. For every non-cured individual, we simulate time to receive their next dose based on the model presented in Equation (3.1). Covariates in the time-to-event model are indicators for reminder and clinic-based interventions, time trend indicators for each time period j and gender. Because the NORVAX study is ongoing and

our simulated data mimic these data, individuals are censored at the end of Period 4, corresponding to the most recent data pull. We used the inverse cumulative density function (CDF) method to simulate from a piecewise Weibull distribution with cure proportions, with details in Appendix B. A single dataset was simulated with $N = 26430$ unique patients from the NORVAX study. We fit the Bayesian multistate cure model described in Section 3.3.2. Initiation and completion outcomes are estimated with posterior means using Equation 3.4. Intervention effects are then estimated by setting $\boldsymbol{X_{tx}} = 0$ and calculating the differences between $\zeta_i - \zeta_i^*$ and $\zeta_c - \zeta_c^*$.

Due to difficulties in estimating cure proportions when times to event are long, we first conducted a simulation with short time-to-event values to ensure the model was estimating parameters accurately. In this simulation, individuals are observed beyond their 18th birthday and 2-year visit interval, which is in contrast to the study inclusion/exclusion criteria. Results shown in Table 3.1. For the transition between 0 and 1 dose, the 95% credible intervals all contain the true values. This is also the case for the transition between 1 and 2 doses, but posterior means are further from true values. This can be explained by the higher percentage of censoring for this transition. Censoring in State 1 occurs for many individuals who enter the study in State 1 during the later part of the study and for individuals who enter the study in State 0, receive a dose, and are censored in State 1 at the end of the study.

The second simulation more closely mirrors the NORVAX data. Both time to receive dose and cure proportions were increased from the previous simulation, and exit times were based on the study inclusion and exclusion criteria. If a simulated dose occurred after a patient's 18^{th} birthday or beyond 2 years from an encounter at the health system, that individual was censored. Results are shown in Table 3.2 in the column labeled "Simulation 1". In this simulated data, all credible intervals for the intervention effects contain the true value. However, posterior intervals for several of the intercepts, including γ_{01} and b_{10}, and some covariates effects do not include the true value. We attribute the bias to several factors. The study features a large number of individuals entering the study throughout the duration of the study. Late entry into the study leads to shortened observed times and biased survival

	Parameter	True Value	Model Value
0 to 1 dose	Transition Intensity		
	γ_{01}	5.10	5.08 (4.97, 5.18)
	α_{01}	0.90	0.89 (0.86, 0.93)
	Period 2	0.10	0.04 (-0.11, 0.19)
	Period 3	0.50	0.42 (0.23, 0.61)
	Period 4	0.30	0.23 (-0.05, 0.50)
	Male (transition intensity)	-0.10	-0.05 (-0.16, 0.07)
	Reminder	-0.40	-0.34 (-0.51, -0.17)
	Clinic-based	-0.60	-0.63 (-0.85, -0.42)
	Cure Proportion		
	b_{00}	-1.00	-1.05 (-1.18,-0.93)
	Male (cure)	-0.20	-0.19 (-0.39, 0.01)
1 to 2 dose	Transition Intensity		
	γ_{12}	5.10	5.18 (5.10, 5.24)
	α_{12}	1.10	1.10 (1.07, 1.13)
	Period 2	0.10	0.07 (-0.02, 0.15)
	Period 3	0.50	0.45 (0.33, 0.57)
	Period 4	0.30	0.16 (-0.01, 0.34)
	Male	-0.10	0.02 (-0.04, 0.08)
	Reminder	-0.40	-0.43 (-0.54, -0.33)
	Clinic-based	-0.60	-0.47 (-0.60, -0.34)
	Cure Proportion		
	b_{10}	-1.00	-0.89 (-1.00, -0.81)
	Male	-0.10	-0.02 (-0.14, 0.10)
	σ_c^2	0.02	0.02 (0.00, 0.04)

Table 3.1: Simulation Multistate Results. Covariate effects for time-to-event model are reported as log hazard ratios

parameters, particularly intercepts (Betensky and Mandel, 2015).

Another reason for biased estimates comes from the high percentage of right-censored observations. Right-censoring in NORVAX comes from several sources. The study eligibility criteria contribute substantially to individuals being right-censored; according to the eligibility criteria, individuals exit the study upon reaching their 18^{th} birthday, and are considered "inactive" patients and removed from the study if they go two years without a health encounter. The second criteria is particularly noteworthy, as individuals who leave the system are likely a combination of 1) individuals who may have moved out of the geographic area and 2) individuals who simply have not come back for health visits in two years. It is well-

	Parameter	True Value	Sim. 1	Sim. 2
0 to 1 dose	Transition Intensity			
	γ_{01}	6.20	5.82 (5.61, 6.03)	5.88 (5.80, 6.13)
	α_{01}	0.80	0.82 (0.78, 0.85)	0.81 (0.80, 0.86)
	Period 2	0.10	0.25 (0.11, 0.40)	0.26 (0.21, 0.43)
	Period 3	0.50	0.86 (0.68, 1.06)	0.98 (0.89, 1.25)
	Period 4	0.30	0.42 (0.17, 0.67)	0.67 (0.51, 1.12)
	Male	-0.10	0.04 (-0.14, 0.24)	0.09 (0.00, 0.35)
	Reminder Intvn	-0.40	-0.36 (-0.52, -0.20)	-0.24 (-0.31, -0.01)
	Clinic Intvn	-0.60	-0.78 (-0.98, -0.58)	-0.78 (-0.88, -0.49)
	Cure Proportion			
	b_{00}	-0.20	-0.05 (-0.21, 0.09)	-0.20 (-0.27, -0.01)
	Male (cure)	-0.70	-0.28 (-0.48, -0.08)	-0.23 (-0.32, 0.05)
1 to 2 dose	Transition Intensity			
	γ_{12}	5.90	5.89 (5.72, 6.05)	5.89 (5.84, 6.05)
	α_{12}	1.10	1.12 (1.10, 1.14)	1.11 (1.10, 1.13)
	Period 2	0.10	0.13 (0.08, 0.18)	0.13 (0.11, 0.18)
	Period 3	0.50	0.42 (0.35, 0.49)	0.40 (0.37, 0.48)
	Period 4	0.30	0.17 (0.06, 0.27)	-0.08 (-0.13, 0.08)
	Male	-0.05	0.00 (-0.04, 0.03)	-0.01 (-0.02, 0.03)
	Reminder Intvn	-0.40	-0.40 (-0.47, -0.32)	-0.39 (-0.42, -0.31)
	Clinic Intvn	-0.60	-0.55 (-0.63, -0.47)	-0.46 (-0.50, -0.36)
	Cure Proportion			
	b_{10}	-0.70	-0.51 (-0.56, -0.46)	-0.54 (-0.56, -0.49)
	Male	0.05	0.00 (-0.07, 0.06)	-0.01 (-0.03, 0.07)
	σ_I	0.10	0.05 (0.01, 0.20)	0.05 (0.02, 0.19)

Table 3.2: Simulation Multistate Results. Covariate effects for time-to-event model are reported as log (hazard ratios)

known that censored observations in time-to-event studies can bias estimates of the Weibull survival parameters, both for likelihood maximization and Bayesian techniques (Ducrosa and Pamphile, 2018). Heavy censoring has also been shown to introduce bias into cure models, especially when estimating intercepts (Lin and Huang, 2019).

It has been observed that cure models are biased when censored observations are not observed long enough to differentiate between non-cured and cured individuals (Kearns et al., 2021; Stedman et al., 2014). When the population is a mixture of cured and non-cured individuals, censored observations can belong to either group. Heavy censoring has been shown to bias cure model parameter estimates, and there is no straightforward relationship

between censoring patterns and patterns of bias for estimates of either survival or cure parameters (Othus et al., 2020). Further discussion of identifiability in cure models has been explored elsewhere (Hanin and Huang, 2014). A non-parametric method for estimating cure proportions has been developed (Escobar-Bach et al., 2021). However, this method is not unbiased and maintains consistency only under certain restrictions. Further, including covariates in this nonparametric estimation approach would be challenging. It has been observed that no modeling or prior selection can overcome insufficient follow-up time for estimation of cure proportions, unless the cure proportion is known ahead of time (Felizzi et al., 2021).

To address the issues of late entry and insufficient followup time for estimating cure proportions, we conducted another simulation restricting the sample to individuals who entered the study in the first two time periods, thus allowing for a longer observation time. Results from fitting the model to these simulated data are shown in the "Simulation 2" column of Table 3.2. These results show improvements in the estimation of cure and survival intercepts, and slightly worse estimation of certain covariate effects including intervention effects.

Table 3.3 shows estimates of study population-level initiation and completion percentages, calculated using posterior means in Equation 3.4, at three time points, corresponding to the end of Periods 2, 3 and 4. Columns labeled "Observed" initiation and completion are the observed percentage of individuals at time t who are study eligible and have received at least one dose or two doses, respectively, in the simulated data. "Anticipated values" are the model-based estimates initiation and completion percentages calculated using the specified simulation parameter values. We obtained model-based estimates of population percentages, using results from both "Simulation" 1 and 2.

"Anticipated" initiation and completion percentages were close to the observed percentages at all time points. For Simulation 1, the estimated initiation percentage was higher than the observed percentage, and completion percentage was underestimated at all time points. Initiation and completion percentages were estimated more accurately using Simulation 2.

The poorer performance of Simulation 1 can be largely attributed to the heavier censoring, leading to estimation of a shorter survival time for transitioning from State 0 to 1, and an overestimation of the cure intercept for transitioning from States 1 to 2.

	Initiation $\zeta_i(t)$				Completion $\zeta_c(t)$			
Time	Observed	Ant. Values	Sim. 1	Sim. 2	Observed	Ant. Values	Sim. 1	Sim. 2
730 days	82.9%	83.2%	83.7%	83.7%	38.3%	38.8%	36.4%	36.9%
1095 days	81.5%	82.2%	82.2%	82.0%	40.0%	40.4%	38.4%	38.8%
1460 days	81.7%	82.2%	82.9%	82.0%	43.5%	44.0%	42.2%	43.5%

Table 3.3: Simulation Results: Study-Population Percentages

To estimate intervention effects on initiation and completion percentages, we fit the model in Equation 3.4 to obtain ζ_i^* and ζ_c^* by setting $\boldsymbol{X}_{tx} = 0$, for both Simulation 1 and Simulation 2. Results are shown in Table 3.4. Intervention effects were overestimated for initiation and underestimated for completion. These biases can be partially attributed to biased estimates of intervention effects from the multistate model; for example, posterior estimates for the log hazard ratio for the clinic-based intervention from States 1 to 2 for Models 1 and 2 were -0.55 and -0.46, respectively, compared to true value of -0.60. However, the main source of bias is from estimation of the intercepts of both time-to-event and cure parameters. Both Simulations 1 and 2 underestimate survival times for transitioning from State 0 to 1, leading to larger estimated intervention effects in initiation. Cure percentages are overestimated for the transitions from States 1 to 2 in both Simulations 1 and 2 compared to true value. Larger estimated cure percentage leads to underestimation of the true intervention effects on completion.

