Addressing Issues with the Stepped Wedge Design of Cluster-Randomized Clinical Trials

Tanya S. Granston

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

James Hughes, Chair

Peter Gilbert

Barbra Richardson

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

Addressing Issues with the Stepped Wedge Design of Cluster-Randomized Clinical Trials

Tanya S. Granston

Chair of the Supervisory Committee:
Professor James Hughes
Biostatistics

The stepped wedge design (SWD) of cluster-randomized trials has been growing in popularity and is increasingly being used to efficiently evaluate the rollout of interventions in a community setting. The design is especially useful in resource limited settings where it may not be feasible to introduce interventions all at once and where there are substantial clusters/communities/groups to provide answers regarding intervention effects. This paper reviews recent work on, and applications of, the SWD. Issues unique to the design are raised and two such issues, lagged effects and interval censoring in the context of the SWD, are addressed. The effects of frailty on hazard ratio estimation are also briefly analyzed.

If an intervention is not fully implemented or not fully effective in all study units in the time step in which it is assigned, i.e., before the next time step, there are implications

for effect estimation and power with the SWD and these implications have not been well studied. Additionally, methods for estimating and modeling lags in effect in this context are needed. A two-step method is proposed to estimate lagged effects and the estimator is evaluated in simulations. Suggestions are made for calculating power when a lagged effect is anticipated or when there are delays in intervention rollout. The method is applied to the stepped wedge trial of expedited partner treatment in Washington State.

There has been little work on issues and methods of analysis in stepped wedge trials with time-to-event endpoints. In this paper, the finite sample behavior of the hazard function and hazard ratio estimate in the presence of frailty are studied and analytic methods to address interval censoring of time-to-event endpoints in the context of the SWD are developed. An EM Algorithm approach with random cluster effects and true failure times as joint latent variables is proposed to address interval censoring in discrete time. A method for computing the power of stepped wedge trials with time-to-event endpoints is proposed and the effect of interval censoring on power is investigated.

Finally, the implications of this work on applied research and plans for future work are discussed.

Contents

List of Tables

List of Figures

ACKNOWLEDGMENTS

I would like to thank the members of my committee, Jim Hughes, Peter Gilbert, Barbra Richardson, James Dai, and Grace John-Stewart for their time and patience. Their advice and suggestions have helped to make this dissertation something of which I can be proud. I'd additionally like to thank Jim Hughes for the many lessons and his humor.

I might not have made it through the program without the support of my cohort, particularly during the first two years, and I'm eternally grateful to them for that.

Susanne May has been very supportive, offering really helpful advice at some difficult times, and I'm grateful to her. My friends Amy Laird, Dara Farrell, and Sheena Jacob have been invaluable throughout this process and I thank them for the support and the distractions.

Chris Feilbach was my C tutor towards the end, helping to speed up my simulations. Thank you very much, Chris.

Last, but certainly not least, I thank my mother, Sonia Samuels, my brother, Leonard Granston, Jr., and the rest of my family and friends, especially those in Jamaica, for believing in me and supporting me from thousands of miles away.

I acknowledge the financial support provided by NIH grant AI29168 for the research described in this dissertation.

1 — THE STEPPED WEDGE DESIGN

1.1 Introduction

In a stepped wedge design (SWD) trial the intervention or treatment is introduced to each unit (person, clinic, community) at a random time. Each unit spends time in the control state but unidirectionally crosses over into the intervention state creating a stepped wedge schematic as seen in Figure 1.1. This allows us to construct both within- and between-unit estimates of intervention effects. The design is particularly useful in situations where it is unethical to withhold treatment but where time and/or finances preclude the provision of the treatment to all units at once (e.g., to evaluate the effectiveness during rollout of a treatment which has been shown to be efficacious in an individually randomized trial or in a different setting). The design has been described and critiqued previously [1]-[6], applied in several trials [7]-[13], and is under consideration for future trials, including phase IV HIV vaccine trials [14].

We review recent research on, and applications of, the SWD and outline several directions for future research. In Section 1.2 we review recent results and discussions relevant to stepped wedge trials, including applications of the design. In Section 1.3 we examine key aspects concerning the design and analysis of SWD trials. We conclude with a brief discussion about future directions in research on the SWD.

Randomization Unit	Time 1	2	3	4	5
1	0	1	1	1	1
2	0	1	1	1	1
3	0	0	1	1	1
4	0	0	1	1	1
5	0	0	0	1	1
6	0	0	0	1	1
7	0	0	0	0	1
8	0	0	0	0	1

Figure 1.1 Stepped Wedge Trial Design indicating treatment (1) assignment (two units per time step) and control (0) assignment.

1.2. Stepped Wedge Trials - Overview and Recent Results

Stepped wedge trials may be conducted using individuals or clusters as the unit of randomization; the outcome may be measured cross-sectionally (e.g., means or prevalences) or longitudinally (e.g., incidences). At the individual level, the intervention may be provided at a point in time (e.g., vaccination) or may consist of an ongoing program (e.g., daily medication). Similarly, from the community perspective, the intervention may be viewed as a one-time event or an ongoing program. These variations can be illustrated by describing several recent applications of this design.

The Gambia Hepatitis Study [7] evaluated the efficacy of a 4-dose hepatitis B vaccine (HBV) regimen (administered to infants over a 9-month period beginning just after birth) for reducing the incidence of hepatocellular cancer (HCC) and other chronic liver diseases manifested and observed in adulthood. Participants consisted of newborns born in and after July, 1986 (through the end of the study) in The Gambia. During the four-year program

rollout all infants were administered the (pre-existing) standard course of vaccinations but, every three months, HBV vaccination was permanently added to the routine vaccinations in one randomly selected geographical area (health district). Therefore, from the perspective of the SWD, the health districts formed the clusters and the babies born in each three-month period in each cluster formed a cohort. Cancer registries were set up to survey cases of HCC, liver cirrhosis, and other chronic liver diseases over the next 30-40 years. These rates can be compared between those subjects born in a geographic area/time period when the HBV vaccine was available and those born when the vaccine was not available to estimate the efficacy of the vaccine.

Nielsen *et al.* [8] used a non-randomized SWD study to evaluate how Vitamin A supplementation impacts mortality of children in emergency situations. Participants were children aged six months to five years living in the capital of Guinea-Bissau from October, 1998 through the end of the war in 1999. Approximately every three months when field assistants of the Bandim Heath Project conducted visits to monitor children, the children were offered Vitamin A. Delays in delivery of Vitamin A to children created the SWD in this study. A Cox proportional hazards model was used to compare hazards of death associated with Vitamin A supplementation adjusted for the age of the child and month of delivery of the Vitamin A.

Grant *et al.* [9] describes an individually randomized stepped wedge trial of tuberculosis (TB) preventive therapy in HIV-positive men in South Africa. The study took place in one health clinic, which provided care for employees of a gold mining company. Since it was

logistically impossible to start all potential participants on the prevention program simultaneously, employees previously testing positive for HIV were given a randomly determined time to be interviewed for participation. Those agreeing to be part of the study and showing no active TB were given isoniazid and followed for incident TB episodes. The incidence of TB was then compared between the two phases (before and after first clinic visit) using a Poisson random effects model. Adjustment was made for calendar time to account for the increasing incidence of TB expected with progression of HIV disease.

The THRio study [10] sought to determine whether training and implementation of a program to screen and treat latent TB with isoniazid in HIV-infected individuals in Rio de Janeiro would reduce the incidence of latent TB in this population. There are 29 HIV clinics under the purview of the health department in Rio de Janeiro and the training and implementation program was introduced in two randomly chosen clinics every two months until it was standard of care in all 29 clinics. A study cohort of HIV-infected persons over the age of 15 was assembled at the start of the study period in each clinic and these participants were followed and tested for TB during the study implementation period.

The SWD was also used in study of the effect of anti-retroviral therapy (ART) integration into antenatal care clinics on the percent of ART treatment eligible pregnant women who initiate ART within 60 days of HIV diagnosis [11]. The units were eight clinics in the Lusaka district of Zambia, one of which integrated ART provision into antenatal care every month. Logistic regression accounting for clinic and time was used to estimate the adjusted odds ratio for ART enrollment and initiation.

Mhurchu *et al.* [12] describes a cluster randomized controlled SWD trial investigating the effects of a free school breakfast program (SBP) on the attendance, academic achievement, and short-term hunger of children, aged 5 to 13 years, in 14 low socioeconomic area New Zealand schools. Three-four schools were randomly assigned to begin the free SBP every term (of four terms). The primary analysis made use of a logistic mixed effect model of the odds ratio for school attendance adjusted for age, gender, ethnicity, and school term for time trends.

In another example we describe a trial of expedited partner therapy (EPT) giving antibiotics (or vouchers for antibiotics) to persons diagnosed with gonorrhea or chlamydia to give to their partners in Washington State. In this ongoing trial EPT is being evaluated using a SWD [13, 15] with counties as the unit of randomization. Prevalences of gonorrhea and chlamydial infection before and after implementation are being evaluated through sentinel surveillance at selected health care sites.

Each of these examples raises unique issues with respect to the design and analysis of a stepped wedge trial. We review some of these issues in Section 1.3.

1.3. Stepped Wedge Trials - Issues with Design and Analysis

1.3.1 Intervention Effect Estimation

Analysis of data from a stepped wedge trial essentially involves comparing outcomes at times when the units are under treatment to outcomes at times when the units are not under treatment. The outcome may be measured using repeated cross-sectional samples (e.g., the

Washington state EPT trial) or by following cohorts longitudinally (e.g., the THRio study). Information on the treatment effect arises from two sources: i) comparisons between units receiving and not receiving the intervention at a given time, and ii) comparisons before and after treatment is initiated within a unit. A naïve estimate that uses only the before-after (within-unit) information will be biased if there is an underlying time trend in the response (i.e., the intervention effect will be confounded with the time trend). An estimate that utilizes only between-unit information (e.g., by combining repeated cross-sectional estimates of the treatment effect) is inefficient as we demonstrate in Chapter 2. Thus, an estimate of the treatment effect that is both unbiased and efficient must combine both sources of information. Here we describe some approaches that have been used for estimating intervention effects in stepped wedge trials.

1.3.1.1 Mean/Prevalence Outcomes

In the context of repeated cross-sectional surveys, Hussey and Hughes [1] use a mixed effects model [16] that incorporates both within- and between-unit information on the treatment effect, accounts for within-unit correlation, and explicitly models a time effect. Although Hussey and Hughes focus on cluster randomization and a cluster-level analysis of prevalence with an identity link (so that the model parameters estimate risk differences), they note that individual level data could be analyzed similarly using generalized linear mixed models (GLMM) [17]. An individual level analysis is preferable when cluster sample sizes vary and, for binary outcomes, can be used to provide estimates of relative risks or odds ratios, if desired. Scott [14] developed a generalized estimating equations (GEE) [18]

approach to estimating intervention effects using repeated cross-sectional incidences within clusters, arguing that a population-averaged effect estimate is more relevant to public health and public health policy. She developed both model-based and model-free permutation methods to test the statistical significance of the resulting estimate as a way to adjust for the well-known bias in GEE estimates when the number of clusters is small [19]. These methods provide a robust approach to cluster-randomized trial analysis in general and the SWD in particular and can be used with individual-level or cluster-level data.

1.3.1.2 Incidence Outcomes

If the outcome of interest is disease incidence, then, typically, the endpoint will be measured in cohorts. These cohorts may be identified once at the beginning of the study, or new cohorts may be identified at each time step. In the latter case, the data may be analyzed using methods similar to those described above for cross-sectional data. For example, in the Gambia Hepatitis study [7], cohorts were effectively formed by the children born in each 3-month period (time step) during the study. The outcome (incidence of hepatocellular cancer and chronic liver diseases) can then be compared between cohorts who were given HBV and those who were not. Adjustment to allow for variable person-time within each unit and within each time step is possible using an offset term in an individual level or a cluster level analysis.

When cohorts are formed at the start of the trial and the outcome can only occur once (e.g., death, HIV infection), then high risk individuals are more likely to experience the event early in the trial (when most clusters are in the control condition) and mostly low risk indi-

viduals will remain late in the trial (when most clusters are in the intervention condition). Even if the intervention is ineffective, a naïve analysis that simply compares incidence during the control and intervention periods would suggest that the treatment is effective in this case. This bias has been termed healthy survivor bias. The solution is straightforward - the analysis must control for time. For example, in the THRio study, Moulton *et al.* [10] follows clinic-specific cohorts over time and propose controlling for time by comparing outcomes at each time step to the risk set at that same time using a partial likelihood approach. They propose pooling the estimated intervention effect over all time steps (under a proportional hazards assumption) while accounting for within-cluster correlation using bootstrapping or robust variance estimates. Thus, any underlying time trend will be absorbed into the baseline hazard. However, one subtlety remains. If risk varies among individuals and the effect of the intervention depends on the risk level of the individual (e.g., the intervention is more effective in high risk individuals) then the changing composition of the risk set over time in a cohort will lead to a changing effect size over time (i.e., a treatment by time-on-study interaction or, more generally, a time-dependent effect). Whether one considers this a bias is debatable but it clearly complicates interpretation and the estimated intervention effect (if averaged over time) will depend strongly on the design and duration of the study. Importantly, however, such time-dependent effects are identifiable and estimable - see Section 1.3.2.

1.3.2 Lags and Time-Dependent Effects

Time-dependent effects may arise in a number of ways in a stepped wedge trial. In the context of our discussion of healthy survivor bias above, we described how a treatment by time-on-study interaction can arise if a cohort of participants is monitored for the duration of the trial and the treatment effect varies by level of risk of the participants. Any influence of time-on-study on the effect of the intervention can be estimated by the inclusion of an intervention by time-on-study interaction term in the analysis; this is an important area for further research.

We also consider treatment by time-on-intervention interactions. For example, if the intervention is an ongoing program then one might experience increasing effectiveness of the program (due to increased provider experience) or decreasing effectiveness of the program (due to participant fatigue) over time. Even if the treatment is a one-time intervention (e.g., vaccination) the possibility of a treatment by time-on-intervention interaction exists (e.g., waning effectiveness of vaccination) although this effect may occur beyond the time frame of the study.

Delayed intervention effects represent one type of treatment by time-on-intervention interaction. Unrecognized, delayed intervention effects can result in an effective intervention appearing to have no effect and have serious implications for the power of a study [1]. If anticipated, however, it may be possible to incorporate the delayed intervention effect into the study design. The Gambia Hepatitis study [7] provides an example of such a circumstance. As noted, chronic liver disease and hepatocellular cancer were key outcomes in that study

but, because of the nature of these diseases, any effect of vaccination in newborns on the outcome would not be seen until years after vaccination (well into adulthood). A simplistic measure of incidence of liver disease in the population immediately after introduction of the vaccination program would show no effect of the intervention. The study dealt with this anticipated delayed intervention effect by design. In the national cancer registry the birthdate and birthplace of each person with liver disease is recorded and, therefore, the investigators will be able to properly attribute each outcome measurement to the appropriate time step and intervention status (i.e., was the vaccination program in place in the health district when the individual was born). This is equivalent to identifying a new cohort at each time step, exposing them to the intervention or control condition as appropriate, and following them for the outcome.

It may not always be feasible to handle issues of this nature through design, especially if the exact nature of the lag is not known or if the lag is not anticipated. If a delayed intervention effect is not anticipated in the design, then it can be incorporated into the analysis. Hussey and Hughes [1] and Moulton *et al.* [10] have made suggestions for dealing with lags through analysis. If the nature of the lag effect is known (e.g., the intervention is not fully implemented as originally scheduled, and the degree of implementation is known or measureable), Hussey and Hughes [1] suggest using fractional values for the intervention indicator to represent the expected percentage of effectiveness of the intervention after certain time steps and thereby provide an unbiased estimate of the long term intervention effect. They found, however, that power is reduced in the presence of a lag effect and the reduction is greater

for longer lags. Similarly, Moulton *et al.* [10] suggested performing an as-treated (instead of intent-to-treat (ITT)) analysis by focusing on individuals and tracking them from study entry through follow up for active TB, while considering their actual intervention status (rather than their randomized intervention status) as a time-dependent covariate in order to handle delayed intervention effects.

If the intervention is initiated at each time step as planned but takes more time than expected to implement, a lag in the effect of the intervention (relative to the time step) is created, apart from any lag in the mechanism of action of the intervention. As above, ignoring this slow introduction and carrying out an ITT analysis has the potential to attenuate the estimated (long-term) treatment effect and reduce power. One way to deal with this is to adjust the ITT analysis for the percentage of the cluster that has completed introduction, if that information is available. While this approach can account for incomplete introduction, the adjustment may not adequately revert the attenuation, especially since the relationship between the percent completeness of the implementation and the effect size may not be linear. The ITT analysis can also be adjusted for the time it takes to completely introduce the intervention to each individual. In the context of a cluster-randomized trial, where individuals are likely to be more similar in rollout completion time and, perhaps to a lesser degree, risk within each cluster than across clusters, it is perhaps more practical to adjust for the mean (or some other summary measure) introduction completion time per cluster. Again, this adjustment may not be adequate. The slow introduction can also be handled by doing an as-treated analysis (using time of complete introduction rather than

time randomized, similar to what was suggested by Moulton *et al.* [10]). However, as with any as-treated analysis, there is a potential for bias in the effect estimate if there is differential implementation of the intervention by risk. For example, in a trial measuring the effect of circumcision on incidence of a sexually transmitted disease, more promiscuous and susceptible men may be more reluctant to show up for scheduled appointments to receive the intervention. Any favorable solution to this dilemma would involve recovering power while limiting the potential for bias and, again, is an important area for future research.

If prior information about the nature of the lag is not known (e.g., the effect of the intervention lags behind the introduction of the intervention), then it is theoretically possible to estimate the lag effect from the data (i.e., a treatment by time-on-treatment interaction) and this is a research topic we consider in Chapter 3.

1.3.3 Outcome Measurement Overlaps Two or More Steps

In the SWD trials described above, the intervention is introduced in a time step and the outcome is measured at the same time or before the start of the next time step. In the case of incident disease trials, however, the outcome may be measured in intervals that overlap two or more time steps. This is an example of interval censoring. If the intervention is introduced at discrete time steps and the outcome cannot be unambiguously assigned to an individual time step then we have discrete time interval censoring. More generally, one might envision an intervention being rolled out (introduced) into communities in an opportunistic fashion that does not result in evenly spaced time steps. Simultaneously, rolling or open cohorts (where participants in a trial enter and leave the study cohort at any time during the trial)

may be formed to track outcome events. If the outcome is assessed by periodic testing of the members of the cohort, then this design will likely produce outcome measurement intervals that overlap the introduction of the intervention. For example, imagine a trial where the intervention is rolled out more or less bimonthly but at some point is delayed by a couple of weeks. A cohort is formed at the beginning of the trial and the outcome (incident disease) is measured in each participant bimonthly (beginning one week after the start of the study (see Figure 1.2 for the timeline). This creates a scenario where the 3rd and 4th outcome measures cannot be definitively attributed to a time step and, for some units, the outcome measurement cannot be definitively attributed to control or intervention condition. It is

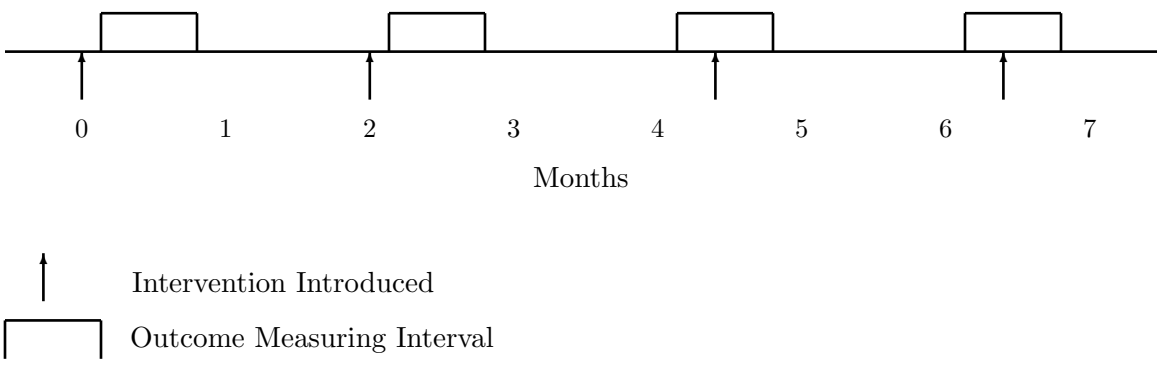

Figure 1.2 Outcome Measurements Overlapping Time Steps.

important in this design that the intervention rollout be independent of the outcome measurement process (i.e., introduction of the intervention does not increase or decrease the probability of censoring or observing an outcome).

The common problem in these situations is that it is not known in which time steps all outcome measures occurred. Thus, the intervention status at the time that the outcome occurs is not known for some individuals/clusters. When this happens, it becomes more difficult to assess intervention effects and properly account for time trends. In addition, powering a trial with this type of design is not straightforward. Current methods of analyzing SWD trial data are inadequate to handle these situations. Methods for analyzing interval-censored data are considered in Chapter 4.

The SWD of cluster-randomized trials is growing in popularity. Though the intervention in the SWD eventually becomes available to all participants in the study, the design is still relevant due to the need to efficiently evaluate the rollout of interventions that are individually efficacious in a community setting, as well as the need to maintain the golden standard of a randomized study. However, there remain limitations and outstanding issues with this design, which could potentially limit its usefulness where it's most needed. We address lagged effects and interval censoring in the context of the SWD in this paper. In the future, there are interesting extensions to the design that need to be explored. Particularly, we will consider how a SWD can be used to evaluate combination interventions. Sometimes, the effect of multiple interventions and their combinations are of interest in public health studies. For example, malaria control programs include both drug treatment for affected individuals and bed nets and mosquito control for prevention. Trachoma prevention interventions use the SAFE intervention strategy (Surgery, Antibiotics (azithromycin), controlling Flies, and Environmental sanitation). Increasingly, HIV prevention programs are

considering multi-component HIV prevention interventions [20, 21, 22] and the hope is that these multi-component interventions will have a synergistic effect on HIV acquisition, i.e., the combination of interventions will have an effect that exceeds the additive effect of the individual interventions on acquisition of HIV.

2 — EFFECT OF IGNORING WITHIN-CLUSTER INFORMATION IN THE SWD

One of the advantages of the SWD over traditional designs is the ability to get an estimate of the effect of the intervention by comparing outcomes within clusters across time in addition to comparing outcomes between clusters at a particular time. This is not possible in a traditional parallel design, for example. Both sources of information are important in the SWD. A comparison of outcomes only within clusters at time steps when clusters are in contrasting intervention conditions ignores potential time trends and will provide biased estimates of the effect of the intervention. On the other hand, comparison of outcomes only between clusters at each time step, though adjusting for potential time trends, can be inefficient particularly when outcomes vary between clusters. We investigate the effect of ignoring within-cluster (before-after) information on the standard error of the estimated intervention effect in a hypothetical cluster-randomized stepped wedge trial.

Consider two models estimating intervention effect, β_1, with I clusters, T time steps, and N individuals per cluster.

Model 1 (Ignoring within-cluster information): For each $j = 2, \ldots, T-1$

$$Y_{ijk} = \beta_{0j} + \beta_{1j} X_{ij} + \alpha_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, I; \; k = 1, \ldots, N,$$

where β_{0j} is the overall mean of Y in the control group at time j, β_{1j} is the fixed effect of intervention on Y at time j, X_{ij} indicates intervention status in cluster i at time j, α_{ij} is the random effect of cluster i on Y at time j with $\alpha_{ij} \sim \mathrm{N}(0, \tau^2)$, and $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$ is the distributional error of Y independent of α_{ij}. The estimates of the cross-sectional effects of the intervention, $\tilde{\beta}_{1j},\ j = 2, \ldots, T-1$, are combined to estimate β_1 by $\tilde{\beta}_1 = \sum_{j=2}^{T-1} \frac{w_j}{w} \tilde{\beta}_{1j}$ where $w_j = \left[\left(\frac{\sigma^2}{N} + \tau^2 \right) \left(\frac{1}{n_{X_j}} + \frac{1}{I - n_{X_j}} \right) \right]^{-1}$, $w = \sum_{j=2}^{T-1} w_j$, and n_{X_j} represents the number of clusters in treatment at time j. Effectively, this estimate is simply the inverse variance weighted mean of the cross-sectional effect estimates at each time step and is hereafter referred to as the cross-sectional (CS) estimate. The sum includes only data from time steps 2 through T-1 since only these times include clusters in opposing treatment status. We also note that the weighting is not necessarily optimal since it does not account for the correlation between the $\tilde{\beta}_{1j}$.

Model 2 (Including within-cluster information):

$$Y_{ijk} = \beta_0 + \beta_1 X_{ij} + \delta_j + \alpha_i + \epsilon_{ijk}, \quad i = 1, \ldots, I;\ j = 1, \ldots, T;\ k = 1, \ldots, N.$$

Here, β_0 is the overall mean Y in the control group at time $j = 0$, β_1 is the fixed effect of intervention in cluster i at time j, X_{ij} indicates intervention status in cluster i at time j, δ_j is the fixed effect of time j ($\delta_1 = 0$ for identifiability), α_i is the random effect of cluster i on Y with $\alpha_i \sim \mathrm{N}(0, \tau^2)$, and $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$ independent of α_i. Model 2 is the model proposed in Hussey and Hughes [1] at the individual level. The estimate of the intervention effect, $\hat{\beta}_1$, is hereafter referred to as the Hussey and Hughes (HH) estimate.

We compared the standard error (SE) of the CS estimate ($\tilde{\beta}_1$) to that of the HH estimate ($\hat{\beta}_1$) for varying degrees of correlation between individuals from the same cluster (intra-class correlation coefficient (ICC) = 0.00, 0.10, 0.20, 0.30, 0.40, and 0.50) under the assumptions that the within- and between-cluster variances are known. The between-cluster variance is then computed from the within-cluster variance to ensure the above ICCs (using the formula ICC = $\frac{\text{between-cluster variance}}{\text{between-cluster variance+within-cluster variance}}$). The relative efficiency of Model 2 relative to Model 1 (the variance of the CS estimate divided by the variance of the HH estimate) was computed for performance comparison of the estimates based on the following design: 30 clusters were randomized to begin receiving the intervention at each of 5 of $T = 6$ total time steps (6 clusters per time step, with the first 6 clusters in treatment condition at time 2 and all clusters in the treatment condition at time 6). Clusters were of equal size of 50.

It can be shown (see Appendix A2.1) that the variance of the CS estimate is

$$\text{Var}\left(\tilde{\beta}_1\right) = \frac{I\left(\frac{\sigma^2}{N}+\tau^2\right)\left(I\sum_{ij}X_{ij} - \sum_j\left(\sum_i X_{ij}\right)^2\right) + 2\tau^2 I\sum_{j<k}\left(I\sum_i X_{ij} - \left(\sum_i X_{ij}\right)\left(\sum_i X_{ik}\right)\right)}{\left(I\sum_{ij}X_{ij} - \sum_j\left(\sum_i X_{ij}\right)^2\right)^2}$$

and the variance of the HH estimate, defined in Hussey and Hughes [1] is

$$\text{Var}\left(\hat{\beta}_1\right) = \frac{I\frac{\sigma^2}{N}\left(\frac{\sigma^2}{N}+T\tau^2\right)}{\left[I\sum_{ij}X_{ij} - \sum_j\left(\sum_i X_{ij}\right)^2\right]\frac{\sigma^2}{N} + \left[\left(\sum_{ij}X_{ij}\right)^2 + IT\sum_{ij}X_{ij} - T\sum_j\left(\sum_i X_{ij}\right)^2 - I\sum_i\left(\sum_j X_{ij}\right)^2\right]\tau^2},$$

where $X_{ij} = 0$ (1) if cluster i is on control (treatment) at time j and sums over time are taken from time 2 to T-1 in the variance of the CS estimate and from time 1 to T in the variance

of the HH estimate. Note that the variance of the CS estimate equals the variance of the HH estimate when there is no between-cluster variance ($\tau^2 = 0$), since $I \sum_i X_{ij} = \left(\sum_i X_{ij}\right)^2$ at times $j = 1$ and $j = T$.

The SE and relative efficiency (RE) comparisons are shown in Table 2.1. We note that, except for the case ICC = 0 when they are equal, the SEs of the CS estimate from Model 1 are always greater than those of the HH estimate from Model 2 with the discrepancy increasing with increasing ICC. This indicates that the HH estimate from Model 2 is at least as efficient as the CS estimate from Model 1 and that ICC has more of an effect on the SEs of the CS estimate than the SEs of the HH estimate. This demonstrates that a cross-sectional analysis can be an inefficient approach to take when attempting to account for time trends.

Table 2.1 Comparing Standard Error (SE) and Relative Efficiency (RE) of Treatment Effect Estimates from a Model Averaging Cross-Sectional Estimates (Model 1) with Estimates from a Model Explicitly Accounting for Time Trend (Model 2).

ICC(%)	Model 1	Model 2		Model 2*	
	SE	SE	RE	SE	RE
0	0.071	0.071	1.000	0.071	1.000
10	0.273	0.098	7.703	0.112	5.984
20	0.402	0.099	16.379	0.114	12.489
30	0.522	0.100	27.528	0.114	20.858
40	0.649	0.100	42.418	0.115	32.017
50	0.794	0.100	63.251	0.115	47.642

* Model 2 ignoring time points when clusters are either all in control or all in intervention status.