3.6 Application to NORVAX Data

We applied the proposed multistate cure model to data from the NORVAX study using the individual-level time to receipt of HPV doses as the observations. Based on the biases observed in the simulation study, we fit models to different subsets of the data. Dataset 1 uses all data from the start of the study through the end of Period 4 and uses the study

	Initiation $\zeta_i(t)$		
Time (Days)	Observed	Simulation 1	Simulation 2
730	82.9%	83.7%	83.7%
1095	81.5%	82.2%	82.0%
1460	81.7%	82.9%	82.0%
	Completion $\zeta_c(t)$		
Time (Days)	Observed	Simulation 1	Simulation 2
730	38.3%	36.4%	36.9%
1095	40.0%	38.4%	38.8%
1460	43.5%	42.2%	43.5%

Table 3.4: Simulation Population Intervention Effects. Columns labeled "Reminder" set the reminder intervention log hazard ratio to 0, columns labeled "Clinic" set the clinic-based intervention log hazard ratio to 0, and columns labeled "Both" set the log hazard ratio for both interventions to 0.

eligibility criteria regarding age and active patient status. To address the issue of late entry and insufficient followup time for estimating cure proportions, Dataset 2 only includes individuals who enter the study in the first two years, corresponding to Periods 1 and 2. Model 3 restricts to the same individuals but also relaxes the eligibility criteria. Dataset 3 continues to observe individuals beyond their 18^{th} birthday and does not stop observing individuals if they have gone more than two years without an encounter at a clinic. Relaxing these criteria reduces the amount of early censoring and only allows individuals to be censored at the end of the study.

Table 3.5 shows results from fitting the model to the different subsets of data. In all cases, both interventions are associated with a negative log hazard ratio for both transition intensities, corresponding to reducing the time to receive either dose. Dataset 1 estimates a shorter time-to-event but higher cure proportion for the transition between States 0 and 1 compared with Dataset 2. For the transition between States 1 and 2, survival times and cure proportions are estimated to be slightly lower in Dataset 2 compared with Model 1. Dataset 3 estimates shorter survival times but higher cure proportions for both transition intensities. In this subset, individuals who would have been censored are observed for longer amounts of time and contribute more information to the cure proportion part of the model.

	Parameter	Data. 1	Data. 2	Data. 3
0 to 1	Transition Intensity			
	γ_{01}	6.43 (6.22, 6.66)	6.90 (6.44, 7.54)	5.76 (5.61, 5.91)
	α_{01}	0.78 (0.74, 0.81)	0.73 (0.68, 0.78)	0.81 (0.77, 0.85)
	Period 2	-0.45 (-0.62, -0.28)	-0.42 (-0.60, -0.23)	-0.24 (-0.40, -0.07)
	Period 3	-0.15 (-0.35, 0.05)	0.69 (0.34, 1.04)	0.61 (0.31, 0.92)
	Period 4	-0.86 (-1.11, -0.61)	-0.22 (-0.73, 0.28)	0.21 (-0.30, 0.75)
	Male	-0.17 (-0.4, 0.05)	-0.44 (-0.97, 0.06)	-0.18 (-0.37, 0.00)
	Reminder Intvn	-0.27 (-0.43, -0.10)	-0.27 (-0.51, -0.04)	-0.44 (-0.67, -0.21)
	Clinic Intvn	-0.02 (-0.23, 0.18)	-0.28 (-0.66, 0.10)	-0.01 (-0.41, 0.40)
	Cure Proportion			
	b_{00}	-0.08 (-0.29, 0.09)	-0.69 (-2.38, -0.03)	1.66 (1.56, 1.75)
	Male	-0.12 (-0.34, 0.11)	0.20 (-0.50, 1.53)	-0.45 (-0.57, -0.33)
1 to 2	Transition Intensity			
	γ_{12}	5.92 (5.84, 6.01)	5.91 (5.83, 5.99)	5.78 (5.71, 5.85)
	α_{12}	1.10 (1.08, 1.12)	1.06 (1.03, 1.09)	1.03 (1.00, 1.05)
	Period 2	-0.12 (-0.19, -0.05)	-0.14 (-0.21, -0.06)	-0.17 (-0.25, -0.10)
	Period 3	0.30 (0.21, 0.39)	0.37 (0.24, 0.49)	0.30 (0.18, 0.41)
	Period 4	0.08 (-0.05, 0.21)	-0.15 (-0.35, 0.05)	-0.20 (-0.39, -0.01)
	Male	-0.03 (-0.09, 0.03)	-0.04 (-0.10, 0.03)	-0.03 (-0.08, 0.03)
	Reminder Intvn	-0.14 (-0.23, -0.05)	-0.19 (-0.29, -0.08)	-0.17 (-0.26, -0.07)
	Clinic Intvn	-0.27 (-0.37, -0.17)	-0.24 (-0.39, -0.10)	-0.24 (-0.37, -0.11)
	Cure Proportion			
	b_{10}	-0.59 (-0.67, -0.53)	-0.59 (-0.67, -0.52)	-0.03 (-0.09, 0.04)
	Male	0.07 (-0.02, 0.16)	0.03 (-0.07, 0.14)	0.01 (-0.08, 0.10)
	σ_I	0.07 (0.01, 0.16)	0.06 (0.01, 0.15)	0.03 (0.00, 0.10)

Table 3.5: Model Applied to NORVAX Data

Observed and model-based estimates of initiation and completion percentages are shown in Table 3.6. For Datasets 1 and 2, observed percentages are based on the full eligible study population, whereas for Data 3, observed percentages are based on the relaxed criteria. In general, the observed and model-estimated percentages are closest for Data 3. This is consistent with the expectation that less censoring will lead to more accurate parameter estimates.

Model-based estimates for the effects of the two interventions on initiation and completion percentages are shown in Table 3.7. Overall, the estimated intervention effects are small, in most cases less than one percentage point. Dataset 2 estimates larger intervention effects for the reminder intervention for both initiation and completion at all times. This may

Table 3.6: NORVAX Study: Study-Population Level Percentages

With censoring from study criteria

	Initiation ζ_i			Completion ζ_c		
Time	Observed	Dataset 1	Dataset 2	Observed	Dataset 1	Dataset 2
End Period 2	79.0%	82.5%	83.2%	54.4%	54.8%	55.7%
End Period 3	81.0%	83.0%	82.3%	55.6%	53.6%	53.8%
End Period 4	85.2%	84.6%	84.1%	60.1%	53.8%	54.9%

Without study criteria censoring

	Initiation ζ_i		Completion ζ_c	
Time	Observed	Dataset 3	Observed	Dataset 3
End of Period 2	68.9%	69.8%	46.0%	47.8%
End of Period 3	71.2%	71.7%	51.8%	52.8%
End of Period 4	72.4%	72.6%	54.8%	55.1%

be attributed to an estimated lower cure proportion for transitioning out of State 0 and a larger estimated effect of the reminder intervention for the transition between States 0 and 1. For the clinic-based intervention, we see a higher intervention effect from Dataset 2 for the initiation percentage and a higher intervention effect from Dataset 1 for the completion percentage. The small intervention effect for the clinic-based intervention on initiation in Dataset 1 (0.02% at the end of the third period and 0.13% at the end of the fourth period) is attributable to the small estimated log hazard ratio for the transition intensity between States 0 and 1 (-0.02, hazard ratio = 0.98). For the clinic-based intervention, the difference in intervention effects on completion are not very different between Models 1 and 2 across all time points.

Estimates from the model fit to Dataset 3 are the most accurate in when compared to observed completion and initiation percentages, as seen in Table 3.6, but estimates lower intervention effects compared to Datasets 1 and 2. The smaller intervention effects can be attributed to the higher estimated cure proportion for both transitions, particularly from States 0 to 1.

		Initiation Tx Effect ζ_i-ζ_i^*			Completion Tx Effect ζ_i-ζ_i^*		
Time	Dataset	Reminder	Clinic	Both	Reminder	Clinic	Both
End of Period 2	Dataset 1	0.24%	–	0.24%	0.25%	–	0.25%
	Dataset 2	0.29%	–	0.29%	0.36%	–	0.36%
	Dataset 3	0.10%	–	0.10%	0.14%	–	0.14%
End of Period 3	Dataset 1	0.24%	0.02%	0.26%	0.33%	0.51%	0.84%
	Dataset 2	0.30%	0.19%	0.49%	0.44%	0.46%	0.89%
	Dataset 3	0.10%	0.00%	0.11%	0.15%	0.18%	0.33%
End of Period 4	Dataset 1	0.32%	0.13%	0.46%	0.70%	1.14%	1.83%
	Dataset 2	0.68%	0.76%	1.42%	0.92%	1.13%	2.04%
	Dataset 3	0.17%	0.05%	0.22%	0.24%	0.30%	0.54%

Table 3.7: Estimates of Intervention Effects on Initiation and Completion Percentages. Columns labeled "Reminder" set the reminder intervention log hazard ratio to 0, columns labeled "Clinic" set the clinic-based intervention log hazard ratio to 0, and columns labeled "Both" set the log hazard ratio for both interventions to 0. Estimates are percentage point change in initiation and completion attributable to the interventions.

3.7 Discussion

We have proposed a Bayesian multistate cure model for a HPV vaccination promotion trial using a stepped wedge design. Parameter estimates from the multistate cure model can be converted to transition probabilities and these transition probabilities can be used to quantify intervention effects as changes in study population-level initiation and completion percentages. Our modeling approach is a novel contribution that allows for use of individual-level time-to-dose data to estimate intervention effects on study population-level initiation and completion percentages.

The estimated intervention effects in our data application are smaller than what was projected when the NORVAX study was designed. The COVID-19 pandemic began midway through Period 3 and impacted the number of individuals receiving vaccinations. The number of patients receiving their next dose has been increasing in the study clinics but as of the end of Period 4, it still has not recovered to pre-pandemic levels. It is hoped with more data collected in the future that we see stronger intervention effects.

In our model, we only allowed for time-varying covariates that change values in alignment with changes in study design periods, such as the implementation of interventions. However,

this could be generalized to allow for covariates to change values at other time periods; the solution involves breaking the observation time of individuals into other segments with constant covariate values. We also defined study population-level initiation and completion outcomes in Section 3.4 that are specific to our HPV study, but other outcomes could be defined based on a percentage of individuals in certain states for other studies.