One other drawback to a cross-sectional analysis is that it makes use of only the data from time steps where there is at least one cluster on control and at least one cluster on treatment (otherwise there's no way to get a time-specific estimate of effect), discarding

valuable data from the first (if a baseline step) and last time steps. If the comparison of the CS estimate is made with the HH estimate at the time points with at least one treatment and one control cluster (arguably a more fair comparison), then the HH estimate from Model 2 is only slightly less efficient than before compared to the CS estimate from Model 1 (when there is correlation between individuals in a cluster) as a result of now slightly higher SEs for the HH estimate from Model 2. These higher SEs are a result of the loss of information with the exclusion of baseline (if used) and last time steps when clusters are all in either control or intervention status. This reinforces the importance of the contribution of within-cluster comparison information to intervention effect estimation.

The SWD provides an opportunity to maximize use of the resulting data for intervention effect estimation. There are significant drawbacks to ignoring the between-unit or the within-unit information. We have shown that relying solely on between-cluster comparisons in a SWD cluster-randomized trial can result in very inefficient estimates of the intervention effect.

3 — ESTIMATING LAGGED EFFECTS IN THE SWD OF CLUSTER RANDOMIZED CLINICAL TRIALS

3.1 Introduction

The stepped wedge design (SWD) of cluster-randomized clinical trials has previously been defined and critiqued [1, 2]. Briefly, in a SWD trial the intervention or treatment is introduced to each cluster (e.g., community, county, group) at a random time. Each cluster spends time in the control state before crossing over into the intervention state, creating a stepped wedge schematic as seen in Figure 3.1. Note that crossover is unidirectional and a cluster stays in the intervention state once it crosses over. The outcome measure is collected from each cluster either at the cluster or individual level during each time step, which provides information on the intervention effect based on both within- and between-unit comparisons.

		Time				
		1	2	3	4	5
Cluster	1	0	1	1	1	1
	2	0	0	1	1	1
	3	0	0	0	1	1
	4	0	0	0	0	1

Figure 3.1 Stepped Wedge Trial Design indicating treatment (1) assignment (one unit per time step) and control (0) assignment.

For the purposes of analysis and for calculating power, it is typically assumed that the intervention has been fully introduced and has reached its maximal effect during the time step in which the intervention was assigned. For example, if the schematic in Figure 3.1 refers to a SWD trial investigating the effect of a vaccine on the incidence of a particular infection, a standard calculation of the power of such a trial assumes that all individuals (or the maximum possible) in cluster 1 have been vaccinated and the vaccination has reached its maximal effect within time step 2, and that the effect is maintained for the duration of the study. In other words, standard analyses and power calculations assume there is no delay in the implementation or effect of the intervention and no waxing or waning of the effect over time. Therefore, if there is a delay and it is ignored, there are implications for effect estimation and power.

In general, a delayed effect occurs when the full effect of the intervention is not realized in the time step in which the intervention is introduced. We outline implications for estimating the effect of the intervention below.

A delayed effect can occur in two ways: (a) when the intervention is fully delivered within the interval in which it is introduced but the mechanism by which the intervention acts is not fully operational within the interval (which we'll call "a lag in effect"); and/or (b) when there is a delay in the delivery of the intervention (which we'll call a "coverage delay"). For example, vaccines work in general by inducing the production of antibodies, which attack infectious agents when they are encountered in the future. In this example, a lag in effect occurs when the vaccine has been given to every individual in the assigned

cluster but it hasn't had a chance to induce the maximal production of antibodies before the next time step, so the individual is not fully protected. A coverage delay occurs when not all individuals (or, at least, not the maximum achievable number of individuals) in the cluster have been vaccinated before the next time step starts. If the delay is anticipated and its nature is known, a lag in effect can be avoided by designing wider time steps to allow for full (or greater) impact of the intervention within the time step where the intervention is introduced; coverage delays can often be avoided by better planning around intervention implementation, including contingency planning.

The Gambia Hepatitis Study [7] provides an example of dealing with a delay in effect by careful study design. The Gambia Hepatitis Study was designed to determine the effect of hepatitis B vaccination on the incidence of hepatocellular cancer and other chronic liver disease in The Gambia. The aim was to implement a course of hepatitis B vaccination in infants (along with the normal vaccination course) in clusters defined by 17 geographical regions surrounding 17 national health centers over a period of four years. Beginning in July, 1986, one region was randomly assigned to begin vaccinating all babies born in that region with the hepatitis B vaccine. Regions were stepped into the program at 3-month intervals. However, any effect that the intervention (hepatitis B vaccination) might have on the outcome (hepatocellular cancer and other chronic liver diseases) would not be expected for 30 to 40 years after vaccination. Therefore, identification data were collected on all the babies (including handprints and footprints) and a national liver cancer surveillance system was set up to assign future cases of liver disease back to a time step and cluster

so that appropriate estimates of the effect of the national hepatitis vaccination program on chronic liver disease outcome could be obtained. While an intent-to-treat analysis might appropriately ignore a coverage delay (type (b)), a per-protocol analysis may incorporate coverage delays into the analysis. If detailed information on the nature of the delay is available then, following the suggestions of Hussey and Hughes [1], one can use a fractional treatment covariate in the analysis to represent the proportion of the maximal coverage that has been provided. The Expedited Partner Therapy (EPT) Trial [15] in Washington State used this approach in a per-protocol analysis to estimate the effect of expedited partner treatment and partner services on rates of gonorrhea and chlamydial infection in Washington State.

If the nature of the delay is not known it can be estimated but methods of doing so and their consequence for power are not well understood. Our aim is to estimate the long-term stable intervention effect (the long-term effect of the intervention) by explicitly modeling and estimating delays. We focus on a type (a) delay ("lag in effect") in this paper and present a two-step method of estimation. We assess bias, precision and power for estimating the long-term effect of the intervention and the duration of the lag in effect. In Section 3.2 of this paper, we detail the method of estimation and we report results of simulations in Section 3.3. We apply this method to data from the EPT Trial [15] in Section 3.4 and Section 3.5 outlines a method of calculating power in a SWD trial when a lag in effect is anticipated. We discuss the implications of this work in Section 3.6 and present a plan for future research in this context.

3.2 Methods

If there is no lag in effect, we can fit a model for prevalences (e.g., a random effect model [16] or a generalized estimation equations (GEE) model [18]), which provides an estimate of relative risk. Consider a cluster-randomized SWD trial designed to investigate the effect of an intervention on an event outcome Y. A cross-sectional sample is used to determine the outcome at each of J time steps in I clusters based on K_{ij} individuals at time step j in cluster i. We use K_{ij} to acknowledge the possibility that the number of individuals could vary by cluster and time step. Y_{ijk} then indicates the event in cluster i at time step j for the k_{ij}^{th} individual. The random effect model can be represented at the cluster level as

$$\log\left(p_{ij}|X_{ij}\right) = \alpha + \beta_j + \theta X_{ij} + \nu_i; \qquad i = 1, \ldots, I;\ j = 1, \ldots, J, \tag{3.1}$$

where p_{ij} is the prevalence in cluster i at time j, binary outcome, $Y_{ijk} \sim \text{Bern}\left(p_{ij}\right)$, α is the log population prevalence at time step $j = 1$ when all clusters are in the control condition, β_j is the log RR associated with time on study j ($\beta_1 = 0$ for identifiability), θ is the log RR associated with the intervention, X_{ij} is a 0/1 indicator which is 1 when cluster i is in the intervention condition, and $\nu_i \sim \text{N}\left(0, \tau^2\right)$ is a random effect for cluster i. The inclusion of the random effect, ν_i, accounts for within-cluster dependency.

A lag in effect can be estimated in two steps. In the first step, we estimate the effect of the intervention at different lengths of time on the intervention, i.e., a time-on-intervention effect. In the second step, we fit the trajectory suggested by the step 1 time-on-intervention effects in a non-linear model to get estimates of the long-term effect of the intervention and

the duration of the lag.

Step 1

In this step we fit a model for prevalences that provides an estimate of time-specific relative risks and can be represented at the cluster level as

$$\log\left(p_{ijl}|X_{il}\right) = \alpha + \beta_j + \theta_l X_{il} + \nu_i; \qquad i = 1,\ldots,I;\ j = 1,\ldots,J;\ l = 0,\ldots,J-1, \quad (3.2)$$

where p_{ijl} is the prevalence in cluster i at time j after l time steps on intervention, binary outcome $Y_{ijlk} \sim \text{Bern}\left(p_{ijl}\right)$, α is the log population prevalence at time step $j = 1$ when all clusters are in the control condition (i.e., $l = 0$), β_j is the log RR associated with time on study j ($\beta_1 = 0$ for identifiability), θ_l is the log RR l time steps after the intervention introduction (i.e., intervention is introduced at $l = 1$ and $\theta_0 = 0$ for identifiability), X_{il} is a 0/1 indicator which is 1 when cluster i has been in the intervention condition for l time steps and 0 otherwise, and $\nu_i \sim \text{N}\left(0, \tau^2\right)$ is a random effect for cluster i. As with the no-lag model (3.1), a GEE model [18] may be used in this step for marginal estimates of time-on-intervention effects. Model (3.2) can be fit in most general-purpose statistical packages and estimates of log RRs $\left(\hat{\theta}_L = \left(\hat{\theta}_1,\ldots,\hat{\theta}_{J-1}\right)'\right)$ and their estimated variance-covariance matrix $\left(\hat{\Sigma}_L = \hat{\text{Var}}\left(\hat{\theta}_L\right)\right)$ can be retrieved for use in the second step.

The estimated log RRs, $\hat{\theta}_L$, can be plotted against time on intervention to form a trajectory over time that gives us insight into the long-term effect of the intervention and the duration of the lag. Though time on intervention is discrete in model (3.2), Figure 3.2 depicts a few different trajectories that the log RRs could take over continuous time. In example

(1), log RR goes from 0 before intervention assignment to a long-term log RR of log(0.5). In examples (2) and (3), the intervention has no effect in the long run after having some initial effect. We focus on example (1) in this paper as examples (2) and (3) may require too complex a second step model, particularly in designs with relatively few time steps. However, the methods outlined here may, in principle, be used with any identifiable model for the time trend. We envision the exponential function in example (1) as a potential model for drug effect following a pharmacokinetic model of serum concentration of the drug.

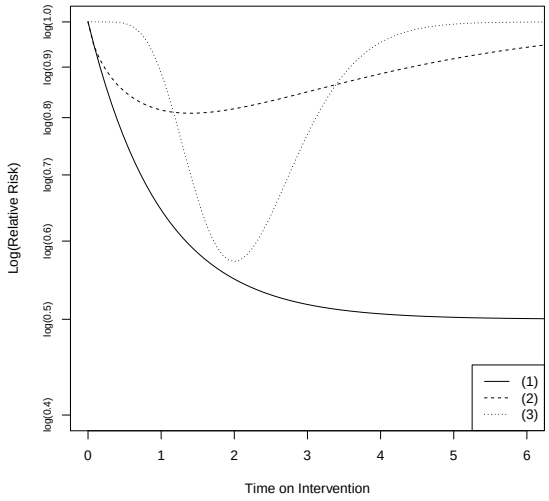

Figure 3.2 Examples of trajectories of time-specific intervention effects over time.

Step 2

In step 2, we will use the time-specific log RR estimates to fit a model relating time on intervention, l, to effect size. In the example described in Section 3.4 we use a non-linear regression of the form

$$\hat{\theta}_l = \theta\left(1 - e^{-\frac{l}{d}}\right) + \epsilon_l; \qquad l = 1, \ldots, J-1;\ (\epsilon_1, \ldots, \epsilon_{J-1})' \sim \text{MVN}\left(\mathbf{0}, \hat{\Sigma}_L\right), \tag{3.3}$$

where $\hat{\theta}_l$ represents the estimated log RR after l time steps (obtained in Step 1), θ is the long-term log RR, d is a lag duration parameter, and ϵ_l are the normally distributed errors on the time-specific effect estimates, the variances of which have been estimated in Step 1 so that $\hat{\text{Var}}(\epsilon_l)$ is the element in the l^{th} row and the l^{th} column of $\hat{\Sigma}_L$. We model the time-on-intervention estimates on the log scale because of the more flexible range of log RR values, which can be negative and converge to normality faster. The model in equation (3.3) depicts a trajectory with log RR assumed to be 0 (RR = 1) at $l = 0$ and the time-specific log RR attaining the long-term log RR at $l = \infty$. d is a positive real number and large (small) values of d determine if there is a slow (fast) rate of convergence to the long-term effect. One way to think of d in (3.3) is that the intervention is about $\left(1 - e^{-\frac{1}{d}}\right) * 100$ percent effective at the end of the 1^{st} time step. There is no lag in the intervention effect when $d = 0$.

Applying maximum likelihood estimation (MLE) theory to step 1 (estimating equation theory if a GEE model is used in step 1), $\hat{\theta}_L \sim \text{MVN}(\theta_L, \Sigma_L)$ asymptotically in the number of clusters. If these asymptotic normality assumptions hold then, from model (3.3), the likelihood

$$L\left(\theta, d|\hat{\theta}_L\right) \propto e^{-\frac{1}{2}\left(\hat{\theta}_L - \theta\mu_d\right)' \hat{\Sigma}_L^{-1}\left(\hat{\theta}_L - \theta\mu_d\right)}, \tag{3.4}$$

where $\mu_d = \left(1\text{-}e^{-\frac{1}{d}}, \ldots, 1\text{-}e^{-\frac{J-1}{d}}\right)'$. A closed-form expression for the MLE of θ is $\frac{\hat{\theta}_L' \hat{\Sigma}_L^{-1} \hat{\mu}_d}{\hat{\mu}_d' \hat{\Sigma}_L^{-1} \hat{\mu}_d}$.

There is no closed-form expression for the MLE of d. Therefore, we proceed with estimation by iterating between the formula for the estimate of θ and estimating d numerically based on (3.4) (conditional on $\hat{\theta}_L$) via a Newton-Raphson type algorithm. We note that if $\theta = \log(1) = 0$, d is not estimable, which may cause difficulty in the estimation of these parameters when $\hat{\theta}$ is close to 0. In particular, the estimation procedure may be sensitive to starting values of the estimate of θ. However, the trajectory of the $\hat{\theta}_L$ will provide insight into which side of 0 the initial value should be. If the $\hat{\theta}_L$ are all near 0, suggesting no overall effect of the intervention, then the 2^{nd} step model is unnecessary.

The estimated variance-covariance matrix ($\hat{\Sigma}_L$) from Step 1 is a component of the likelihood function (3.4), from which the information matrix of the long-term intervention effect (θ) and the duration of the lag (d) may be derived to compute standard error estimates for $\hat{\theta}$ and $\hat{d}$. Additionally, standard normal quantiles ($\frac{\alpha}{2}^{\text{th}}$ and $\left(1-\frac{\alpha}{2}\right)^{\text{th}}$) may be used along with the estimated standard errors for $\hat{\theta}$ to determine the coverage probability of the $(1-\alpha)*100\%$ confidence interval for θ, which measures the actual confidence of this approach to estimating θ. However, this method will underestimate the variance of the 2^{nd} step estimates as it ignores the additional variation in estimating θ and d from the Step 1 estimates. Therefore, robust estimates of standard errors were obtained using a parametric bootstrap in the following way.

1. Re-sample predicted outcomes from the model in Step 1 (3.2) 1000 times and re-estimate to obtain 1000 new estimates of θ_L and Σ_L.

2. Execute Step 2 on the 1000 estimates of θ_L and Σ_L to obtain 1000 estimates of θ and

d.

3. Use the variance of the 1000 estimates of θ and d, respectively, as their estimated variances.

In other words, re-sample predicted outcomes based on (3.2), execute Steps 1 and 2 1000 times for a joint empirical distribution of $\hat{\theta}$ and $\hat{d}$ in each simulation, and estimate the variance of the parameter estimates with the empirical variance. We also rely on the empirical distribution of $\hat{\theta}$ to estimate the actual coverage probability. The 0.025^{th} and 0.975^{th} quantiles of these 1000 estimates of θ in each simulation can be used to determine the actual coverage of the 95% confidence interval for this two-step approach to estimation of the long-term effect of the intervention. See Section 3.3.3 for more details. All analyses were performed using `R` version 3.0.1.

3.3 Simulations

We simulated cluster-randomized SWD trials measuring the effect of an intervention on a prevalence outcome in each of three scenarios, 1) when there is no lag in effect (i.e., $d = 0$ so that $\theta_1 = \ldots = \theta_{J-1} = \theta$), 2) when there is a lag in effect and it is ignored, and 3) when there is a lag in effect and it is estimated as proposed above. In scenarios (2) and (3) we considered a short lag ($d = 0.5$) where the full effect of the intervention is attained fairly early on in the trial and a long lag ($d = 1.4$) when the full effect is attained towards or after the end of the trial. We repeated each of these 5 combinations for varying number of clusters ($I = 12$, 24 and 48), number of time steps ($J = 5$, 7 and 9), number of individuals sampled at each time

step in each cluster (K = 20, 50 and 100), and between-cluster variances on the log scale (τ^2 = 0.0, 0.01, 0.04, 0.09, 0.16 and 0.25) to investigate their effect on bias and precision for the intervention effect estimate and bias of the estimate of the duration of the lag. It was assumed that the intervention lowered prevalence by 50% ($\log \text{RR} = \theta = \log(0.5)$) in the long run from a baseline prevalence of 10% in all scenarios. The short and long lags resulted in log RRs of $\log(0.55)$ and $\log(0.70)$ at the first time step, respectively. One thousand simulations were run for each scenario. The following summarizes the results of these simulations. Less than 0.02% of simulations were excluded because parameter estimates were on the boundary set for the parameters. Results were similar across different values of τ^2, therefore we present results for $\tau^2 = 0$.

3.3.1 Bias of the Estimate of θ

When there was no lag in the effect of the intervention, there was little or no bias in the estimation of θ regardless of I, J, or K. When there was a lag in effect and the lag was ignored, there was a distinct bias in estimating θ that was smaller when the lag duration is short (compared to when the lag duration was long) and which does not appear to decrease with increasing I, J, or K. However, when the lag was estimated, though the bias in estimating θ was still smaller when the lag duration was short (compared to a long lag duration), the bias decreased with increasing I, J, and K. See Tables 3.1-3.3.

To investigate how I affects bias of the intervention effect estimate when the lag was ignored versus when it was estimated, we limited the focus to the case when $J = 7$ and $K = 100$. As indicated in Table 3.1, the intervention effect tends to be overestimated when the lag

was ignored but slightly underestimated and much less biased when the lag was estimated.

Table 3.1 Intervention effect estimates (true log RR = log(0.5)) and percent bias by number of clusters (I) when the short and long lag durations are ignored versus estimated. Number of time steps (J) = 7, number of individuals sampled per cluster per time step (K) = 100, and $\tau^2 = 0$.

	$I = 12$		$I = 24$		$I = 48$	
	$\hat{\theta}$	% Bias	$\hat{\theta}$	% Bias	$\hat{\theta}$	% Bias
No Lag	log(0.498)	0.65	log(0.499)	0.28	log(0.499)	0.25
Ignore Short Lag	log(0.519)	5.33	log(0.517)	4.94	log(0.522)	6.10
Ignore Long Lag	log(0.614)	29.58	log(0.617)	30.42	log(0.617)	30.23
Estimate Short Lag	log(0.489)	3.21	log(0.494)	1.65	log(0.499)	0.19
Estimate Long Lag	log(0.458)	12.76	log(0.483)	5.02	log(0.494)	1.88

When the lag in effect was ignored, the percent bias in estimating θ was larger when d was larger, ranging from about 5% to 6% when $d = 0.5$ to 30% when $d = 1.4$. Estimation of the lag greatly improves the estimation of the intervention effect. When the short lag was estimated, the log RR was estimated to be log(0.489) and log(0.499) at $I = 12$ and $I = 48$, respectively, a decrease in percent bias from 3.21% to 0.19%. A much bigger decrease in percent bias of the intervention effect estimate was seen when the long lag was estimated, going from 12.76% at $I = 12$ (log RR estimate = log(0.458)) to 1.88% at $I = 48$ (log RR estimate = log(0.494)).

To investigate how J affects bias of the intervention effect when the lag was ignored versus when it was estimated, we limited the focus to the case when $I = 24$ and $K = 100$. The bias in the intervention effect estimate was smaller when the trial was carried out in 9 time steps rather than 7 or 5, whether the lag was ignored or estimated and whether the lag duration was short or long. However, the bias was smaller and decreased faster with

increasing number of time steps when the lag was estimated rather than ignored as depicted in Table 3.2. If the short lag was estimated, the log RR estimate was log(0.485) with a 4.44% bias at $J = 5$ but increased to log(0.499) with a 0.17% bias at $J = 9$. Estimating the long lag yielded a log RR estimate of log(0.415) (26.90% bias) when $J = 5$, which increased to log(0.499) (0.17% bias) when $J = 9$.

Table 3.2 Intervention effect estimates (true log RR = log(0.5)) and percent bias by number of time steps (J) when the short and long lag durations are ignored versus estimated. Number of clusters (I) = 24, number of individuals sampled per cluster per time step (K) = 100, and $\tau^2 = 0$.

	$J = 5$		$J = 7$		$J = 9$	
	$\hat{\theta}$	% Bias	$\hat{\theta}$	% Bias	$\hat{\theta}$	% Bias
No Lag	log(0.501)	0.34	log(0.499)	0.28	log(0.499)	0.23
Ignore Short Lag	log(0.528)	7.96	log(0.517)	4.94	log(0.517)	4.86
Ignore Long Lag	log(0.648)	37.35	log(0.617)	30.42	log(0.598)	25.94
Estimate Short Lag	log(0.485)	4.44	log(0.494)	1.65	log(0.499)	0.17
Estimate Long Lag	log(0.415)	26.90	log(0.483)	5.02	log(0.499)	0.17

As with the number of clusters and the number of time steps used in the trial, the bias of the estimate of θ decreased with increasing number of individuals sampled per cluster per time step (K) and is smaller when the lag is shorter. Percent bias ranged from about 5% when the short lag was ignored to about 30% when the long lag was ignored, regardless of K. However, if the short lag was estimated, percent bias went from 7.13% at $K = 20$ to 1.65% at $K = 100$ and decreased from 38.40% to 5.02% when the long lag was estimated. We saw in Tables 3.1 and 3.2 that, when K = 100, bias in the intervention effect estimate is less when the lag is estimated rather than ignored. In Table 3.3, this pattern is broken when $K = 20$, with bias being larger when the lag is estimated (and is the case for other τ^2 as

well), underscoring the importance of the number of individuals used in the trial to getting good estimates of the long-term intervention effect when there is a lag in effect.

Table 3.3 Intervention effect estimates (true log RR = log(0.5)) and percent bias by individuals sampled per cluster per time step (K) when the short and long lag durations are ignored versus estimated. Number of clusters (I) = 24, number of time steps (J) = 7, and $\tau^2 = 0$.

	$K = 20$		$K = 50$		$K = 100$	
	$\hat{\theta}$	% Bias	$\hat{\theta}$	% Bias	$\hat{\theta}$	% Bias
No Lag	log(0.500)	0.12	log(0.505)	1.51	log(0.499)	0.28
Ignore Short Lag	log(0.520)	5.60	log(0.518)	5.09	log(0.517)	4.94
Ignore Long Lag	log(0.619)	30.69	log(0.618)	30.68	log(0.617)	30.42
Estimate Short Lag	log(0.476)	7.13	log(0.493)	2.13	log(0.494)	1.65
Estimate Long Lag	log(0.384)	38.40	log(0.462)	11.26	log(0.483)	5.02

3.3.2 Bias of the Estimate of d

When the short lag ($d = 0.50$) is estimated, we get only moderately biased estimates of the duration regardless of τ^2 and the number of clusters (I) used as seen in Table 3.4. Again, J is fixed at 7, K is fixed at 100, and results are presented for $\tau^2 = 0$ since they are similar for all τ^2. The biggest percent bias seen was 7.47% at $I = 12$, which decreased to 6.07% at $I = 48$. We also see better estimates of the long lag with increasing I though there is a tendency for overestimation. The long lag duration ($d = 1.40$) was estimated to be about 1.92 with a percent bias of 37% when $I = 12$ and about 1.47 with a percent bias of 4.90% when $I = 48$.

Table 3.4 Short ($d = 0.5$) and long ($d = 1.4$) lag duration estimates and percent bias by number of clusters (I). Number of time steps (J) = 7, number of individuals sampled per cluster per time step (K) = 100, and $\tau^2 = 0$.

	$I = 12$		$I = 24$		$I = 48$	
	$\hat{d}$	% Bias	$\hat{d}$	% Bias	$\hat{d}$	% Bias
Short Lag	0.537	7.47	0.466	6.76	0.470	6.07
Long Lag	1.92	36.99	1.59	13.78	1.47	4.90

Fixing I at 24, K at 100, and τ^2 at 0, we see also see decreases in the percent bias of estimates of the short and long lag durations with increasing number of time steps. See Table 3.5. The percent bias in the short lag estimate was 6.74% when $J = 5$ and decreased to 2.80% when $J = 9$. The effect of the number of time steps is much more pronounced with estimation of the long lag duration. Percent bias decreased from 63.06% when $J = 5$ to 4.26% when $J = 9$.

Table 3.5 Short ($d = 0.5$) and long ($d = 1.4$) lag duration estimates and percent bias by number of time steps (J). Number of clusters (I) = 24, number of individuals sampled per cluster per time step (K) = 100, and $\tau^2 = 0$.

	$J = 5$		$J = 7$		$J = 9$	
	$\hat{d}$	% Bias	$\hat{d}$	% Bias	$\hat{d}$	% Bias
Short Lag	0.534	6.74	0.466	6.75	0.486	2.80
Long Lag	2.28	63.06	1.59	13.78	1.46	4.26

As with I and J, percent bias in the lag estimates decreased with increasing K and was less when d was smaller. See Table 3.6.

Table 3.6 Short ($d = 0.5$) and long ($d = 1.4$) lag duration estimates and percent bias by number of individuals sampled per cluster per time step (K). Number of clusters (I) = 24, number of time steps (J) = 7, and $\tau^2 = 0$.

	$K = 20$		$K = 50$		$K = 100$	
	$\hat{d}$	% Bias	$\hat{d}$	% Bias	$\hat{d}$	% Bias
Short Lag	0.732	46.48	0.494	1.19	0.466	6.75
Long Lag	3.92	180.20	1.95	39.18	1.59	13.78

3.3.3 Coverage and Standard Error of the Estimate of θ

When there was no lag in effect or when there was a lag but it was ignored, coverage is defined as the percentage of simulations in which the true log RR (θ) falls within the interval

$$\left[\hat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{\text{Var}\left(\hat{\theta}\right)},\ \hat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{\text{Var}\left(\hat{\theta}\right)}\right],$$

where $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)^{\text{th}}$ quantile of the standard normal distribution, and is a measure of the confidence in the estimate of the log RR. $\text{Var}\left(\hat{\theta}\right)$ was the typical model estimate of the variance of the log RR estimate from the model in (3.1) and can be retrieved from standard statistical software. When we estimated the long-term log RR as outlined in Section 3.2 (i.e., while estimating any lag in effect), coverage is defined as the percentage of simulations in which the true log RR (θ) falls within the interval

$$\left[z^*_{\frac{\alpha}{2}},\ z^*_{1-\frac{\alpha}{2}}\right],$$

where $z^*_{\frac{\alpha}{2}}$ and $z^*_{1-\frac{\alpha}{2}}$ are the $\left(\frac{\alpha}{2}\right)^{\text{th}}$ and $\left(1-\frac{\alpha}{2}\right)^{\text{th}}$ quantiles, respectively, of the empirical distri-

bution of $\hat{\theta}$ generated by bootstrap in each simulation. For significance level $\alpha = 0.05$, we expect that $\mathrm{P}\left(\hat{\theta} - 1.96\sqrt{\mathrm{Var}\left(\hat{\theta}\right)} \leq \theta \leq \hat{\theta} + 1.96\sqrt{\mathrm{Var}\left(\hat{\theta}\right)}\right) = 0.95$ and $\mathrm{P}\left(z^*_{\frac{\alpha}{2}} \leq \theta \leq z^*_{1-\frac{\alpha}{2}}\right) = 0.95$ to maintain a type I error rate of 5%.

When there was no lag in effect, coverage was maintained at about the nominal 95% regardless of I, J, or K. Coverage was slightly less than nominal if the short lag was ignored but dropped substantially when the long lag was ignored, a consequence of the large bias in estimating θ when the long lag was ignored. Estimation of the short lag duration brought coverage back up to nominal. While estimation of the long lag duration resulted in substantial recovery of the coverage probability, the nominal percentage was only achieved at the highest levels of I, J, and K. See Tables 3.7-3.9.

Table 3.7 Coverage of the intervention effect estimate by number of clusters (I). Number of time steps (J) = 7, number of individuals sampled per cluster per time step (K) = 100, and $\tau^2 = 0$.

	$I = 12$	$I = 24$	$I = 48$
No Lag	95.22	95.40	95.09
Ignore Short Lag	93.81	93.10	88.98
Ignore Long Lag	53.66	22.00	03.61
Estimate Short Lag	95.10	95.60	
Estimate Long Lag	92.90	94.80	

Coverage decreased with increasing number of clusters when the lag was ignored. However, coverage increased with increasing number of clusters when the lag was estimated. See Table 3.7.