The simulation study showed that heavy censoring and insufficient follow-up time can lead to biased estimates of parameters in the multistate cure model. The patients in the NORVAX study have relatively long time intervals between doses. When there are long average times-to-dose, it is necessary to observe patients for a long duration to get more accurate estimates of cure probabilities. Estimating intervention effects on study population initiation and completion relies on accurate estimation of survival and cure model intercepts, which can be biased with large amounts of censoring and the inability to distinguish between cured and non-cured individuals.

The NORVAX study uses a factorial design and includes a condition in which individuals can receive both interventions simultaneously. We fit models with interactions terms in preliminary work. Estimates of an interaction effect were very small and thus we did not include an interaction term in the final model. As more data are collected, an interaction between the two interventions may materialize and necessitate estimation in the model.

There are several limitations to the proposed model in addition to the censoring and follow-up challenges. We used fixed, time-invariant regression coefficients in the time-to-event model in Equation 3.1, but one could model more complex relationships with time. The time-to-event model can incorporate time dependency by replacing the vector of regression coefficients in Equation 3.1 with $\boldsymbol{\beta_{cd}}(t)$. We also fixed the covariate vector in the formulation of the cure proportion with respect to time. The cure proportion could incorporate time-dependent covariates by substituting in $\boldsymbol{b_c}(t)$ in Equation 3.2. Models with time-varying coefficients and cure proportions have been explored elsewhere (Dong et al., 2021; Beretta and Heuchenne, 2019), as have models with time-varying covariates incorporated into the cure proportion (Dirick et al., 2019). Inclusion of time-varying coefficients in both time-to-event

and cure model would make estimation of the transition probabilities more difficult. Finally, we assumed a common clinic-level random intercept u_i across all transitions. The model could be extended to accommodate different hierarchical structures, such as individuals nested within clinics or transition-specific clinic-level intensities.

CHAPTER 4

Discussion

In this dissertation, we have presented methods to address several gaps in the stepped wedge design literature. We first proposed methods for conducting power calculations for SWDs with multiple interventions and a continuous outcome, which has been published (Sundin and Crespi, 2022). We then presented a multistate cure model that was applied to model the number of doses for HPV vaccine received by patients in a stepped wedge design trial. Parameter estimates from the multistate cure model were then used to quantify intervention effects as changes in the percentage of study population that have initiated or completed their HPV vaccine regimen.

For power calculation for SWDs with multiple interventions, several areas of future work are discussed in Section 2.6. One area of further interest is continued development of methods for power for SWD trials with non-continuous outcomes. There is sparse literature for design and analysis of SWD trials with non-continuous outcomes and a single intervention, let alone multiple interventions. Standard errors for intervention effects can be difficult to calculate analytically for models with non-continuous outcomes; power calculations for such trials currently tend to use simulation (Barker et al., 2017).

Combining the two methodological developments proposed in this dissertation for future projects would provide an exciting, challenging opportunity to conduct power calculations for studies using multistate cure models. Power and sample size methods have been developed for proportional hazards mixture cure models (Wang et al., 2012) and Markov multistate models (Cassarly et al., 2017). Neither of these approaches consider stepped wedge designs, nor do they consider semi-Markov models. It would also be of interest to conduct power

calculations on changes in study population percentages due to interventions instead of hazard ratios acting on transition intensities.

Further areas of work for the multistate model are discussed in Section 3.7. A prominent area for future work is the issue of censoring and insufficient followup times for mixture cure models. These features of the data lead to biased parameter estimatation that cannot be easily overcome using current statistical methods (Bernhardt, 2016; Jiang et al., 2017). Existing non-parametric estimators that attempt to overcome heavily censored data are not unbiased and do not allow for inclusion of covariates (Escobar-Bach et al., 2021). Cure models have ample room for further methodological study, particularly for estimating cure proportions with censored observations that are not easily identifiable as cured or not.

Appendix A - Derivation of Standard Errors

Derivation of Standard Errors of Treatment Effect Estimates

The variance-covariance matrix of the fixed effects in a linear mixed model is found by taking the inverse of $\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}$ where $\boldsymbol{Z}$ is the fixed effects design matrix and $\boldsymbol{V}$ is the variance-covariance matrix of the outcome variable. In model (2.4), $\boldsymbol{Z}$ is the $IT\times(T+3)$ design matrix and $\boldsymbol{V}$ is the $IT\times IT$ variance-covariance matrix of the outcome. Here, we find closed form expressions for the variance-covariance matrix of the treatment effect estimates, $\hat{\theta}_1, \hat{\theta}_2,$ and $\hat{\theta}_3$ for this model by first finding an expression for $\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}$ and then using block matrix inversion techniques to get the desired elements of $(\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z})^{-1}$. We use the repeated cross-sectional model. For the nested exchangeable and cohort models, the appropriate elements of $\boldsymbol{V}$ can be substituted.

Defining the Precision Matrix

Assuming that clusters are independent, for the repeated cross-sectional model, the matrix $\boldsymbol{V}$ has block diagonal structure with elements $\boldsymbol{V_i} = \sigma_c^2\boldsymbol{I_T}+\sigma_\alpha^2\boldsymbol{1_T}\boldsymbol{1'_T}$, where $\boldsymbol{I_T}$ is a $T\times T$ identity matrix and $\boldsymbol{1_T}$ is a $T\times 1$ vector of 1's. Due to the block diagonal structure of $\boldsymbol{V}$, we can write the precision matrix as

$$\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z} = \sum_{i=1}^{I}\boldsymbol{Z_i'}\boldsymbol{V_i}^{-1}\boldsymbol{Z_i}.$$

Using the Sherman-Morrison formula (Sherman and Morrison, 1949; Bartlett, 1951) for the inverse of a matrix of this form, we obtain

$$\boldsymbol{V_i}^{-1} = \frac{1}{\sigma_c^2(\sigma_c^2+T\sigma_\alpha^2)}\left[(\sigma_c^2+T\sigma_\alpha^2)\boldsymbol{I_T}-\sigma_\alpha^2\boldsymbol{1_T}\boldsymbol{1}_{\boldsymbol{T}}^{'}\right].$$

The submatrix $\boldsymbol{Z_i}$ is the $T\times(T+3)$ subset of $\boldsymbol{Z}$ corresponding to cluster i. Thus

$$\boldsymbol{Z_i'}\boldsymbol{V_i}^{-1}\boldsymbol{Z_i} = \frac{1}{\sigma_c^2(\sigma_c^2+T\sigma_\alpha^2)}\left[(\sigma_c^2+T\sigma_\alpha^2)\boldsymbol{Z_i'}\boldsymbol{Z_i}-\sigma_\alpha^2\boldsymbol{Z_i'}\boldsymbol{1_T}\boldsymbol{1}_{\boldsymbol{T}}^{'}\boldsymbol{Z_i}\right]. \tag{4.1}$$

In the following, the vectors $\boldsymbol{X}, \boldsymbol{W},$ and $\boldsymbol{XW}$ denote the columns of the design matrix corresponding to treatment 1, treatment 2, and the interaction term, respectively. Let $\boldsymbol{X_i}$

be the $(T\times 1)$ vector that corresponds to cluster i, and $\boldsymbol{X_{i,-T}}$ be the $(T-1)\times 1$ vector for cluster i that does not include the value of $\boldsymbol{X_i}$ at time T. We use similar notation for $\boldsymbol{W_i}$ and $\boldsymbol{W_{i,-T}}$ for treatment 2, and we use $\boldsymbol{XW_i}$ and $\boldsymbol{XW_{i,-T}}$ for the interaction. Using the summation of submatrices in (4.1), we can write the precision matrix $\boldsymbol{Z'V^{-1}Z}$ as a $(T+3)\times(T+3)$ symmetric matrix whose lower triangular elements are

$$\boldsymbol{Z'V^{-1}Z} = \begin{bmatrix} Tf & & & & & \\ f\mathbf{1_{T-1}} & (f+gT)\boldsymbol{I_{T-1}}-g\mathbf{1_{T-1}}\mathbf{1'_{T-1}} & & & & \\ y_1 & \sum_{i=1}^{I}\frac{\boldsymbol{X'_{i,-T}}}{\sigma_c^2}-\sigma_\alpha^2 h_1\mathbf{1'_{T-1}} & l_1-z_1 & & & \\ y_2 & \sum_{i=1}^{I}\frac{\boldsymbol{W'_{i,-T}}}{\sigma_c^2}-\sigma_\alpha^2 h_2\mathbf{1'_{T-1}} & q_1 & l_2-z_2 & & \\ y_3 & \sum_{i=1}^{I}\frac{\boldsymbol{(XW)'_{i,-T}}}{\sigma_c^2}-\sigma_\alpha^2 h_3\mathbf{1'_{T-1}} & q_2 & q_3 & l_3-z_3 & \end{bmatrix}.$$

The term $\mathbf{1_{T-1}}$ is a $(T-1)\times 1$ vector of 1's and $\boldsymbol{I_{T-1}}$ is a $(T-1)$ identity matrix, and we define

$$a=\frac{1}{\sigma_c^2+T\sigma_\alpha^2},\quad b=\frac{1}{\sigma_c^2},\quad c=ab,$$

$$X^T=\sum_{j=1}^{T}X_{ij},\quad W^T=\sum_{j=1}^{T}W_{ij},\quad (XW)^T=\sum_{j=1}^{T}X_{ij}W_{ij},$$

$$X^{IT}=\sum_{i=1}^{I}\sum_{j=1}^{T}X_{ij},\quad W^{IT}=\sum_{i=1}^{I}\sum_{j=1}^{T}W_{ij},\quad (XW)^{IT}=\sum_{i=1}^{I}\sum_{j=1}^{T}X_{ij}W_{ij},$$

$$f=Ia,\quad g=Ic\sigma_\alpha^2,$$

$$y_1=aX^{IT},\quad y_2=aW^{IT},\quad y_3=a(XW)^{IT},$$

$$h_1=cX^{IT}\quad h_2=cW^{IT},\quad h_3=c(XW)^{IT},\quad z_1=c\sigma_\alpha^2\sum_{i=1}^{I}(X^T)^2,$$

$$z_2=c\sigma_\alpha^2\sum_{i=1}^{I}\left(W^T\right)^2,$$

$$z_3=c\sigma_\alpha^2\sum_{i=1}^{I}\left((XW)^T\right)^2,\quad l_1=bX^{IT},\quad l_2=bW^{IT},\quad l_3=b(XW)^{IT},$$

$$q_1=l_3-c\sigma_\alpha^2\sum_{i=1}^{I}\left(X^T\right)\left(W^T\right),\quad q_2=l_3-c\sigma_\alpha^2\sum_{i=1}^{I}\left((XW)^T\right)\left(X^T\right),$$

$$q_3 = l_3 - c\sigma_\alpha^2 \sum_{i=1}^{I} \left((XW)^T\right)\left(W^T\right).$$

The terms q_2 and q_3 make use of the relationship $X_{ij}^2 = X_{ij}$ and $W_{ij}^2 = W_{ij}$.