There was no pattern in the coverage with increasing number of time steps when the lag was ignored. When the short lag was estimated, the coverage was maintained at about 95%.

There was an increase in coverage (to just over 95%) with increasing number of time steps when the long lag was estimated. When the long lag was estimated, coverage increases from 91.90% when 5 times steps were used to 95.80% when 9 time steps were used. See Table 3.8.

Table 3.8 Coverage of the intervention effect estimate by number of time steps (J). Number of time clusters (I) = 24, number of individuals sampled per cluster per time step (K) = 100, and $\tau^2 = 0$.

	$J = 5$	$J = 7$	$J = 9$
No Lag	94.79	95.40	95.50
Ignore Short Lag	91.46	93.10	91.60
Ignore Long Lag	22.60	22.00	23.10
Estimate Short Lag	94.80	95.60	94.89
Estimate Long Lag	91.90	94.80	95.80

Similar to I, coverage decreased with increasing K when the lag was ignored. However coverage increased with increasing number of individuals sampled per cluster per time step when the lag was estimated. See Table 3.9. When the long lag was estimated, coverage increased from 90.60% when $K = 20$ to 94.80% when $K = 100$.

Table 3.9 Coverage of the intervention effect estimate by number of individuals sampled per cluster per time step (K). Number of time clusters (I) = 24, number of time steps (J) = 7, and $\tau^2 = 0$.

	$K = 20$	$K = 50$	$K = 100$
No Lag	95.24	95.00	95.40
Ignore Short Lag	94.25	93.27	93.10
Ignore Long Lag	81.42	50.10	22.00
Estimate Short Lag	94.10	94.90	95.60
Estimate Long Lag	90.60	93.90	94.80

As a measure of the accuracy of the log RR estimate ($\hat{\theta}$), we present estimates of the standard errors (SEs) of $\hat{\theta}$ computed as the square root of the variance of the log RR es-

timates over all simulations. As a comparative measure of accuracy (and to evaluate the consistency of the SEs) we present the average standard deviation over all simulations, where the standard deviation in each simulation is the square root of the variance of $\hat{\theta}$ estimated by parametric bootstrap.

Estimated standards errors for $\hat{\theta}$ are about the same whether there is no lag or there is a lag but it is ignored. See Tables 3.10-3.12. This is because the method of estimation is the same in all three scenarios, though the invalidity of ignoring the lag is evident in the bias in estimation of θ. Estimating the lag increases the standard error estimates (relative to estimates that do not incorporate a lag) because the two-step method we propose adds variability from the Step 1 estimation of time-on-intervention effects, though it decreases bias in estimating θ. We note that standard error estimates are, in most cases, more than twice as large when the lag duration is long than when the lag duration is short.

As expected, increasing the number of clusters used in the trial decreases the standard error of $\hat{\theta}$. See Table 3.10.

Table 3.10 Standard error (SE) computed over all simulations and average of the estimated SD at each simulation ($\overline{\text{SD}}$) by number of clusters (I). Number of time steps (J) = 7, number of individuals sampled per cluster per time step (K) = 100, and $\tau^2 = 0$.

	$I = 12$		$I = 24$		$I = 48$	
	SE	$\overline{\text{SD}}$	SE	$\overline{\text{SD}}$	SE	$\overline{\text{SD}}$
No Lag	0.1131	0.1142	0.0810	0.0806	0.0572	
Ignore Short Lag	0.1136	0.1130	0.0798	0.0800	0.0542	
Ignore Long Lag	0.1099	0.1087	0.0767	0.0768	0.0570	
Estimate Short Lag	0.1496	0.1934	0.0975	0.1037	0.0695	
Estimate Long Lag	0.4002	0.4702	0.1951	0.2973	0.1155	

Similarly, standard error estimates decrease with increasing number of time steps (Table

3.11). However, with estimation of the lag duration, the gaps between the SEs and the $\overline{\text{SD}}$s increase with increasing number of time steps. This is expected since there are more step 1 estimates as J increases, which will increase the resulting variability going into step 2 of our proposed method of estimating θ.

Table 3.11 Standard error (SE) computed over all simulations and average of the estimated SD at each simulation ($\overline{\text{SD}}$) by number of time steps (J). Number of clusters (I) = 24, number of individuals sampled per cluster per time step (K) = 100, and $\tau^2 = 0$.

	$J = 5$		$J = 7$		$J = 9$	
	SE	$\overline{\text{SD}}$	SE	$\overline{\text{SD}}$	SE	$\overline{\text{SD}}$
No Lag	0.1019	0.1007	0.0810	0.0806	0.0697	0.0694
Ignore Short Lag	0.0997	0.0994	0.0798	0.0800	0.0683	0.0689
Ignore Long Lag	0.0949	0.0945	0.0767	0.0768	0.0663	0.0667
Estimate Short Lag	0.1583	0.1809	0.0975	0.1037	0.0812	0.0824
Estimate Long Lag	0.5497	0.5448	0.1951	0.2973	0.1209	0.1563

Standard error estimates also decrease with increasing number of individuals sampled per cluster per time step. See Table 3.12. The gaps between the SEs and the $\overline{\text{SD}}$s narrow with increasing K, suggesting consistency of the standard error estimates of $\hat{\theta}$ as number of individuals sampled per cluster per time step increases.

Table 3.12 Standard error (SE) computed over all simulations and average of the estimated SD at each simulation ($\overline{\text{SD}}$) by number of individuals sampled per cluster per time step (K). Number of clusters (I) = 24, number of time steps (J) = 7, and $\tau^2 = 0$.

	$K = 20$		$K = 50$		$K = 100$	
	SE	$\overline{\text{SD}}$	SE	$\overline{\text{SD}}$	SE	$\overline{\text{SD}}$
No Lag	0.1841	0.1812	0.1161	0.1141	0.0810	0.0806
Ignore Short Lag	0.1767	0.1791	0.1167	0.1132	0.0798	0.0800
Ignore Long Lag	0.1711	0.1723	0.1122	0.1088	0.0767	0.0768
Estimate Short Lag	0.3413	0.3911	0.1490	0.1907	0.0975	0.1037
Estimate Long Lag	0.7538	0.6285	0.4029	0.4505	0.1951	0.2973

Note the greater decrease in standard error that accompanies a small increase in the number of time steps (J) compared to a doubling in the number of clusters (I) or the number of individuals sampled per cluster per time step (K).

These results suggest that the parametric bootstrap method of estimating the standard error of $\hat{\theta}$ is appropriate, as the average standard error computed in this way over all simulations approaches the standard deviation of the estimates over all simulations as I and K increase.

3.4 Application to Washington State EPT Trial

Expedited partner therapy (EPT) is the treatment of sex partners of persons who have been diagnosed with a treatable sexually transmitted disease without the partners themselves being evaluated by a clinician. One of the aims of the Washington State EPT Trial was to estimate the effect of EPT on the prevalence of chlamydial infection among women 14-25 years of age who were tested in clinics throughout the state supported by the Centers for Disease Controls Infertility Prevention Project between 2006 and 2010. EPT had previously been evaluated in four randomized controlled trials and was shown to decrease the rates of chlamydial and gonorrheal infections. Therefore, this latest trial was designed as a stepped wedge, community-randomized trial in 22 counties and 132 health clinics nested within those counties. There were 14 time steps of 2-4 month intervals excluding baseline. We wish to determine if there may be a lag in any effect that EPT might have on chlamydia rates in these counties.

We first fit a random effect model ignoring a possible lag in the effect of EPT. This model was adjusted for a fixed effect of time on study and with a random intercept for counties and a random intercept for clinics nested within counties. That is, on the individual level (since county and clinic sizes are not approximately equal),

$$\log\left(p_{rstu}|X_{rt}\right) = \alpha + \beta_t + \theta X_{rt} + \nu_r + \omega_{s(r)};$$
$$r = 1, \ldots, 22;\ s = 1, \ldots, 132;\ t = 0, \ldots, 14;\ u = 1, \ldots, n_{s(r)}, \qquad (3.5)$$

where, in person u in clinic s of county r at time t, p_{rstu} is the risk of chlamydia where chlamydial infection at time t in clinic s of county $r \sim \text{Bern}\left(p_{rstu}\right)$, α is the log risk of chlamydia at time step $t = 0$ when all counties are in the control condition, β_t is the log RR associated with time on study t ($\beta_0 = 0$ for identifiability), θ is the log RR associated with EPT, X_{rt} is a 0/1 indicator which is 1 when county r is using EPT at time t, $\nu_r \sim \text{N}\left(0, \sigma_r^2\right)$ is a random effect for county r, $\omega_{s(r)} \sim \text{N}\left(0, \sigma_{s(r)}^2\right)$ is random nested effect of clinic s within county r, and $n_{s(r)}$ is the number of individuals in clinic s of county r. The log RR associated with EPT was estimated to be log(0.953) with a standard error of 0.0447. The two-sided size 0.05 Wald test fails to reject the null hypothesis that EPT has no effect on the prevalence of chlamydia in this population (p-value = 0.28). Therefore, based on the model in (3.5), there is no convincing evidence that EPT reduces the prevalence of chlamydial infection. However, if there is a lag in the effect of EPT and we use the model in (3.5) ignoring the lag, EPT could appear to have no effect.

Step 1 of our two-step approach to estimating any lag in the long-term EPT effect begins

with a random effect model of chlamydia prevalence on time on intervention, time on study, and with a random intercept for counties and a random intercept for clinics nested within counties, i.e., the model,

$$\log\left(p_{rstuv}|X_{rv}\right) = \alpha + \beta_t + \theta_v X_{rv} + \nu_r + \omega_{s(r)};$$

$$r = 1, \ldots, 22;\ s = 1, \ldots, 132;\ t = 0, \ldots, 14;\ u = 1, \ldots, n_{s(r)};\ v = 1, \ldots, 14, \tag{3.6}$$

where all components are defined as in equation 3.5, except p_{rstuv} is the risk of chlamydia after v time steps on EPT, θ_v is the log RR associated with EPT v time steps after the start of EPT (i.e., EPT is introduced at $v = 1$ and $\theta_0 = 0$ for identifiability), X_{rv} is a 0/1 indicator which is 1 when county r has been using EPT for v time steps and 0 otherwise. The estimates of the effects of time on intervention and their corresponding variance-covariance matrix were retrieved for step 2.

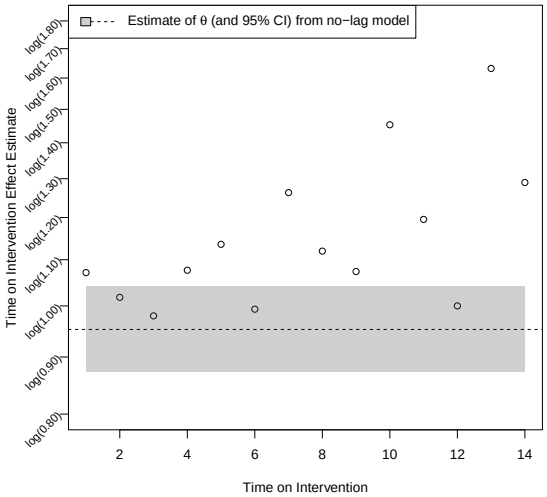

Figure 3.3 Time on EPT log(RR) estimates and the no-lag model estimate of overall log(RR) associated with EPT (with 95% confidence interval).

As can be seen in Figure 3.3, the time on intervention estimates increase with increasing time on intervention and there is little or no information about an upper bound (what would be the long-term EPT effect). Therefore, the 2nd step model did not converge.

Two things are of note regarding the results of this lagged analysis. First, the 2nd step of the two-step approach to estimating the long-term log RR does what is expected based on the time on intervention effect estimates obtained from the 1st step. There is an increasing trend in the time on intervention effect estimates with increasing time on intervention, which occurs above log(1), the assumed log RR at time on intervention = 0. These effects suggest that the intervention harms and does so increasingly over time, contradicting the results of the no-lag analysis. Second, all the time on intervention effect estimates are higher than the estimated log RR ignoring any lag, which is expected to be a weighted average of the time on intervention effect estimates and, hence, should lie somewhere in their range. This is different from what we saw in simulations, which leads us to believe that there is an assumption that was made in our model (and, therefore, in the simulations) that is violated in these data.

In simulations, estimates of time on study effects are about the same in the model ignoring a lag and the model estimating a lag. This means that, in using the two-step approach to decompose the intervention effect into its time on intervention effects, time on study effects are more or less preserved. Interestingly, this does not happen with this EPT data, where the time on study estimates change meaningfully depending on whether the no lag or lag model is fit. See Table 3.13.

Table 3.13 Time on study effect estimates from the no-lag and lagged models of chlamydial infection.

	Time on Study Effect Estimates						
	1	2	3	4	5	6	7
No-Lag Model	0.077	0.112	-0.021	-0.016	-0.093	-0.105	-0.122
Lagged Model	0.003	0.008	-0.003	-0.009	-0.018	-0.013	-0.030
	8	9	10	11	12	13	14
No-Lag Model	-0.196	-0.133	-0.246	-0.140	-0.124	-0.116	-0.006
Lagged Model	-0.032	-0.021	-0.051	-0.031	-0.020	-0.052	-0.022

To examine what would happen if we attempted to preserve the time on study effects from model (3.5) in model (3.6), we employed a method of using the estimates of time on study effects from model (3.5) as an offset in step 1 (instead of adjusting for time on study). However, such an approach does not account for the fact that the offset has been estimated when computing standard errors for the time on intervention estimates and, hence, for the long-term intervention effect and lag duration estimates. Therefore, we derived jackknifed (leave one county out) standard error estimates for the long-term intervention effect and lag duration estimates. That is, we investigated whether there was a lag in the effect of EPT as follows:

1. Let $\hat{\beta}_t,\ t = 1, \ldots, 14$ represent the 14 time on study effect estimates from the model in (3.5).

2. Fit the step 1 model as

$$\log\left(p_{rstuv}|X_{rv}\right) = \alpha + \hat{\beta}_t + \theta_v X_{rv} + \nu_r + \omega_{s(r)};$$

$$r = 1, \ldots, 22;\ s = 1, \ldots, 132;\ t = 0, \ldots, 14;\ u = 1, \ldots, n_{s(r)};\ v = 0, \ldots, 14,$$

and extract model estimates $\hat{\theta}_v;\ v = 1, \ldots, 14$ and their variance-covariance matrix, $\hat{\Sigma}$, for step 2.

3. Fit the step 2 model as

$$\hat{\theta}_v = \theta\left(1 - e^{-\frac{v}{d}}\right) + \epsilon_v; \qquad v = 1, \ldots, 14;\ (\epsilon_1, \ldots, \epsilon_{14})' \sim \text{MVN}\left(\mathbf{0}, \hat{\Sigma}\right), \tag{3.7}$$

and extract $\hat{\theta}$ and $\hat{d}$ as estimates of the long-term EPT effect and lag duration, respectively.

4. Obtain jackknifed estimates of the variance of $\hat{\theta}$ and $\hat{d}$. Repeat steps 2. and 3. 22 times leaving one county out each time. Extract $\hat{\theta}^r$ and $\hat{d}^r;\ r = 1, \ldots, 22$ as estimates of the long-term EPT effect and lag duration, respectively, with county r left out of the analysis. Compute jackknifed estimates of $\text{Var}(\hat{\theta}) = \frac{21}{22}\sum_{r=1}^{22}\left(\hat{\theta}^r - \hat{\theta}_\cdot\right)^2$ and $\text{Var}(\hat{d}) = \frac{21}{22}\sum_{r=1}^{22}\left(\hat{d}^r - \hat{d}_\cdot\right)^2$, where $\hat{\theta}_\cdot$ and $\hat{d}_\cdot$ are the average of the $\hat{\theta}^r$ and $\hat{d}^r$, respectively.

Figure 3.4 shows the result of this analysis. If we estimate a lag using the offset approach detailed above, the log(RR) is estimated to be log(0.950) with jackknifed standard error 0.0584. The lag duration is estimated to be 0.723 (about 75% of the effect of EPT on the log scale attained by the 1st time step) with jackknifed standard error 0.7262. Therefore, based on this approach which retains the time on study effects from the no-lag model, we can conclude that there is no strong evidence here of an effect of EPT since it doesn't appear that the estimate ignoring any potential lag is attenuated.

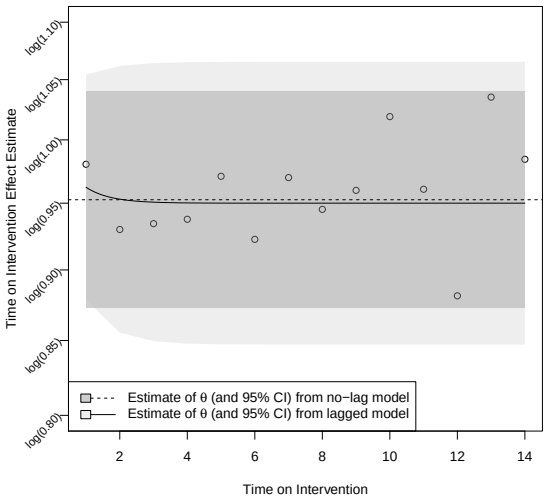

Figure 3.4 With the "offset" analysis: No-lag model estimate of overall log(RR) associated with EPT and time on intervention effect estimates and their estimated trajectory over time to the long-term log(RR) associated with EPT (with 95% confidence intervals).

3.5 Calculating Power if There is a Lag in Effect or Coverage Delay

Consider a cluster-randomized SWD trial of a prevalence outcome Y using I clusters, J time steps, and K individuals in each cluster. If an anticipated delay in the intervention effect (type (a) delay) can be modeled as in Section 3.2, we may wish to draw inference about the long-term intervention effect (θ) and test the hypothesis

$$H_0 : \theta = 0 \text{ vs. } H_a : \theta = \theta_a. \tag{3.8}$$

The power for such a test could be approximated using a simulation study conducted as follows.

1. Generate M sets of $Y_{ijlk} \sim \text{Bern}(p_{ijl}|X_{il}) = \text{Bern}(e^{\alpha+\beta_j+\theta_l X_{il}+\nu_i}|X_{il});\ i = 1, \ldots, I;\ j = 1, \ldots, J;\ l = 0, \ldots, J-1$ for the number of clusters, number of time steps, and average number of individuals to be sampled per time step in each cluster in the study. That is, generate M datasets based on the model in (3.2). It is recommended that M be at least 1000.

2. The mean function in (3.3), $\theta\left(1 - e^{-\frac{l}{d}}\right)$, may be used to define $\theta_l;\ l = 0, \ldots, J-1$, the trajectory anticipated over the course of the study. Note that there must be some idea about the lag duration, d, to be experienced during the study. Based on the results of our simulations above, it is recommended that the long-term effect of the intervention, θ_a, is attained during the study.

3. Execute Step 1 (model 3.2) on the M sets of Y_{ijlk}'s for M estimates of θ_L and Σ_L.

4. Execute Step 2 on the M estimates of θ_L and Σ_L to obtain M estimates of θ and d.

5. For each of the M fitted models in part 3, resample the fitted outcomes B times and refit model (3.2) using the B resampled outcomes. Steps 1 and 2 can then be executed on the B resampled outcomes. This step is needed to obtain an empirical distribution for each estimate of θ from part 4. The distribution can be used to determine the $\left(\frac{\alpha}{2}\right)^{\text{th}}$ and $\left(1 - \frac{\alpha}{2}\right)^{\text{th}}$ quantiles of the distribution. It is recommended that B be at least 1000 for a representative empirical distribution of $\hat{\theta}$.

6. Power (with α type I error) can be approximated by the percentage of times out of M simulations θ_a falls outside the $\left(\frac{\alpha}{2}\right)^{\text{th}}$ and $\left(1-\frac{\alpha}{2}\right)^{\text{th}}$ quantiles.

Power can be approximated for a type (b) or coverage delay in a single step. If there is a coverage delay, the intervention effect (corresponding to 100% coverage) can be estimated with a model of the form

$$\log\left(p_{ij}|X_{ij}\right) = \alpha + \beta_j + \theta X_{ij} + \nu_i; \qquad i = 1, \ldots, I;\ j = 1, \ldots, J, \tag{3.9}$$

where p_{ij} is the prevalence of outcome Y in cluster i at time j, α is the log population prevalence at time step $j = 1$ when all clusters are in the control condition, β_j is the log RR associated with time on study j ($\beta_1 = 0$ for identifiability), θ is the log RR associated with the intervention, X_{ij} is an indicator which is equal to 1 when cluster i is receiving 100% coverage of the intervention at time j and 0 when cluster i is receiving no intervention coverage at time j, and $\nu_i \sim \mathrm{N}\left(0, \tau^2\right)$ is a random effect for cluster i. We recommend (following Hussey and Hughes [1]) that the intervention indicator in (3.9) be replaced with fractions that represent the coverage of the intervention in each cluster at each time step, i.e., the percentage of the cluster in which the intervention has been fully introduced, relative to the number for which the full intervention effect is being estimated. We note that this means using a post-randomization measure of intervention coverage as the true measure of intervention introduction, which may be subject to measurement error. conditioning on this post-randomization covariate may lead to imbalanced prognostic factors between the two treatment groups, in which case estimates of the parameters do not measure causal effects

of treatment assignment.

With this type of delay, power can be approximated with a two-tailed, size α Wald test of (3.8) with

$$\Phi\left(\frac{|\theta_a|}{\sqrt{\text{Var}\left(\tilde{\theta}_a\right)}} - z_{1-\frac{\alpha}{2}}\right), \tag{3.10}$$

where Φ is the cumulative distribution function of the standard normal distribution, $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)^{\text{th}}$ quantile of the standard normal distribution, and $\text{Var}\left(\tilde{\theta}_a\right)$ can be estimated as follows:

a) Let $\mathbf{Z}$ represent the (IJ)x$(J+1)$ design matrix corresponding to the parameter vector $\eta = (\alpha, \beta_2, \ldots, \beta_J, \theta_a)$ for the model in (3.9), with fractions replacing intervention indicators. Let $\mathbf{V}$ represent the $IJxIJ$ block diagonal variance-covariance matrix for cluster mean outcomes across the J time steps. If the link function in (3.9) is the identity link, each JxJ block of $\mathbf{V}$ looks like

$$\begin{bmatrix} \frac{p(1-p)}{K}+\tau^2 & \tau^2 & \ldots & \tau^2 \\ \tau^2 & \frac{p(1-p)}{K}+\tau^2 & \tau^2 & \vdots \\ \vdots & \tau^2 & \ddots & \tau^2 \\ \tau^2 & \ldots & \tau^2 & \frac{p(1-p)}{K}+\tau^2 \end{bmatrix}$$

where p is the baseline prevalence of outcome Y, which makes $\frac{p(1-p)}{K}$ the within-cluster variance of the cluster mean outcome, and τ^2 is the between-cluster variance of outcome Y.

It is unlikely that τ^2 will be known but power can be computed for postulated values of τ^2. If the link function in (3.9) is not the identity link (where the probability modeled and the cluster effect are on the same scale), then the within-cluster variances would need to be converted to variance on the same scale as the between-cluster variances, the linear scale. This can be achieved using the Delta Method or a first-order Taylor series expansion. For example, if the log link is used in (3.9) (hence, estimating the relative risk), the within-cluster variance on the linear scale would be $\frac{1}{p^2}\frac{p(1-p)}{K} = \frac{1-p}{pK}$, and each JxJ block matrix of $\mathbf{V}$ looks like

$$\begin{bmatrix} \frac{1-p}{pK}+\tau^2 & \tau^2 & \dots & \tau^2 \\ \tau^2 & \frac{1-p}{pK}+\tau^2 & \tau^2 & \vdots \\ \vdots & \tau^2 & \ddots & \tau^2 \\ \tau^2 & \dots & \tau^2 & \frac{1-p}{pK}+\tau^2 \end{bmatrix}.$$

Note that the between-cluster variance is defined by the link function, i.e., $\text{Var}\left(g\left(p_{ijl}\right)\right) = \tau^2$ for link function g. This means that if the between-cluster variance is known (or postulated), for example, on the probability scale it would have to adjusted for the link function scale for correct definition of $\mathbf{V}$. Conversely, if the between-cluster variance had been previously estimated using a random effects [16] or a GEE [18] model using a link function other than the identity link, then it will have to be adjusted for the probability scale if the link function used in (3.9) is the identity link. Yelland, *et al.* [23] provided suggestions for converting the between-cluster variance to the probability scale for the log and logit link functions, which essentially utilize a Taylor series expansion as well.

b) Using weighted least squares, $\mathrm{Var}(\hat{\eta})$ can be estimated as $\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}$. The variance estimate for power calculation in (3.8) is then the appropriate element of $\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}$, in this case the $(J+1)^{\text{th}}$ element.

If there is no delay expected in the intervention effect or in the coverage and assuming the same number of individuals in each cluster at each time step, Hussey and Hughes [1] provides an explicit formula for the variance of the intervention effect estimate.

3.6 Discussion

In simulations, we have investigated the bias and precision of the intervention effect estimate when ignoring lags in the intervention effect and when estimating the lags using a two-step method. We found that bias of the intervention effect estimate is substantially decreased when the lag duration is estimated rather than ignored, with an exception when a very small number of individuals are sampled in each cluster at each time step. The differences in bias, comparing when the lag is ignored versus estimated, are more profound when the lag duration is longer, which is not surprising given that the true effect of the intervention after just 1 time step is closer to the long-term effect of the intervention when the lag duration is short. This, however, underscores how effective this two-step approach to estimation of the long-term intervention effect is when the lag duration is unknown, provided the data are rich enough. The results were similar regardless of the between-cluster variance, which occurs because the step 1 random effects model makes both within- and between-cluster comparisons for effect estimation.

The opposite was true for the precision of the intervention effect estimate. The standard error of the estimate of θ when the lag duration was ignored was lower than the standard error when the lag duration was estimated. This is to be expected since there are more parameters being estimated with the two-step approach. The result is that, though standard errors are larger with the two-step approach, the estimates of θ are less biased and we have seen that small increases in the number of time steps used can reduce the standard error of the estimate to a level similar to when there is no lag in effect. We also investigated the percent bias in the estimate of the lag duration and found that bias was smaller when the lag duration was smaller. The estimates of the lag durations were based on the mean of the estimates over all simulations in each scenario, which we noted had skewed distributions particularly when the number of individuals sampled was small.

We also explored the effect of key cluster-randomized SWD characteristics (the number of clusters, the number of time steps, and the number of individuals per cluster per time step) on the bias and precision of the intervention effect estimate and the bias of lag duration estimates. While MLEs are asymptotically unbiased, we saw finite sample bias of the estimates in these simulations, which decreased with increasing I, J, and K. When the lag in the intervention effect was ignored, bias of the intervention effect was high and seemed unaffected by I, J, and K. If the lag duration was estimated, however, bias of the intervention effect and lag duration estimates decreased with increasing number of clusters, number of time steps, and number of individuals per cluster per time step. The greatest impact on bias of the intervention effect estimate came with small increases in the number of time steps

used in the design (Table 3.2). For the short lag duration ($d = 0.5$), the percent bias of the long-term intervention effect is 4.44% at the lowest J of 5. This means that a good estimate of the intervention effect (if we use 5% bias as a cutoff) is achieved in just 5 time steps in this scenario. The longest time on intervention that is observed is J - 1 = 4 by which time $1 - e^{-\frac{4}{0.5}} > 99\%$ of the intervention effect is achieved. For the long lag duration ($d = 1.4$), the percent bias in estimating θ is 26.9% with 5 time steps with $1 - e^{-\frac{4}{1.4}} = 94.3\%$ of the full intervention effect achieved by the longest time on intervention. Increasing J by just two time steps reduces the percent bias to 5.02% and $1 - e^{-\frac{6}{1.4}} = 98.6\%$ of the full intervention effect is achieved by the longest time on intervention. When $J = 9$, the percent bias is 0.17% and $1 - e^{-\frac{8}{1.4}} = 99.7\%$ of the full intervention effect is achieved by the longest time on intervention. This was consistent across all values of τ^2 and suggests that at least approximately 99% of the full intervention effect has to be attained by the end of the study to be able get good estimates of the long-term intervention effect. Note that this is contingent on there being enough clusters (in this case ≥ 24) and enough individuals sampled (in this case 100).

The percent bias of the long-term intervention effect estimate fell to within 2% (1% in some cases) at the highest levels of I and J, and within 6% at the highest level of K. The percent bias of the lag duration estimate fell to within 7% at the highest levels of I and J and within 14% at the highest level of K. We note that, if the estimate of the lag duration is determined by the median of the estimates over all simulations in each scenario (as opposed to the mean since the distributions are skewed), percent bias of the lag duration estimate fell to within 4% at the highest levels of I and J and within 9% at the highest level of K.

Simulation results will vary by effect size and baseline prevalence of the outcome. However, Table 3.4 suggests that, in anticipation of a short lag (about 86% of the long-term effect of the intervention attained by the end of the 1st time step), the lag can be estimated with about the same bias using at least 12 clusters, i.e., there is no appreciable reduction in percent bias of the effect estimate by increasing I. The long-term effect of the intervention is attained relatively soon after the 1st time step, however, a good (within 5% bias) estimate of the lag duration is still not seen until the study is conducted with 9 time steps (Table 3.5). The number of individuals used in each cluster at each time step also needs to be as large as possible for good estimates of the lag duration. Table 3.6 showed us that the percent bias was about 180% when K was 20 and about 14% when K was 100. These results are for when $\tau^2 = 0$ (though similar for $\tau^2 > 0$), therefore the effective sample size is $I = 24$ times K, which emphasizes the importance of a large K in estimating lag durations. The interest might only lie in getting good estimates of the long-term intervention effect, which seems to occur just from estimating the lag duration.