Blocking the Precision Matrix

We partition the precision matrix into a 2×2 block matrix where $(\boldsymbol{Z'V^{-1}Z})_{11}$ is the $T\times T$ submatrix, $(\boldsymbol{Z'V^{-1}Z})_{21} = (\boldsymbol{Z'V^{-1}Z})'_{12}$ is the $T\times 3$ submatrix and $(\boldsymbol{Z'V^{-1}Z})_{22}$ is the 3×3 submatrix corresponding to the precision of the parameters of interest: $\hat{\theta}_1, \hat{\theta}_2,$ and $\hat{\theta}_3$. Using block matrix inversion (Lu and Shiou, 2000),

$$(\boldsymbol{Z'V^{-1}Z})_{22}^{-1} = \left((\boldsymbol{Z'V^{-1}Z})_{22} - (\boldsymbol{Z'V^{-1}Z})_{21}(\boldsymbol{Z'V^{-1}Z})_{11}^{-1}(\boldsymbol{Z'V^{-1}Z})_{12}\right)^{-1}.$$

We first obtain $(\boldsymbol{Z'V^{-1}Z})_{11}^{-1}$ using another variation of block matrix inversion (Lu and Shiou, 2000) and Schur complements, yielding

$$(\boldsymbol{Z'V^{-1}Z})_{11}^{-1} = \frac{1}{(f+gT)}\begin{bmatrix} \frac{(g+f)}{f} & -\mathbf{1'_{T-1}} \\ -\mathbf{1_{T-1}} & (\boldsymbol{I_{T-1}} + \mathbf{1_{T-1}1'_{T-1}}) \end{bmatrix}.$$

Let $\boldsymbol{B} = \boldsymbol{I_{T-1}} + \mathbf{1_{T-1}1'_{T-1}}$ and $\boldsymbol{M} = (\boldsymbol{Z'V^{-1}Z})_{21}(\boldsymbol{Z'V^{-1}Z})_{11}^{-1}(\boldsymbol{Z'V^{-1}Z})_{12}$. We can find

$$\boldsymbol{M} = \frac{1}{(f+gT)}\begin{bmatrix} y_1 & \sum_{i=1}^{I} \frac{\boldsymbol{X'_{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1'_{T-1}} \\ y_2 & \sum_{i=1}^{I} \frac{\boldsymbol{W'_{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1'_{T-1}} \\ y_3 & \sum_{i=1}^{I} \frac{\boldsymbol{(XW)'_{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1'_{T-1}} \end{bmatrix} \begin{bmatrix} \frac{(g+f)}{f} & -\mathbf{1'_{T-1}} \\ -\mathbf{1_{T-1}} & \boldsymbol{B} \end{bmatrix} * \begin{bmatrix} y_1 & \sum_{i=1}^{I} \frac{\boldsymbol{X'_{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1'_{T-1}} \\ y_2 & \sum_{i=1}^{I} \frac{\boldsymbol{W'_{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1'_{T-1}} \\ y_3 & \sum_{i=1}^{I} \frac{\boldsymbol{(XW)'_{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1'_{T-1}} \end{bmatrix}'$$

$$= \frac{1}{(f+gT)} \begin{bmatrix} y_1 & \sum_{i=1}^{I} \frac{\boldsymbol{X}_{\boldsymbol{i},-\boldsymbol{T}}'}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}}' \\ y_2 & \sum_{i=1}^{I} \frac{\boldsymbol{W}_{\boldsymbol{i},-\boldsymbol{T}}'}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}_{\boldsymbol{T-1}}' \\ y_3 & \sum_{i=1}^{I} \frac{(\boldsymbol{XW})_{\boldsymbol{i},-\boldsymbol{T}}'}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1}_{\boldsymbol{T-1}}' \end{bmatrix} *$$

$$\begin{bmatrix} \frac{g+f}{f} y_1 - \mathbf{1}_{\boldsymbol{T-1}}' \left(\sum_{i=1}^{I} \frac{\boldsymbol{X}_{\boldsymbol{i},-\boldsymbol{T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}} \right) & -y_1 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{X}_{\boldsymbol{i},-\boldsymbol{T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}} \right) \\ \frac{g+f}{f} y_2 - \mathbf{1}_{\boldsymbol{T-1}}' \left(\sum_{i=1}^{I} \frac{\boldsymbol{W}_{\boldsymbol{i},-\boldsymbol{T}}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}_{\boldsymbol{T-1}} \right) & -y_2 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{W}_{\boldsymbol{i},-\boldsymbol{T}}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}_{\boldsymbol{T-1}} \right) \\ \frac{g+f}{f} y_3 - \mathbf{1}_{\boldsymbol{T-1}}' \left(\sum_{i=1}^{I} \frac{(\boldsymbol{XW})_{\boldsymbol{i},-\boldsymbol{T}}}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1}_{\boldsymbol{T-1}} \right) & -y_3 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{(\boldsymbol{XW})_{\boldsymbol{i},-\boldsymbol{T}}}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1}_{\boldsymbol{T-1}} \right) \end{bmatrix}'.$$

In the second matrix in this product, we simplify the term $-\mathbf{1}_{\boldsymbol{T-1}}' \left(\sum_{i=1}^{I} \frac{\boldsymbol{X}_{\boldsymbol{i},-\boldsymbol{T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}} \right)$, and similarly simplify similar terms in the other rows, by rewriting this term as

$$= \left(\sum_{i=j}^{T-1} \sum_{i=1}^{I} \frac{X_{ij}}{\sigma_c^2} - h\sigma_\alpha^2 \mathbf{1}_{\boldsymbol{T-1}}{}' \mathbf{1}_{\boldsymbol{T-1}} \right) = \left(\sum_{i=j}^{T-1} \sum_{i=1}^{I} \frac{X_{ij}}{\sigma_c^2} - \sigma_\alpha^2 (T-1) \sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{\sigma_c^2 (\sigma_c^2 + T\sigma_\alpha^2)} \right)$$

$$= \left(\sum_{i=j}^{T} \sum_{i=1}^{I} \frac{\sigma_c^2 X_{ij} + T\sigma_\alpha^2 X_{ij}}{\sigma_c^2 (\sigma_c^2 + T\sigma_\alpha^2)} - \frac{\sigma_\alpha^2 T X_{ij} - \sigma_\alpha^2 X_{ij}}{\sigma_c^2 (\sigma_c^2 + T\sigma_\alpha^2)} \right) - \sum_{i=1}^{I} \frac{X_{iT}}{\sigma_c^2}$$

$$= \left(\sum_{i=j}^{T} \sum_{i=1}^{I} \frac{X_{ij}}{(\sigma_c^2 + T\sigma_\alpha^2)} \right) + \left(\sum_{i=j}^{T} \sum_{i=1}^{I} \frac{\sigma_\alpha^2 X_{ij}}{\sigma_c^2 (\sigma_c^2 + T\sigma_\alpha^2)} \right) - \sum_{i=1}^{I} \frac{X_{iT}}{\sigma_c^2}$$

$$= y_1 + \sigma_\alpha^2 h_1 - \sum_{i=1}^{I} \frac{X_{iT}}{\sigma_c^2}.$$

Substituting these simplifications back into $\boldsymbol{M}$ yields $\boldsymbol{M} = \frac{1}{(f+gT)} \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix}$ with elements

$$m_{11} = y_1 \left(\frac{g+f}{f} y_1 + \left(-y_1 - \sigma_\alpha^2 h_1 + \sum_{i=1}^{I} \frac{X_{iT}}{\sigma_c^2} \right) \right) +$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{X}_{\boldsymbol{i},-\boldsymbol{T}}'}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}}' \right) \left(-y_1 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{X}_{\boldsymbol{i},-\boldsymbol{T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}} \right) \right)$$

$$m_{12} = y_1 \left(\frac{g+f}{f} y_2 + \left(-y_2 - \sigma_\alpha^2 h_2 + \sum_{i=1}^{I} \frac{W_{iT}}{\sigma_c^2} \right) \right) +$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{X}_{\boldsymbol{i},-\boldsymbol{T}}'}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}}' \right) \left(-y_2 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{W}_{\boldsymbol{i},-\boldsymbol{T}}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}_{\boldsymbol{T-1}} \right) \right)$$

$$m_{13} = y_1 \left(\frac{g+f}{f} y_3 + \left(-y_3 - \sigma_\alpha^2 h_3 + \sum_{i=1}^{I} \frac{(XW)_{iT}}{\sigma_c^2} \right) \right) +$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{X}'_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}'_{\boldsymbol{T-1}} \right) \left(-y_3 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{(\boldsymbol{XW})_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1}_{\boldsymbol{T-1}} \right) \right)$$

$$m_{21} = y_2 \left(\frac{g+f}{f} y_1 - \left(y_1 + \sigma_\alpha^2 h_1 - \sum_{i=1}^{I} \frac{X_{iT}}{\sigma_c^2} \right) \right) +$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{W}'_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}'_{\boldsymbol{T-1}} \right) \left(-y_1 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{X}_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}} \right) \right)$$

$$m_{22} = y_2 \left(\frac{g+f}{f} y_2 - \left(y_2 + \sigma_\alpha^2 h_2 - \sum_{i=1}^{I} \frac{W_{iT}}{\sigma_c^2} \right) \right) +$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{W}'_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}'_{\boldsymbol{T-1}} \right) \left(-y_2 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{W}_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}_{\boldsymbol{T-1}} \right) \right)$$

$$m_{23} = y_2 \left(\frac{g+f}{f} y_3 - \left(y_3 + \sigma_\alpha^2 h_3 - \sum_{i=1}^{I} \frac{(XW)_{iT}}{\sigma_c^2} \right) \right) +$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{W}'_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}'_{\boldsymbol{T-1}} \right) \left(-y_3 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{(\boldsymbol{XW})_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1}_{\boldsymbol{T-1}} \right) \right)$$

$$m_{31} = y_3 \left(\frac{g+f}{f} y_1 - \left(y_1 + \sigma_\alpha^2 h_1 - \sum_{i=1}^{I} \frac{X_{iT}}{\sigma_c^2} \right) \right) +$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{XW}'_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1}'_{\boldsymbol{T-1}} \right) \left(-y_1 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{X}_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}} \right) \right)$$

$$m_{32} = y_3 \left(\frac{g+f}{f} y_2 - \left(y_2 + \sigma_\alpha^2 h_2 - \sum_{i=1}^{I} \frac{W_{iT}}{\sigma_c^2} \right) \right) +$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{XW}'_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1}'_{\boldsymbol{T-1}} \right) \left(-y_2 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{W}_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}_{\boldsymbol{T-1}} \right) \right)$$

$$m_{33} = y_3 \left(\frac{g+f}{f} y_3 - \left(y_3 + \sigma_\alpha^2 h_3 - \sum_{i=1}^{I} \frac{(XW)_{iT}}{\sigma_c^2} \right) \right) +$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{XW}'_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1}'_{\boldsymbol{T-1}} \right) \left(-y_3 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{(\boldsymbol{XW})_{\boldsymbol{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_3 \mathbf{1}_{\boldsymbol{T-1}} \right) \right).$$

Simplifying the diagonal elements of $\boldsymbol{M}$

The three diagonal elements of $\boldsymbol{M}$ have the same form, so we work with m_{11} and apply the form to m_{22} and m_{33}. Let $\eta_1 = y_1 + \sigma_\alpha^2 h_1 - \sum_{i=1}^{I} \frac{X_{iT}}{\sigma_c^2}$. Then we can rewrite the diagonal term m_{11} as

$$= \frac{1}{(f+gT)} \left(\frac{y_1^2(f+g)}{f} - 2y_1\eta_1 + \left(\sum_{i=1}^{I} \frac{\boldsymbol{X'_{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \boldsymbol{1'_{T-1}} \right) \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \boldsymbol{1_{T-1}} \right) \right).$$

We begin by expanding the term $\left(\sum_{i=1}^{I} \frac{\boldsymbol{X'_{i,-T}}}{\sigma_c^2} - \frac{l_1 - y_1}{T} \boldsymbol{1'_{T-1}} \right) \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \frac{l_1 - y_1}{T} \boldsymbol{1_{T-1}} \right)$

$$= \left(\sum_{i=1}^{I} \frac{\boldsymbol{X'_{i,-T}}}{\sigma_c^2} - \frac{l_1}{T} \boldsymbol{1'_{T-1}} \right) \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \frac{l_1}{T} \boldsymbol{1_{T-1}} \right) +$$

$$\frac{y_1}{T} \boldsymbol{1'_{T-1}} \boldsymbol{B} \frac{y_1}{T} \boldsymbol{1_{T-1}} + 2 \frac{y_1}{T} \boldsymbol{1'_{T-1}} \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \frac{l_1}{T} \boldsymbol{1_{T-1}} \right).$$

We simplify each of these three summands. We will use the relationship $\sum_{i=1}^{I} \sum_{j=1}^{T-1} \frac{X_{ij}}{\sigma_c^2} = l_1 - \sum_{i=1}^{I} \frac{X_{iT}}{\sigma_c^2}$. Moving left to right, we simplfiy by the first summand by

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{X'_{i,-T}}}{\sigma_c^2} - \frac{l_1}{T} \boldsymbol{1'_{T-1}} \right) \boldsymbol{B} \left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \frac{l_1}{T} \boldsymbol{1_{T-1}} \right)$$

$$= \sum_{j=1}^{T-1} \left(\sum_{i=1}^{I} \frac{X_{ij}}{\sigma_c^2} \right)^2 + l_1^2 - 2l_1 \sum_{i=1}^{I} \frac{X_{iT}}{\sigma_c^2} + \left(\sum_{i=1}^{I} \frac{X_{iT}}{\sigma_c^2} \right)^2 - 2l_1^2 + 2l_1 \sum_{i=1}^{I} \frac{X_{ij}}{\sigma_c^2} + l_1^2 - \frac{l_1^2}{T}$$

$$\sum_{j=1}^{T} \left(\sum_{i=1}^{I} \frac{X_{ij}}{\sigma_c^2} \right)^2 - \frac{l_1^2}{T} = w_1 - \frac{l_1^2}{T},$$

where $w_1 = \sum_{j=1}^{T} \left(\sum_{i=1}^{I} \frac{X_{ij}}{\sigma_c^2} \right)^2$. Similarly, let $w_2 = \sum_{j=1}^{T} \left(\sum_{i=1}^{I} \frac{W_{ij}}{\sigma_c^2} \right)^2$ and $w_3 = \sum_{j=1}^{T} \left(\sum_{i=1}^{I} \frac{(XW)_{ij}}{\sigma_c^2} \right)^2$.

Simplifying the second summand yields

$$\frac{y_1}{T} \boldsymbol{1'_{T-1}} \boldsymbol{B} \frac{y_1}{T} \boldsymbol{1_{T-1}} = \frac{y_1^2}{T^2} \boldsymbol{1'_{T-1}} (\boldsymbol{I_{T-1}} + \boldsymbol{1_{T-1}} \boldsymbol{1'_{T-1}}) \boldsymbol{1_{T-1}} =$$

$$\frac{y_1^2}{T^2}\big((T-1)+(T-1)\mathbf{1}_{\boldsymbol{T-1}}^{'}\mathbf{1}_{\boldsymbol{T-1}}\big) \;=\; \frac{y_1^2}{T}(T-1).$$

Simplifying the third summand yields

$$\frac{2y_1}{T}\mathbf{1}_{\boldsymbol{T-1}}^{'}\boldsymbol{B}\left(\sum_{i=1}^{I}\frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2}-\frac{l_1}{T}\mathbf{1}_{\boldsymbol{T-1}}\right)=$$

$$\frac{2y_1}{T}(\mathbf{1}_{\boldsymbol{T-1}}^{'}+(T-1)\mathbf{1}_{\boldsymbol{T-1}}^{'})\left(\sum_{i=1}^{I}\frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2}-\frac{l_1}{T}\mathbf{1}_{\boldsymbol{T-1}}\right)$$

$$=2y_1\mathbf{1}_{\boldsymbol{T-1}}^{'}\left(\sum_{i=1}^{I}\frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2}-\frac{l_1}{T}\mathbf{1}_{\boldsymbol{T-1}}\right)$$

$$=2y_1\left(-\sum_{i=1}^{I}\frac{X_{iT}}{\sigma_c^2}-\frac{l_1}{T}+\frac{y_1}{T}-\frac{y_1}{T}\right)=2y_1\left(\frac{y_1}{T}+\sigma_\alpha^2h_1-\sum_{i=1}^{I}\frac{X_{iT}}{\sigma_c^2}\right).$$

Substituting these expressions back into m_{11}, we obtain

$$=\frac{1}{(f+gT)}\left(\frac{y_1^2(f+g)}{f}-2y_1\eta_1+\left(\sum_{i=1}^{I}\frac{\boldsymbol{X}_{\boldsymbol{i,-T}}^{'}}{\sigma_c^2}-\sigma_\alpha^2h_1\mathbf{1}_{\boldsymbol{T-1}}^{'}\right)\boldsymbol{B}\left(\sum_{i=1}^{I}\frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2}-\sigma_\alpha^2h_1\mathbf{1}_{\boldsymbol{T-1}}\right)\right)$$

$$=\frac{1}{(f+gT)}(\frac{y_1^2(f+g)}{f}-2y_1(y_1+\sigma_\alpha^2h_1+ \tag{4.2}$$

$$\sum_{i=1}^{I}\frac{X_{iT}}{\sigma_c^2})+w_1-\frac{l_1^2}{T}+2y_1\left(\frac{y_1}{T}+\sigma_\alpha^2h_1-\sum_{i=1}^{I}\frac{X_{iT}}{\sigma_c^2})+\frac{y_1^2}{T}(T-1)\right)$$

$$=\frac{1}{(f+gT)}\left(y_1^2\left(\frac{f+gT}{fT}\right)+w_1-\frac{l_1^2}{T}\right) \;=\; \frac{y_1^2}{fT}+\frac{1}{(f+gT)}\left(w_1-\frac{l_1^2}{T}\right).$$

The second and third diagonal elements of the matrix $\boldsymbol{M}$, m_{22} and m_{33}, will have the same form with the corresponding values of $y_1, \boldsymbol{X_{i,-T}}, l_1,$ and h_1.

Simplifying the off-diagonal elements of $\boldsymbol{M}$

To simplify the off-diagonal values, we consider m_{21} and apply the results to the other off-diagonals. We have

$$m_{21}=\frac{1}{(f+gT)}(y_2\left(\frac{g+f}{f}y_1-y_1-\sigma_\alpha^2h_1+\sum_{i=1}^{I}\frac{X_{iT}}{\sigma_c^2}\right)+$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{W_{i,-T}}'}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}_{\boldsymbol{T-1}}'\right)\left(-y_1 \mathbf{1}_{\boldsymbol{T-1}} + \boldsymbol{B}\left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}}\right)\right)).$$

We simplify terms, moving left to right. The first term simplifies to

$$y_2\left(\frac{g+f}{f}y_1 - y_1 - \sigma_\alpha^2 h_1 + \sum_{i=1}^{I}\frac{X_{iT}}{\sigma_c^2}\right) \;=\; y_1 y_2 \frac{f+g}{f} - y_1 y_2 - \sigma_\alpha^2 h_1 y_2 + y_2 \sum_{i=1}^{I}\frac{X_{iT}}{\sigma_c^2}.$$

We now consider the second term and multiply $\left(\sum_{i=1}^{I} \frac{\boldsymbol{W_{i,-T}'}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}_{\boldsymbol{T-1}}'\right)$ through. The first product simplifies to

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{W_{i,-T}'}}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}_{\boldsymbol{T-1}}'\right)(-y_1 \mathbf{1}_{\boldsymbol{T-1}}) \;=\; -y_1 \sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2} + (T-1)\sigma_\alpha^2 h_2 y_1.$$

The final product can be written out as

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{W_{i,-T}}'}{\sigma_c^2} - \sigma_\alpha^2 h_2 \mathbf{1}_{\boldsymbol{T-1}}'\right)\left(\boldsymbol{B}\left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \sigma_\alpha^2 h_1 \mathbf{1}_{\boldsymbol{T-1}}\right)\right)$$

$$= \left(\sum_{i=1}^{I} \frac{\boldsymbol{W_{i,-T}}'}{\sigma_c^2} - \frac{l_2 - y_2}{T}\mathbf{1}_{\boldsymbol{T-1}}'\right)\left(\left(\boldsymbol{I_{T-1}} + \mathbf{1}_{\boldsymbol{T-1}}\mathbf{1}_{\boldsymbol{T-1}}'\right)\left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \frac{l_1 - y_1}{T}\mathbf{1}_{\boldsymbol{T-1}}\right)\right)$$

$$= \left(\sum_{i=1}^{I} \frac{\boldsymbol{W_{i,-T}}'}{\sigma_c^2} - \frac{l_2}{T}\mathbf{1}_{\boldsymbol{T-1}}' + \frac{y_2}{T}\mathbf{1}_{\boldsymbol{T-1}}'\right) *$$
$$\left(\left(\boldsymbol{I_{T-1}} + \mathbf{1}_{\boldsymbol{T-1}}\mathbf{1}_{\boldsymbol{T-1}}'\right)\left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \frac{l_1}{T}\mathbf{1}_{\boldsymbol{T-1}} + \frac{y_1}{T}\mathbf{1}_{\boldsymbol{T-1}}\right)\right)$$

$$= \left(\sum_{i=1}^{I} \frac{\boldsymbol{W_{i,-T}}'}{\sigma_c^2} - \frac{l_2}{T}\mathbf{1_{T-1}}'\right)\left(\left(\boldsymbol{I_{T-1}} + \mathbf{1_{T-1}}\mathbf{1}'_{\boldsymbol{T-1}}\right)\left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \frac{l_1}{T}\mathbf{1_{T-1}}\right)\right) +$$

$$\frac{y_2}{T}\mathbf{1}'_{\boldsymbol{T-1}}(\boldsymbol{I_{T-1}} + \mathbf{1_{T-1}}\mathbf{1}'_{\boldsymbol{T-1}})\frac{y_1}{T}\mathbf{1_{T-1}} +$$

$$\frac{y_2}{T}\mathbf{1}'_{\boldsymbol{T-1}}(\boldsymbol{I_{T-1}} + \mathbf{1_{T-1}}\mathbf{1}'_{\boldsymbol{T-1}})\left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \frac{l_1}{T}\mathbf{1_{T-1}}\right) +$$