If the anticipated lag duration is large (about 51% of the long-term effect of the intervention reached by the end of the 1st time step), then a very large number of clusters, number of time steps, and number of individuals per cluster per time step are needed for good estimates of the lag duration. All three adjustments do not have to be made simultaneously. Perhaps the easiest and most pertinent design change to make is in the number of time steps used in the trial. Estimation of a long lag requires a long follow-up to track the trajectory of the time-on-intervention effects. Fixing K at 50 and J at 5, the percent bias in estimating the

long lag duration with 24 clusters is 50.1%. Doubling K to 100 decreases the percent bias to 21.4% but increasing J by just 2 time steps (and keeping K at 50) reduces the percent bias to 15.9%. Therefore, increasing the number of time steps can yield better estimation of the lag duration with fewer individuals per cluster per time step. Since an increase in the number of time steps is accompanied by an increase in the number of outcome measurements taken (which may be financially prohibitive), a change in the design of the trial could be made to allow for greater impact of the intervention before the 1^{st} time step, i.e., an expansion of the time between time steps.

We tested the two-step approach to estimating lagged effects on the data from the latest Washington State EPT Trial. We saw results of the estimation of the long-term RR associated EPT that were inconsistent with an effect estimate assuming no lag in the effect of EPT and with suggestions from previous studies. There is no precedent for a suggestion that EPT increases the risk of chlamydial infection nor that this gets worse over time. The results are also inconsistent with what we see in simulations. Time on study effect estimates from the no-lag model are about the same as the time on study effect estimates from the lagged model. Additionally, the estimate of the effect of the intervention from the no-lag model is a weighted average of the time on intervention effect estimates from the lagged model. Therefore, we believe there is something different about this EPT data that is not present in simulated datasets, particularly, a difference relating to the effect of time on study that we have not been able to replicate. Recognizing that the time on study effect estimates from the no-lag model are not preserved in the lagged model as they are in simulations, we

used an offset approach to force the preservation of time on study effect estimates. This method is not recommended as a fix; we merely used it as a tool to correct what we thought might be the underlying difference between the EPT data and what we saw in simulations. There may be other forces at work here, including a differential effect of EPT by county. We recommend that care is taken to represent the right model for the data in the 1st step of this approach.

We presented an approach to calculating power when a type (a) or a type (b) lag is expected. To estimate power for a type (a) lag one must define both the long-term effectiveness (θ) and the percentage of effectiveness of the intervention after 1 time step $\left(1 - e^{-\frac{1}{d}}\right)$. To estimate power for a type (b) lag one must define the percent of coverage expected at each time step. The precision of the intervention effect estimate will decrease with increasing lag durations and power will decrease correspondingly.

The SWD can be adjusted for more efficient estimation of the intervention effect if a lag in the effect is expected. We showed how different SWD characteristics affect intervention effect estimation. However, an interesting future direction regarding design adjustment would be to investigate the effect of these characteristics while fixing the overall cost of the study. If the design can't be modified, estimation of lagged effects in the SWD for better estimation of intervention effects is feasible with the two-step approach presented in this paper. Additionally, there are straightforward ways of approximating power when a lag in the intervention effect or a coverage delay is expected. We limited our investigation into estimating lagged effects to a specific lag function (3.3). This function models the lag in the

intervention effect as a percentage of the long-term intervention effect attained at each time steps. In the future, we could consider lag functions that consider the mechanism by which the intervention acts and specifically model the way an intervention works over time for particular indications, which would mean collaborations with experts in pharmacokinetics and pharmacodynamics.

4 — INTERVAL CENSORING IN THE SWD OF CLUSTER RANDOMIZED CLINICAL TRIALS IN THE PRESENCE OF FRAILTY

4.1 Introduction

The stepped wedge design (SWD) has been used to evaluate, in a community setting, interventions that have shown some potential for benefit in an individual setting [15]. The design has previously been defined and critiqued [1, 2]; briefly, in a SWD trial the intervention or treatment is introduced to each unit (person, clinic, community) at a random time. Each unit spends time in the control state but unidirectionally crosses over into the intervention state creating a stepped wedge schematic as seen in Figure 4.1. Therefore, all units eventually receive the intervention but the order and timing of the introduction of the intervention is randomized. Outcome measures are collected from each unit during each time step allowing for construction of both within- and between-unit estimates of intervention effects. The design is also used in studies, where the timing of introduction of the intervention is not randomized.

		Time				
		1	2	3	4	5
Randomization Unit	1	0	1	1	1	1
	2	0	0	1	1	1
	3	0	0	0	1	1
	4	0	0	0	0	1

Figure 4.1 Schematic of the Stepped Wedge Design indicating treatment (1) and control (0) assignments.

The SWD has been applied in several studies [7, 8, 9, 10, 13, 15], and is under consideration for future studies, including phase IV HIV vaccine trials [14]. However, most of these papers have studied prevalence outcomes measured in cross-sectional surveys. Relatively little has been written about the use of the SWD when the outcome of interest is incidence [10] and no one, to our knowledge, has considered interval censored time-to-event outcomes.

Time-to-event outcomes are often analyzed using a Cox proportional hazards model [24]. When outcomes are correlated (as is typical in cluster-randomized trials) and this correlation is not accounted for in analysis, standard errors for hazard ratio (HR) estimates are invalid and, therefore, inferences are invalid. One way to account for correlated time-to-event outcomes is to employ a frailty model, which is a proportional hazards model with multiplicative random effect called the frailty, that attempts to model the heterogeneity in the outcomes between clusters. Henderson and Oman [25] investigated the effects of ignoring the correlation in time-to-event outcomes on HR estimates in simulations under a parallel design. They showed that HR estimates were attenuated towards the null and that the bias was dependent on the amount of correlation. In Section 4.2, we show analytically that the effects of ignored frailty are similar in the SWD, where treatment assignment is time-dependent.

Interval-censored outcomes are very common in health studies with periodic follow-up/testing for an asymptomatic condition. For instance, imagine a study in which an intervention is being introduced while, simultaneously, rolling or open cohorts (where participants in a study enter and leave the study cohort at any time during the study) are formed to track outcome events. Since enrollment in the open cohort may not be synchronized with the time steps of the intervention introduction, the observation intervals for individuals in the cohort may overlap two or more time steps (i.e., interval censoring). Alternatively, imagine a scenario where a cohort is formed at the beginning of the study and the incident disease outcome is measured at evenly spaced times, say, bimonthly. The intervention is also introduced bimonthly but is delayed at some point during the study. When this happens, the outcomes are interval censored and results in ambiguous assignment of outcome to those time steps, which is particularly problematic when clusters are crossing over from control to intervention within those time steps. This means that there is incomplete information about the outcome time for some observations, which affects the efficiency with which we can estimate intervention effects.

To estimate the intervention effect when there is interval censoring, we propose an expectation-maximization (EM) algorithm [26] using a Cox proportional hazards model [24] to analyze data from a SWD study measuring a time-to-event outcome in discrete time. Though the EM Algorithm will get closer to the maximum likelihood estimate (MLE) with each iteration, it can be slow to converge. Maximization of the observed likelihood, however, is fast and gets closer to the MLE when starting values are close to the maxima. Therefore,

we propose a subsequent maximization of the observed data likelihood using the estimates from the EM Algorithm as starting values to speed up the maximization process. In Section 4.3, we present a model for the hazard function, detail the derivation of the full likelihood, and outline an algorithm for parameter estimation in the case when there is no interval censoring. We then extend the algorithm for parameter estimation in the presence of interval censoring in Section 4.4. In Section 4.5, we report the performance of the estimator of the hazard ratio in simulations. In particular, we evaluate the finite sample bias and standard error of the estimator as the number of clusters used in the trial increases. Section 4.6 outlines a proposal for computing the power in a SWD time-to-event trial and explores the effect of interval censoring on the standard error of the intervention effect estimate in simulations. We conclude with a brief discussion and plans for future work in Section 4.7.

4.2 Ignored frailty in the SWD

If outcomes depend on a cluster effect (frailty) but a marginal model (standard Cox proportional hazards model unconditional on cluster effects) is fit to the data, the effects on the estimated HR can be seen analytically by examining the marginal hazard function derived from the corresponding cluster-specific hazard function. Consider a cluster-randomized, SWD trial measuring a failure time or time-to-event outcome in I clusters. A frailty model of the hazard of the outcome in continuous time can be represented as

$$\lambda_i\left(t|X_{it},\nu_i\right)=\lambda_0(t)\nu_i e^{\beta X_{it}};\qquad \nu_i\sim\Gamma\left(\frac{1}{\tau^2},\tau^2\right);\quad i=1,\ldots,I, \tag{4.1}$$

where $\lambda_i\left(t|X_{it},\nu_i\right)$ is the cluster-specific hazard of the outcome at time t, $\lambda_0(t)$ is the baseline hazard at time t, β is the population intervention effect (the log hazard ratio associated with intervention), X_{it} indicates intervention condition in cluster i at time t, and ν_i is the random effect of cluster i. The Gamma distribution is commonly used for the random effect and is conveniently chosen so that integrating over the random effect is mathematically straightforward. Each random effect has mean 1 and variance τ^2, where a small value of τ^2 means there's a small effect of cluster on outcomes, i.e., hazards are not very different across clusters. A value of ν_i less than 1 means the hazard of outcomes is relatively lower in cluster i, while a value greater than 1 means the hazard is increased.

What happens to the estimated HR if we fit the standard Cox proportional hazards model,

$$\lambda_i^{'}\left(t|X_{it},\nu_i\right)=\lambda_0^{'}(t)e^{\beta^{'}X_{it}};\qquad i=1,\ldots,I, \tag{4.2}$$

instead of fitting the appropriate model in (4.1)? If there is no cluster effect in (4.1), then $\beta=\beta^{'}$ and $\lambda_0(t)=\lambda_0^{'}(t)$. The marginal hazard function derived from (4.1) is

$$\lambda\left(t|X_{it}\right)=\frac{\lambda_0(t)e^{\beta X_{it}}}{1+\tau^2\int_0^t\lambda_0(s)e^{\beta X_{is}}ds}.$$

See the Appendix A4.1 for the derivation. This marginal hazard function reflects the unconditional hazard in the data and attains the hazard modeled in (4.2) only when there truly is no effect of cluster on hazard, i.e., when $\tau^2=0$. Otherwise, the hazard model in (4.2) is misspecified, the degree to which depends on the degree of across cluster variability. That is, the more clusters vary (bigger τ^2), the more (4.2) overestimates the true hazard.

Additionally, for any value of $\tau^2 > 0$, (4.2) increasingly overestimates the true hazard as t increases since the integral in the true marginal hazard function adds non-negative terms with increasing t.

In each cluster i in the SWD, if $X_{it} = 0$, then $X_{is} = 0$ for $s < t$. Therefore, the true marginal hazard when $X_{it} = 0$ is

$$\lambda\left(t|X_{it}=0\right) = \frac{\lambda_0(t)}{1+\tau^2\int_0^t \lambda_0(s)ds}.$$

In any cluster i, if $j_i \leq t$ is the first time step in the intervention condition in cluster i and if $X_{it} = 1$, then $X_{is} = 0$ for $s < j_i$ and $X_{is} = 1$ for $s \geq j_i$. We can define the marginal hazard function based on this j_i conditional on $X_{it} = 1$ as

$$\lambda\left(t|X_{it}=1\right) = \frac{\lambda_0(t)e^{\beta}}{1+\tau^2\int_0^t \lambda_0(s)e^{\beta\mathbb{I}(s\geq j_i)}ds},$$

where $\mathbb{I}$ is an indicator function. If the intervention is rolled out as planned in the SWD, then X_{it} will never take on both 0 and 1 in the same cluster at the same time step. Therefore, hazards at t conditional on treatment condition can only be compared across clusters. At time t in clusters $i^{'}$ and i, the true HR comparing treatment to control is therefore

$$\frac{\lambda\left(t|X_{i't}=1\right)}{\lambda\left(t|X_{it}=0\right)} = \frac{\lambda_0(t)e^{\beta}}{1+\tau^2\int_0^t \lambda_0(s)e^{\beta\mathbb{I}(s\geq j_i)}ds}\frac{1+\tau^2\int_0^t \lambda_0(s)ds}{\lambda_0(t)}$$

$$= e^{\beta}\frac{1+\tau^2\int_0^t \lambda_0(s)ds}{1+\tau^2\int_0^t \lambda_0(s)e^{\beta\mathbb{I}(s\geq j_i)}ds}. \tag{4.3}$$

This HR is e^{β}, the HR expected from (4.2), only when $\tau^2 = 0$. Therefore, again, unless

there truly is no variation in outcomes across clusters, the model in (4.2) is misspecified. When $\tau^2 > 0$, if the intervention decreases the hazard of the event, then the model in (4.2) attenuates the HR towards 1.0 since the integral in the denominator of (4.3) is smaller than the integral in the numerator. The opposite happens if the intervention increases hazards or harms. Additionally, by L'Hôpital's Rule and the Fundamental Theorem of Calculus, the limit of (4.3) as t increases is

$$e^{\beta}\frac{1+\tau^2\lambda_0(t)}{1+\tau^2\lambda_0(t)e^{\beta\mathbb{I}(t\geq j_i)}} = \frac{e^{\beta}}{e^{\beta\mathbb{I}(t\geq j_i)}} = \frac{e^{\beta}}{e^{\beta}} = 1.$$

That is, the true HR approaches the null as t increases, violating the proportional hazards assumption made in (4.2).

Given these observations, a frailty model (or any model accounting for correlated outcomes) is recommended for failure time data from SWD cluster-randomized trials. The EM Algorithm [26] can be used on the full likelihood corresponding to the hazard model in (4.1) where the only latent variable would be the random effect in the absence of interval censoring. In Section 4.3 we outline such an approach and extend the approach to account for the presence of interval censoring in Section 4.4.

4.3 Frailty Model Without Interval Censoring in the SWD

Consider first a SWD trial with time-to event outcome and complete failure time information, i.e., there is no interval censoring. The data are discrete failure times, T, on the K_i individuals in cluster i where $i = 1, \ldots, I$. There are J time steps so each T ranges from 1 to J. Outcomes

measured in a cluster will likely be correlated and, if this correlation is not accounted for in the analysis, inference drawn from such an analysis can be invalid. One way to account for correlated outcomes is to employ a frailty model. In Section 4.2 we saw that, if the true model is a frailty model and the frailty is ignored, HR estimates will be biased and will approach the null as t increases. We will use a frailty model or a hazard model conditional on random effects [16] to attempt to account for the correlation of failure times in a cluster but note that a generalized estimating equations (GEE) [18] model may also be used. A discrete time Cox proportional hazards frailty model can be defined as

$$\lambda_i\left(j|X_{ij},\nu_i\right)=\lambda_0(j)\nu_i e^{\beta X_{ij}};\qquad \nu_i\sim\Gamma\left(\frac{1}{\tau^2},\tau^2\right);\quad i=1,\ldots,I;\quad j=1,\ldots,J, \tag{4.4}$$

where $\lambda_i\left(j|X_{ij},\nu_i\right)$ is the cluster-specific hazard of the outcome at time j, $\lambda_0(j)$ is the baseline hazard at time j, β is the population intervention effect (the log hazard ratio associated with intervention), X_{ij} indicates intervention condition in cluster i at time j, and ν_i is the random effect associated with cluster i.