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{W_{i,-T}}'}{\sigma_c^2} - \frac{l_1}{T}\mathbf{1}'_{\boldsymbol{T-1}}\right)(\boldsymbol{I_{T-1}} + \mathbf{1_{T-1}}\mathbf{1}'_{\boldsymbol{T-1}})(\frac{y_1}{T}\mathbf{1_{T-1}}).$$

We simplify each term in this product. The first term in this product can be simplified as

$$\left(\sum_{i=1}^{I} \frac{\boldsymbol{W_{i,-T}}'}{\sigma_c^2} - \frac{l_2}{T}\mathbf{1}'_{\boldsymbol{T-1}}\right)\left(\boldsymbol{B}\left(\sum_{i=1}^{I} \frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2} - \frac{l_1}{T}\mathbf{1_{T-1}}\right)\right)$$

$$= \sum_{j=1}^{T-1}\left(\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2}\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_c^2}\right) + \left(\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2}\right)\left(\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_c^2}\right) - \frac{l_1}{T}\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2} -$$

$$\frac{l_1(T-1)}{T}\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2} - \frac{l_2}{T}\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_c^2} - \frac{l_2(T-1)}{T}\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_c^2} + \frac{l_1 l_2(T-1)}{T^2} + \frac{l_1 l_2(T-1)^2}{T^2}$$

$$= \sum_{j=1}^{T-1}\left(\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2}\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_c^2}\right) + \left(\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2}\right)\left(\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_c^2}\right) -$$

$$l_1\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2} - l_2\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_c^2} + \frac{(T-1)l_1 l_2}{T}$$

$$= w_{XW} - \sum_{j=1}^{T}\left(\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2}\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_c^2}\right) - \frac{l_1 l_2}{T}, \text{ where } \quad w_{XW} = \sum_{j=1}^{T}\left(\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2}\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_c^2}\right).$$

Let $w_{X(XW)} = \sum_{j=1}^{T}\left(\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_c^2}\sum_{i=1}^{I}\frac{(XW)_{ij}}{\sigma_c^2}\right)$ and $w_{W(XW)} = \sum_{j=1}^{T}\left(\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2}\sum_{i=1}^{I}\frac{(XW)_{ij}}{\sigma_c^2}\right)$ for the other off-diagonal elements. Simplifying the remaining terms,

$$\frac{y_2}{T}\mathbf{1}_{\boldsymbol{T-1}}'\boldsymbol{B}\frac{y_1}{T}\mathbf{1}_{\boldsymbol{T-1}}+\frac{y_2}{T}\mathbf{1}_{\boldsymbol{T-1}}'\boldsymbol{B}\left(\sum_{i=1}^{I}\frac{\boldsymbol{X_{i,-T}}}{\sigma_c^2}-\frac{l_1}{T}\mathbf{1}_{\boldsymbol{T-1}}\right)+$$

$$\left(\sum_{i=1}^{I}\frac{\boldsymbol{W_{i,-T}}'}{\sigma_c^2}-\frac{l_2}{T}\mathbf{1}_{\boldsymbol{T-1}}'\right)\boldsymbol{B}(\frac{y_1}{T}\mathbf{1}_{\boldsymbol{T-1}})$$

$$=\frac{y_1y_2(T-1)}{T}+y_2\left(\frac{y_1}{T}+\sigma_\alpha^2h_1-\sum_{i=1}^{I}\frac{X_{iT}}{\sigma_c^2}\right)+y_1\left(\frac{y_2}{T}+\sigma_\alpha^2h_2-\sum_{i=1}^{I}\frac{Y_{iT}}{\sigma_c^2}\right).$$

After simplifying as much as possible, we yield the form of the off-diagonals:

$$=y_1y_2\frac{f+g}{f}-y_1y_2-\sigma_\alpha^2h_1y_2+y_2\sum_{i=1}^{I}\frac{X_{iT}}{\sigma_c^2}-y_1\sum_{j=1}^{T-1}\sum_{i=1}^{I}\frac{W_{ij}}{\sigma_c^2}+(T-1)\sigma_\alpha^2h_2y_1+$$

$$\frac{y_1y_2(T-1)}{T}+y_2\left(\frac{y_1}{T}+\sigma_\alpha^2h_1-\sum_{i=1}^{I}\frac{X_{iT}}{\sigma_c^2}\right)+y_1\left(\frac{y_2}{T}+\sigma_\alpha^2h_2-\sum_{i=1}^{I}\frac{Y_{iT}}{\sigma_c^2}\right)+w_{xy}-\frac{l_1l_2}{T}$$

$$=y_1y_2\frac{f+gT+fT}{fT}-y_1l_2+T\sigma_\alpha^2h_2y_1+w_{xy}-\frac{l_1l_2}{T}\;=\;y_1y_2\left(\frac{f+gT}{fT}\right)+w_{xy}-\frac{l_1l_2}{T}.$$

Multiplying the off-diagonal term by the constant $\frac{1}{f+gT}$ yields $\frac{y_1y_2}{fT}+\frac{1}{f+gT}\left(w_{xy}-\frac{l_1l_2}{T}\right)$.

Solving for Variances

We now have a simplified expression for $\boldsymbol{M}$. To obtain $(\boldsymbol{Z'V^{-1}Z})_{22}^{-1}$, we calculate $(\boldsymbol{Z'V^{-1}Z})_{22}-\boldsymbol{M}$ and take the inverse of this matrix. We have $(\boldsymbol{Z'V^{-1}Z})_{22}-\boldsymbol{M}$ equal to

$$\begin{bmatrix} l_1-z_1-\frac{y_1^2}{fT}-\frac{1}{f+gT}\left(w_1-\frac{l_1^2}{T}\right) & q_1-\frac{y_1y_2}{fT}-\frac{1}{f+gT}\left(w_{XW}-\frac{l_1l_2}{T}\right) & q_2-\frac{y_1y_3}{fT}-\frac{1}{f+gT}\left(w_{X(XW)}-\frac{l_1l_3}{T}\right) \\ q_1-\frac{y_1y_2}{fT}-\frac{1}{f+gT}\left(w_{XW}\frac{l_1l_2}{T}\right) & l_2-z_2-\frac{y_2^2}{fT}-\frac{1}{f+gT}\left(w_2-\frac{l_2^2}{T}\right) & q_3-\frac{y_2y_3}{fT}-\frac{1}{f+gT}\left(w_{W(XW)}-\frac{l_2l_3}{T}\right) \\ q_2-\frac{y_1y_3}{fT}-\frac{1}{f+gT}\left(w_{X(XW)}\frac{l_1l_3}{T}\right) & q_3-\frac{y_2y_3}{fT}-\frac{1}{f+gT}\left(w_{W(XW)}\frac{l_2l_3}{T}\right) & l_3-z_3-\frac{y_3^2}{fT}-\frac{1}{f+gT}\left(w_3-\frac{l_3^2}{T}\right) \end{bmatrix}$$

$$=\begin{bmatrix} b_{11}\,b_{12}\,b_{13} \\ b_{12}\,b_{22}\,b_{23} \\ b_{13}\,b_{23}\,b_{33} \end{bmatrix}.$$

For a stepped wedge design with only a single intervention (i.e. only θ_1 is included

in the model), the variance of the $\hat{\theta}_1$ would be the reciprocal of the first diagonal term $b_{11} = l_1 - z_1 - \frac{y_1^2}{fT} - \frac{1}{f+gT}\left(w_1 - \frac{l_1^2}{T}\right)$, which can be shown to be the same variance as found by Hussey and Hughes (Hussey and Hughes, 2007). If treatment effects are assumed to be additive and an interaction term is not included (i.e. θ_1 and θ_2 are in the model), we can take the inverse of the upper 2×2 matrix for the variance-covariance matrix of the treatment effect regression coefficients. For the full model with two treatment effects an interaction (i.e. model includes θ_1, θ_2, and θ_3), we take the inverse of the 3×3 matrix, yielding

$$Var(\hat{\theta}_1) = \frac{b_{22}b_{33} - b_{23}^2}{b_{11}(b_{22}b_{33} - b_{23}^2) - b_{21}(b_{12}b_{33} - b_{13}b_{32}) + b_{13}(b_{12}b_{23} - b_{13}b_{22})},$$

$$Var(\hat{\theta}_2) = \frac{b_{11}b_{33} - b_{13}^2}{b_{11}(b_{22}b_{33} - b_{23}^2) - b_{21}(b_{12}b_{33} - b_{13}b_{32}) + b_{13}(b_{12}b_{23} - b_{13}b_{22})},$$

$$Var(\hat{\theta}_3) = \frac{b_{11}b_{22} - b_{12}^2}{b_{11}(b_{22}b_{33} - b_{23}^2) - b_{21}(b_{12}b_{33} - b_{13}b_{32}) + b_{13}(b_{12}b_{23} - b_{13}b_{22})}.$$

Appendix B - Simulation Procedure

Piecewise Weibull Simulation Procedure

Let $S(t) = 1-F(t)$, where $F(t)$ is the cumulative distribution function (CDF) of a probability distribution at time t. It is known that the CDF of a continuous random variable follows a uniform distribution ranging from 0 to 1. Letting $U \sim \text{Unif}(0,1)$, then $F(t) \sim \text{Unif}(0,1)$ and therefore $S(t) = 1-F(t) \sim \text{Unif}(0,1)$. Combining the uniform distribution of U with $S(t) = \exp(-H(t))$, where $H(t)$ is the cumulative hazard function, solving for t in this relationship yields $H^{-1}(\log(1-U)) = t$. The survival time for each individual, t_k, can be solved from generating a uniformly distributed random variable and applying the inverse of the cumulative hazard function $H^{-1}(t)$ (Austin, 2012).

Under the piecewise Weibull formulation, the transition intensity between states c and d during interval $j = 1, \ldots, J$ at some time t with an interval-specific hazard ratio k_j can be expressed as:

$$\lambda_{cj}(t) = \begin{cases} k_1 \alpha \gamma^\alpha t_k^{\alpha-1} \text{ for } 0 = \tau_0 < t \leq \tau_1 \\ k_2 \alpha \gamma^\alpha t_k^{\alpha-1} \text{ for } \tau_1 < t \leq \tau_2 \\ \vdots \\ k_J \alpha \gamma^\alpha t_k^{\alpha-1} \text{ for } \tau_J < t \leq \tau_{J+1} \end{cases}$$

From here, we can construct the cumulative hazard for individual k who enters at time 0 and survives up to or is censored at time t_k in interval $j(k)$, the last interval for which interval for individual k is observed. We can write cumulative hazard for as

$$H_j(t_k) = \begin{cases} H_1 = k_1(\gamma t_k)^\alpha \quad \text{for} \quad 0 < t_k \leq \tau_1 \\ H_2 = H_1 + k_2(\gamma t_k)^\alpha - k_2(\gamma \tau_1)^\alpha \quad \text{for} \quad \tau_1 < t_k \leq \tau_2 \\ H_3 = H_2 + k_3(\gamma t_k)^\alpha - k_3(\gamma \tau_2)^\alpha \quad \text{for} \quad \tau_2 < t_k \leq \tau_3 \\ \vdots \\ H_j = H_{j-1} + k_j(\gamma t_k)^\alpha - k_j(\gamma \tau_{j-1})^\alpha \quad \text{for} \quad \tau_{j(k)-1} < t_k \leq \\ \tau_{j(k)}. \end{cases}$$

We solve for t_k by equating each line in this piecewise cumulative hazard to $-\log(1-U_k)$, where U_k is an instance of the random variable U for individual k. We can then write the inverse cumulative hazard function $H_k^{-1}(t_k)$ as

$$\begin{cases} t_k = \frac{(-\log(1-U_k))^{1/\alpha}}{k_1^{1/\alpha}\gamma} \quad \text{for} \quad 0 < -\log(1-U_k) \leq H_1 \\ t_k = \frac{(-\log(1-U_k) - H_1 + (k_2\gamma(\tau_1))^\alpha)^{1/\alpha}}{k_2^{1/\alpha}\gamma} \quad \text{for} \quad H_1 < -\log(1-U_k) \leq H_2 \\ \vdots \\ t_k = \frac{(-\log(1-U_k) - H_{j(k)-1} + (k_j\gamma(\tau_{j(k)-1}))^\alpha)^{1/\alpha}}{k_{j(k)-1}^{1/\alpha}\gamma} \text{ for } H_{j(k)-1} < -\log(1-U_k) \leq H_{j(k)} \end{cases}$$

To simulate cure proportions, we generate cure indicators for every individual $V_{ck} \sim Binomial(1, \pi_{ck})$. For any individual such that $V_{ck} = 1$, their time to transition out of state c is set to infinity.

Bibliography

Austin, P. C. (2012). Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29):3946–3958.

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

Barker, D., D'Este, C., Campbell, M. J., and McElduff, P. (2017). Minimum number of clusters and comparison of analysis methods for cross sectional stepped wedge cluster randomised trials with binary outcomes: A simulation study. *Trials*, 18.

Bartlett, M. (1951). An inverse matrix adjustment arising in discriminant analysis. *Annals of Mathematical Statistics*, 22(1):107–111.

Bastani, R. (2017). *Comparing Strategies for Health Clinics to Increase HPV Vaccinations in Youth*. Patient-Centered Outcomes Research Institute.

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

Beesley, L. and Taylor, J. (2019). EM algorithms for fitting multistate cure models. *Biostatistics*, 20:416–432.

Beretta, A. and Heuchenne, C. (2019). Variable selection in proportional hazards cure model with time-varying covariates, application to US bank failures. *Journal of Applied Statistics*, 46(9):1529–1549.

Bernhardt, P. W. (2016). A flexible cure rate model with dependent censoring and a known cure threshold. *Statistics in Medicine*, 35(25):4607–4623.

Betensky, R. and Mandel, M. (2015). Recognizing the problem of delayed entry in time-to-event studies: Better late than never for clinical neuroscientists. *Annals of Neurology*, 78:839–844.

Borg, K., Sutton, K., Beasley, M., Tull, F., Faulkner, N., Halliday, J., Knott, C., and Bragg, P. (2018). Communication-based interventions for increasing influenza vaccination rates among aboriginal children: A randomised controlled trial. *Vaccine*, 36:6790–6795.

Botta-Dukat, Z. (2009). Standardized or simple effect size: What should be reported. *British Journal of Psychology*, 100:603–617.

Botta-Dukat, Z. (2016). Cautionary note on calculating standardized effect size (SES) in randomization test. *Community Ecology*, 19:77–83.

Carney, P. A., Hatch, B., Stock, I., Dickinson, C., Davis, M., Larsen, R., Valenzuela, S., Marino, M., Darden, P. M., Gunn, R., et al. (2019). A stepped-wedge cluster randomized trial designed to improve completion of hpv vaccine series and reduce missed opportunities to vaccinate in rural primary care practices. *Implementation Science*, 14(1):1–9.

Cassarly, C., Martin, R., Chimowitz, M., Peña, E. A., Ramakrishnan, V., and Palesch, Y. Y. (2017). Assessing type I error and power of multistate Markov models for panel data—a simulation study. *Communications in Statistics-Simulation and Computation*, 46(9):7040–7061.

CDC (2019). Table 1. recommended child and adolescent immunization schedule for ages 18 years or younger, United States, 2019.

Clements, M. (2019). *Predictions for parametric and penalised multi-state Markov models: R Package*. Karolinska Institutet.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, chapter 2, pages 24–27. L. Erlbaum Associates.

Cook, R. J. and Lawless, J. F. (2018). *Multistate models for the analysis of life history data*. Chapman and Hall/CRC.

Cribari-Neto, F. and Zeileis, A. (2020). *Beta Regression in R: R Package*. Universidade Federal de Pernambuco.

Dilleya, S., Miller, K., and Huha, W. (2020). Human papillomavirus vaccination: Ongoing challenges and future directions. *Gynecologic Oncology*, 156:498–502.

Dirick, L., Bellotti, T., Claeskens, G., and Baesens, B. (2019). Macro-economic factors in credit risk calculations: Including time-varying covariates in mixture cure models. *Journal of Business & Economic Statistics*, 37(1):40–53.

Dong, Q., Peng, Y., and Li, P. (2021). Time to delisted status for listed firms in Chinese stock markets: An analysis using a mixture cure model with time-varying covariates. *Journal of the Operational Research Society*, pages 1–12.

Donner, A. and Klar, N. (2010). *Design and Analysis of Cluster Randomization Trials in Health Research*, chapter 1. John Wiley and Sons Ltd.

Ducrosa, F. and Pamphile, P. (2018). Bayesian estimation of Weibull mixture in heavily censored data setting. *Reliability Engineering and System Safety*, 180:453–462.

Dunnett, C. (1964). A multiple comparison procedure for comparing several treatments with a control. *Biometrics*, 20:482–491.

Durovni, B., Saraceni, V., Moulton, L., Pacheco, A., Cavalcante, S., King, B., Cohn, S., Efron, A., Chaisson, R., and Golub, J. (2013). Effect of improved tuberculosis screening and isoniazid preventive therapy on incidence of tuberculosis and death in patients with HIV in clinics in Rio de Janeiro, Brazil: a stepped wedge, cluster-randomised trial. *The Lancet Infectious Disease*, 13:852–858.

Durovni, B., Saraceni, V., van den Hof, S., Trajman, A., Cordeiro-Santos, M., Cavalcante, S., Menezes, A., and Cobelens, F. (2014). Impact of replacing smear microscopy with xpert mtb/rif for diagnosing tuberculosis in brazil: a stepped-wedge cluster-randomized trial. *PLoS medicine*, 11(12):e1001766.

Escobar-Bach, M., Maller, R., Van Keilgom, I., and Zhao, M. (2021). Estimation of the cure rate for distributions in the Gumbel maximum domain of attraction under insufficient follow-up. *Biometrika*, pages 1–15.

Feldman, H. and McKinlay, S. (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for designs. *Statistics in Medicine*, 11:1685–1704.

Feldman, H. A. and McKinlay, S. M. (1994). Cohort versus cross-sectional design in large field trials: Precision, sample size, and a unifying model. *Statistics in Medicine*, 13:61–78.

Felizzi, F., Paracha, N., Pöhlmann, J., and Ray, J. (2021). Mixture cure models in oncology: A tutorial and practical guidance. *PharmacoEconomics-Open*, pages 1–13.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31:799–815.

Fintzi, J., Bonnett, T., Sweeney, D. A., Huprikar, N. A., Ganesan, A., Frank, M. G., McLellan, S. L., Dodd, L. E., Tebas, P., and Mehta, A. K. (2021). Deconstructing the treatment effect of remdesivir in the adaptive COVID-19 treatment trial-1: Implications for critical care resource utilization. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*.

Ford, W. and Westgate, P. (2020). Maintaining the validity of inference in small-sample stepped wedge cluster randomized trials with binary outcomes when using generalized estimating equations. *Statistics in Medicine*, 39:2779–2792.

Fu, L., Zook, K., Gingold, J., Gillespie, C., Briccetti, C., Cora-Bramble, D., Joseph, J., Haimowitz, R., and Moon, R. (2016). Strategies for improving vaccine delivery: A cluster-randomized trial. *Pediatrics*, 137.

Girling, A. and Hemming, K. (2015). Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Statistics in Medicine*, 35:2149–2166.

Girling, A. J. (2018). Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling. *Statistics in medicine*, 37(30):4652–4664.

Gran, J., Lie, S., Oyeflate, I., Borgan, O., and Aalen, O. (2015). Causal inference in multi-state models–sickness absence and work for 1145 participants after work rehabilitation. *BMC Public Health*, 15.

Grantham, K., Kasza, J., Heritier, S., Hemming, K., and Forbes, A. (2018). Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. *Statistics in Medicine*, 38:1918–1934.

Grayling, M. and Wason, J. (2020). A web application for the design of multi-arm clinical trials. *BMC Cancer*, 20(80).

Grayling, M., Wason, J., and Mander, A. (2017). Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. *Trials*, 18:33.

Grayling, M. J., Mander, A. P., and Wason, J. M. (2019). Admissible multiarm stepped-wedge cluster randomized trial designs. *Statistics in medicine*, 38(7):1103–1119.

Green, S., Liu, P., and O'Sullivan, J. (2002). Factorial design considerations. *Journal of Clinical Oncology*, 20:3424–3230.

Hanin, L. and Huang, L. (2014). Identifiability of cure models revisited. *Journal of Multivariate Analysis*, 130:261–274.

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

Hemming, K., Lilford, R., and Girling, A. J. (2014). Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple level designs. *Statistics in Medicine*, 34:181–196.

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

Hemming, K., Taljaard, M., and Forbes, A. (2017). Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. *Trials*, 18:101.

Hooper, R. and Bourke, L. (2015). Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *bmj*, 350.

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

Hsieh, H.-J., Chen, T., and Chang, S.-H. (2002). Assessing chronic disease progression using non-homogeneous exponential regression Markov models: an illustration using a selective breast cancer screening in taiwan. *Statistics in Medicine*, 21:3369 – 3382.

Hughes, J., Granston, T., and Heagerty, P. (2015). Current issues in the design and analysis of stepped wedge trials. *Contemporary Clinical Trials*, 45(Part A):55–60.

Hurley, L., Beaty, B., Lockhart, S., Gurfinkel, D., Dickinson, L., Roth, H., and Kempe, A. (2019). Randomized controlled trial of centralized vaccine reminder/recall to improve adult vaccination rates in an accountable care organization setting. *Preventive Medicine Reports*, 15.

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

Jiang, W., Sun, H., and Peng, Y. (2017). Prediction accuracy for the cure probabilities in mixture cure models. *Statistical methods in medical research*, 26(5):2029–2041.

Jit, M. (2021). Informing global cost-effectiveness thresholds using country investment decisions: human papillomavirus vaccine introductions in 2006-2018. *Value in Health*, 24(1):61–66.

Kahan, B. (2013). Bias in randomised factorial trials. *Statistics in Medicine*, 32:4540–4549.

Kalbfleisch, J. and Lawless, J. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80(392):863–871.

Kasza, J., Taljaard, M., and Forbes, A. (2019). Information content of stepped-wedge designs when treatment effect heterogeneity and/or implementation periods are present. *Statistics in Medicine*.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 42(4):855–865.

Kearns, B., Stevenson, M., Triantafyllopoulos, K., and Manca, A. (2021). The extrapolation performance of survival models for data with a cure fraction: A simulation study. *Value in Health*, 24:1634–1642.

Kuczmarski, R., Ogden, C., Guo, S., Grummer-Strawn, L., Flegal, K., Mei, Z., Wei, R., Curtin, L., Roche, A., and Johnson, C. (2002). 2000 CDC growth charts for the united states: methods and development. *Vital Health Stat*, 11:1–190.

Le-Rademacher, J. G., Peterson, R. A., Therneau, T. M., Sanford, B. L., Stone, R. M., and Mandrekar, S. J. (2018). Application of multi-state models in cancer clinical trials. *Clinical Trials*, 15(5):489–498.

Li, F. (2019). Design and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure. *Statistics in Medicine*, 39:438–455.

Li, F., Hughes, J., Hemming, K., Taljaard, M., Melnick, E., and Heagerty, P. (2020). Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. *Statistical Methods in Medical Research*, 30:1–28.

Li, F., Turner, E., and Preisser, J. (2018). Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics*.

Lin, L.-H. and Huang, L.-S. (2019). Connections between cure rates and survival probabilities in proportional hazards models. *Stat*, 8(1):e255.

Lindner, S. and McConnell, K. J. (2021). Heterogeneous treatment effects and bias in the analysis of the stepped wedge design. *Health Services and Outcomes Research Methodology*, 21(4):419–438.

Lu, T.-T. and Shiou, S.-H. (2000). Inverses of 2x2 block matrices. *Computers and Mathematics with Applications*.

Lyons, V., Li, L., Hughes, J., and Rowhani-Rahbar, A. (2017). Proposed variations of the stepped-wedge design can be used to accommodate multiple interventions. *Journal of Clinical Epidemiology*, 86:160–167.

Ma, Y., Jenkins, H., Sebastiani, P., Ellner, J., Jones-López, E., Dietze, R., Horsburgh Jr, C., and White, L. (2020). Using cure models to estimate the serial interval of tuberculosis with limited follow-up. *Practice of Epidemiology*, 189.

Magnusson, A., Skaug, H., Nielsen, A., Berg, C., Kristensen, K., Maechler, M., van Bentham, K., and Bolker, B. (2020). *glmmTMB: Generalized Linear Mixed Models using Template Model Builder*.

Martin, J. T., Hemming, K., and Girling, A. (2019). The impact of varying cluster size in cross-sectional stepped-wedge cluster randomised trials. *BMC Medical Research Methodology volume*, 19.

Oelhert, G. (2010). *A First Course in Design and Analysis of Experiments*, chapter 8, pages 170–171. W. H. Freeman.

Othus, M., Bansal, A., Erba, H., and Ramsey, S. (2020). Bias in mean survival from fitting cure models with limited follow-up. *Value in Health*, 8:1034–1039.

O'Keeffe, A., Tom, B., and Farewell, V. (2017). Mixture distributions in multi-state modelling: Some considerations in a study of psoriatic arthritis. *Statistics in Medicine*, 32:600–619.

Perkins, R., Legler, A., Jansen, E., Bernstein, J., Pierre-Joseph, N., Eun, T., Biancarelli, D., Schuch, T., Leschly, K., Fenton, A., Adams, W., Clark, J., Drainoni, M.-L., and Hanchate, A. (2020). Improving HPV vaccination rates: A stepped-wedge randomized trial. *Pediatrics*, 146.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Willigen, B. V., and Ranke, J. (2020). *Linear and Nonlinear Mixed Effects Models*.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reuther, S., Holle, D., Buscher, I., Dortmann, O., Müller, R., Bartholomeyczik, S., and Halek, M. (2014). Effect evaluation of two types of dementia-specific case conferences in German nursing homes (FallDem) using a stepped-wedge design: study protocol for a randomized controlled trial. *Trials*, 15.

Rutten, L., Breitkopf, C., Jennifer, Sauver, Croghan, I., Jacobson, D., Wilson, P., Herrin, J., and Jacobson, R. (2018). Evaluating the impact of multilevel evidence-based implementation strategies to enhance provider recommendation on human papillomavirus vaccination rates among an empaneled primary care patient population: a study protocol for a stepped wedge cluster randomized trial. *Implementation Science*, 13.

Samaniego, F. J. (2010). *A comparison of the Bayesian and frequentist approaches to estimation*, volume 24. Springer.

Sherman, J. and Morrison, W. J. (1949). Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. *Annals of Mathematical Statistics*.

Soulakova, J. (2011). Resampling-based and other multiple testing strategies with application to combination drug trials with factorial designs. *Statistical Methods in Medical Research*, 20:505–521.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Stan Development Team (2022). Stan modeling language users guide and reference manual. Version 2.29.

Stedman, M. R., Feuer, E. J., and Mariotto, A. B. (2014). Current estimates of the cure fraction: a feasibility study of statistical cure for breast and colorectal cancer. *Journal of the National Cancer Institute Monographs*, 2014(49):244–254.

Sundin, P. and Crespi, C. M. (2022). Power analysis for stepped wedge trials with multiple interventions. *Statistics in Medicine*, 41:1498–1512.

Taljaard, M., Teerenstra, S., Ivers, N., and Fergusson, D. (2016a). Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clinical Trials*, 13:459–463.

Taljaard, M., Teerenstra, S., Ivers, N., and Fergusson, D. (2016b). Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. *Clinical Trials*, 13:459–463.

Teerenstra, S., Eldridge, S., Graff, M., de Hoop, E., and Borma, G. F. (2012). A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in Medicine*, 31:2169–2178.

Teerenstra, S., Taljaard, M., Haenen, A., Huis, A., Atsma, F., Rodwell, L., and Hulscher, M. (2019). Sample size calculation for stepped-wedge cluster-randomized trials with more than two levels of clustering. *Clinical Trials*, 16:225–236.

Titman, A. (2011). Flexible nonhomogeneous Markov models for panel observed data. *Biometrics*, 67:780–787.

Titman, A. (2015). Transition probability estimates for non-Markov multi-state models. *Biometrics*, 71:1034–1041.

van der Geest, B., de Graaf, J., Bertens, L., Poley, M., Ista, E., Kornelisse, R., Reiss, I., Steegers, E., and Been, J. (2019). Screening and treatment to reduce severe hyper-

bilirubinaemia in infants in primary care (STARSHIP): a factorial stepped-wedge cluster randomised controlled trial protocol. *BMJ Open*, 9.

Verbeke, G. and Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data*, chapter 2, pages 56–57. Springer-Verlag.

Voldal, E. C., Xia, F., Kenny, A., Heagerty, P. J., and Hughes, J. P. (2022). Random effect misspecification in stepped wedge designs. *Clinical Trials*, page 17407745221084702.

von Cube, M., Schumacher, M., and Wolkewitz, M. (2017). Basic parametric analysis for a multi-state model in hospital epidemiology. *BMC Medical Research Methodology*, 17.

Walker, T. Y., Elam-Evans, L. D., Yankey, D., Markowitz, L. E., Williams, C. L., Fredua, B., Singleton, J. A., and and, S. S. (2019). National, regional, state, and selected local area vaccination coverage among adolescents aged 13–17 years — United States, 2018. *Morbidity and Mortality Weekly Report*, 68.

Wang, S., Zhang, J., and Lu, W. (2012). Sample size calculation for the proportional hazards cure model. *Statistics in medicine*, 31(29):3959–3971.

Wang, Y., Tang, Y., and Zhang, J. (2020). Bayesian approach for proportional hazards mixture cure model allowing non-curable competing risk. *Journal of Statistical Computation and Simulation*, 90(4):638–656.

Whittingham, K., Sanders, M., McKinlay, L., and Boyd, R. N. (2014). Interventions to reduce behavioral problems in children with cerebral palsy: An RCT. *Pediatrics*, 133:1249–1257.

Woertman, W., de Hoop, E., Moerbeek, M., Zuidemac, S., Gerritsen, D., and Teerenstra, S. (2013). Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology*, 66:752–758.

Wu, G., Chang, S.-H., and Chen, T. (2008). A Bayesian random-effects Markov model for

tumor progression in women with a family history of breast cancer. *Biometrics*, 64:1231–1237.

Yiu, S., Farewell, V., and Tom, B. (2017). Clustered multistate models with observation level random effects, mover–stayer effects and dynamic covariates: modelling transition intensities and sojourn times in a study of psoriatic arthritis. *Applied Statistics*, 67:481–500.

Yu, B. and Tiwari, R. C. (2012). A Bayesian approach to mixture cure models with spatial frailties for population-based cancer relative survival data. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 40(1):40–54.

Zhang, P., Shoben, A., Jackson, R., and Fernandez, S. (2020). Variance formulae for multi-phase stepped wedge cluster randomized trial. *Statistics in Medicine*.