In discrete time, $\lambda_i\left(j|X_{ij},\nu_i\right)=\mathrm{P}\left(T=j|T\geq j,X_{i1},\ldots,X_{ij},\nu_i\right)$, therefore the contribution to the conditional likelihood of an individual in cluster i who fails at j is

$$\begin{aligned}\mathrm{P}\left(T=j|X_{i1},\ldots,X_{ij},\nu_i\right)&=\mathrm{P}\left(T=j|T\geq j,X_{i1},\ldots,X_{ij},\nu_i\right)\mathrm{P}\left(T\geq j|X_{i1},\ldots,X_{ij},\nu_i\right)\\&=\lambda_i\left(j|X_{ij},\nu_i\right)\prod_{s=1}^{j-1}\left[1-\lambda_i\left(s|X_{is},\nu_i\right)\right]\\&\approx\lambda_i\left(j|X_{ij},\nu_i\right)e^{-\sum_{s=1}^{j-1}\left(\lambda_i(s|X_{is},\nu_i)+\frac{\lambda_i(s|X_{is},\nu_i)^2}{2}\right)}\end{aligned}$$

where the approximation holds for small hazards. See Appendix A4.2 for details of the

distribution function and the approximation. Individuals who do not experience their event during the study period contribute

$$\mathrm{P}\left(T > J | X_{i1}, \ldots, X_{ij}, \nu_i\right) = \prod_{s=1}^{J} \left[1 - \lambda_i\left(s | X_{is}, \nu_i\right)\right] \approx e^{-\sum_{s=1}^{J}\left(\lambda_i(s|X_{is},\nu_i) + \frac{\lambda_i(s|X_{is},\nu_i)^2}{2}\right)}.$$

Similarly, individuals who drop out at any time, j, during the study (whose failure times are assumed missing completely at random based on the non-informative censoring assumption) contribute

$$\mathrm{P}\left(T > j | X_{i1}, \ldots, X_{ij}, \nu_i\right) \approx e^{-\sum_{s=1}^{j}\left(\lambda_i(s|X_{is},\nu_i) + \frac{\lambda_i(s|X_{is},\nu_i)^2}{2}\right)}.$$

Let $Y_{ijk} = 1$ if the k^{th} individual in cluster i fails at j and 0 otherwise. Let $\mathcal{R}_{ij}$ be the number of individuals at risk for an event in cluster i at time j, C_{ij} be the number of individuals who drop out of the study in cluster i at time j, and let K_i be the number of individuals enrolled in the study in cluster i. Then,

i. $\mathcal{R}_{ij} = K_i - \sum_{s=1}^{j-1} D_{is} - \sum_{s=1}^{j-1} C_{is}$,

ii. $D_{ij} = \sum_{\mathcal{R}_{ij}} Y_{ijk}$ is the number of individuals in cluster i who had their event at j, and

iii. $\bar{D}_i = K_i - \sum_{j=1}^{J} D_{ij} - \sum_{j=1}^{J} C_{ij}$ is the number of individuals in cluster i who didn't have their event in the study period.

Additionally, let $\Lambda_{ij} = \sum_{s=1}^{j} \lambda_0(s) e^{\theta X_{is}}$ and $\Lambda'_{ij} = \sum_{s=1}^{j} \frac{\lambda_0^2(s) e^{2\theta X_{is}}}{2}$. Assuming event times within cluster i are independent conditional on ν_i, the contribution of cluster i to the conditional likelihood is

$$L\left(\theta,\lambda_0(\mathbf{j})|\nu_i\right)=\left(\prod_{j=1}^{J}\left[\lambda_i\left(j|X_{ij},\nu_i\right)e^{\left(\nu_i\Lambda_{ij-1}+\nu_i^2\Lambda'_{ij-1}\right)}\right]^{D_{ij}}\left[e^{\left(\nu_i\Lambda_{ij}+\nu_i^2\Lambda'_{ij}\right)}\right]^{C_{ij}}\right)\left[e^{\left(\nu_i\Lambda_{iJ}+\nu_i^2\Lambda'_{iJ}\right)}\right]^{\bar{D}_i}, \tag{4.5}$$

and the full likelihood is

$$L\left(\theta,\lambda_0(\mathbf{j}),\tau\right)=\prod_{i=1}^{I}\left(\prod_{j=1}^{J}\left[\lambda_i\left(j|X_{ij},\nu_i\right)e^{\left(\nu_i\Lambda_{ij-1}+\nu_i^2\Lambda'_{ij-1}\right)}\right]^{D_{ij}}\left[e^{\left(\nu_i\Lambda_{ij}+\nu_i^2\Lambda'_{ij}\right)}\right]^{C_{ij}}\right)\left[e^{\left(\nu_i\Lambda_{iJ}+\nu_i^2\Lambda'_{iJ}\right)}\right]^{\bar{D}_i}\frac{\nu_i^{\frac{1}{\tau^2}-1}e^{-\frac{\nu_i}{\tau^2}}}{(\tau^2)^{\frac{1}{\tau^2}}\Gamma(\tau^2)}. \tag{4.6}$$

The log likelihood is

$$\begin{aligned}l\left(\theta,\lambda_0(\mathbf{j}),\tau\right)=\sum_{i=1}^{I}\Bigg[\sum_{j=1}^{J}\Big\{&D_{ij}\log\left(\nu_i\right)+D_{ij}\log\left(\lambda_0\left(j\right)e^{\theta X_{ij}}\right)-D_{ij}\left(\nu_i\Lambda_{ij-1}+\nu_i^2\Lambda'_{ij-1}\right)-C_{ij}\left(\nu_i\Lambda_{ij}+\nu_i^2\Lambda'_{ij}\right)\Big\}\\&-\bar{D}_i\left(\nu_i\Lambda_{iJ}+\nu_i^2\Lambda'_{iJ}\right)+\left(\frac{1}{\tau^2}\text{-}1\right)\log\left(\nu_i\right)-\frac{\nu_i}{\tau^2}-\frac{\log\left(\tau^2\right)}{\tau^2}-\log\left(\Gamma\left(\frac{1}{\tau^2}\right)\right)\Bigg].\end{aligned} \tag{4.7}$$

The presence of the latent random effects in (4.4) precludes maximization of the partial likelihood function as an option for parameter estimation (as is done in the unconditional Cox proportional hazard model) since the random effects are unknown. In the absence of interval censoring, the parameters of (4.4) can be estimated by numerical maximization of the marginal likelihood (obtained by integrating (4.5) over ν_i) or by an EM algorithm approach. The random effects ($\nu_i;\ i=1,\ldots,I$) are not of interest in this problem, which makes τ a nuisance parameter, and the EM algorithm approach would treat the random effects as latent variables. Both approaches are fairly straightforward if the random effects

are assumed to have a Gamma distribution since integrating over the random effect yields an explicit, analytic result.

The EM Algorithm approach to estimating the model parameters with known or right-censored failure times but latent random effects would proceed as follows:

Let η denote current values of the parameters $(\theta, \lambda_0(1), \ldots, \lambda_0(J), \tau)$.

a) Provide initial estimates for $\eta = (\theta, \lambda_0(1), \ldots, \lambda_0(J), \tau)$.

b) E-Step Find the expectation of (4.7) under ν_i, conditioned on η and the observed data. That is, find

$$
\begin{aligned}
E\left[l\left(\theta, \lambda_0(\mathbf{j}), \tau\right)|\eta, \mathbf{X}\right] = \sum_{i=1}^{I} \Bigg[\sum_{j=1}^{J} \Big\{ & E\left[\log\left(\nu_i\right)|\eta, \mathbf{X}\right] D_{ij} + D_{ij} \log\left(\lambda_0\left(j\right) e^{\theta X_{ij}}\right) \\
& - D_{ij}\left(E\left(\nu_i|\eta, \mathbf{X}\right)\Lambda_{ij-1} + E\left(\nu_i^2|\eta, \mathbf{X}\right)\Lambda_{ij-1}^{'}\right) \\
& - C_{ij}\left(E\left(\nu_i|\eta, \mathbf{X}\right)\Lambda_{ij-1} + E\left(\nu_i^2|\eta, \mathbf{X}\right)\Lambda_{ij-1}^{'}\right)\Big\} \\
& - \bar{D}_i\left(E\left(\nu_i|\eta, \mathbf{X}\right)\Lambda_{iJ} + E\left(\nu_i^2|\eta, \mathbf{X}\right)\Lambda_{iJ}^{'}\right) + \left(\frac{1}{\tau^2}\text{-}1\right) E\left[\log\left(\nu_i\right)|\eta, \mathbf{X}\right] \\
& - \frac{E\left(\nu_i|\eta, \mathbf{X}\right)}{\tau^2} - \frac{\log\left(\tau^2\right)}{\tau^2} - \log\left(\Gamma\left(\frac{1}{\tau^2}\right)\right)\Bigg]. \qquad (4.8)
\end{aligned}
$$

c) M-Step Maximize (4.8) for updated estimates of the parameters.

The resulting expectation in **b)** can be maximized with respect to η for improved estimates of η. There are options for maximization algorithms but we suggest algorithms that allow constraints on the parameter estimates. This is particularly relevant for

estimating the baseline hazards, which, in discrete time, are restricted to the interval [0,1]. The gradient of (4.8) with respect to η can also be supplied to speed up the maximization procedure. We used the Broyden[27, 28]-Fletcher[29]-Goldfarb[30]-Shanno[31] (BFGS) numerical maximization method, which is an iterative method of finding the roots of functions based on Newton's methods.

d) Estimates from **c)** are used to update the distribution of $\nu_i|\eta, \mathbf{X}$, all of which are used in the next E-Step. Iterate **b)** and **c)** until convergence to final estimates of η.

The final estimates of η are MLEs of the log(HR), discrete baseline hazards, and the standard deviation of the cluster effect.

4.4 Frailty Model With Interval Censoring in the SWD

Now suppose events are interval censored. If an event for the k^{th} individual in the i^{th} cluster is censored in discrete time in the interval $[l_{ik}, r_{ik}]$, then the event is only known to have occurred at some time, j, in that interval. Thus, with interval censoring, the D_{ij}s (and hence $\bar{D}_i$s) are also unobserved latent variables. An EM algorithm approach to estimating the parameters of the model in (4.4) now needs to consider both sources of missingness, the D_{ij}s and the ν_is. Again, let η denote current values of the parameters $(\theta, \lambda_0(1), \ldots, \lambda_0(J), \tau)$. The following outlines the EM algorithm approach to obtaining MLEs of the parameters:

a) Provide initial estimates for $\eta = (\theta, \lambda_0(1), \ldots, \lambda_0(J), \tau)$.

b) E-Step Find the expectation of (4.7) under D_{ij} and ν_i jointly, conditioned on η, the observed data, and the intervals, $[l_{ik}, r_{ik}]$. In other words, find

$$E\left[l\left(\theta,\lambda_0(\mathbf{j}),\tau\right)|\eta,\mathbf{X},l,r\right]=\sum_{i=1}^{I}\Bigg[\sum_{j=1}^{J}\Big\{E\left[D_{ij}\log\left(\nu_i\right)|\eta,\mathbf{X},l,r\right]+E\left(D_{ij}|\eta,\mathbf{X},l,r\right)\log\left(\lambda_0\left(j\right)e^{\theta X_{ij}}\right)$$
$$-E\left(D_{ij}\nu_i|\eta,\mathbf{X},l,r\right)\Lambda_{ij-1}+E\left(D_{ij}\nu_i^2|\eta,\mathbf{X},l,r\right)\Lambda_{ij-1}^{'}$$
$$-E\left(C_{ij}\nu_i|\eta,\mathbf{X},l,r\right)\Lambda_{ij-1}+E\left(C_{ij}\nu_i^2|\eta,\mathbf{X},l,r\right)\Lambda_{ij-1}^{'}\Big\}$$
$$-E\left(\bar{D}_i\nu_i|\eta,\mathbf{X},l,r\right)\Lambda_{iJ}+E\left(\bar{D}_i\nu_i^2|\eta,\mathbf{X},l,r\right)\Lambda_{iJ}^{'}$$
$$+\left(\frac{1}{\tau^2}\text{-}1\right)E\left[\log\left(\nu_i\right)|\eta,\mathbf{X},l,r\right]-\frac{E\left(\nu_i|\eta,\mathbf{X},l,r\right)}{\tau^2}$$
$$-\frac{\log\left(\tau^2\right)}{\tau^2}-\log\left(\Gamma\left(\frac{1}{\tau^2}\right)\right)\Bigg].$$

c) M-Step Maximize the result in **b)** for updated estimates of the parameters.

d) Iterate **b)** and **c)** until convergence.

Since the D_{ij}s are sums of Y_{ijk}s, the distribution function needed for the E-Step is $\mathrm{P}\left(Y_{ijk},\nu_i|\eta,X_{i1},\ldots,X_{ij},l_i,r_i\right)=\mathrm{P}\left(Y_{ijk}|\nu_i,\eta,X_{i1},\ldots,X_{ij},l_i,r_i\right)\mathrm{P}\left(\nu_i|\eta,X_{i1},\ldots,X_{ij},l_i,r_i\right)$. By assumption, conditional on the random effect ν_i, individuals (outcomes and censoring intervals) are independent of each other, i.e., one individual's outcome or interval tells us nothing about another individual's outcome or interval and $Y_{ijk}|\nu_i,\eta,X_{i1},\ldots,X_{ij},l_i,r_i\sim Y_{ijk}|\nu_i,\eta,X_{i1},\ldots,X_{ij},l_{ik},r_{ik}$. It can be shown that (see Appendix A4.3)

$$\mathrm{P}\left(Y_{ijk}, \nu_i | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right) \approx \left[\mathbb{I}_{[l_{ik} \leq j \leq r_{ik}]} \frac{e^{-\nu_i \Lambda_{ij-1} - \nu_i^2 \Lambda'_{ij-1}} - e^{-\nu_i \Lambda_{ij} - \nu_i^2 \Lambda'_{ij}}}{e^{-\nu_i \Lambda_{il_{ik}-1} - \nu_i^2 \Lambda'_{il_{ik}-1}} - e^{-\nu_i \Lambda_{ir_{ik}} - \nu_i^2 \Lambda'_{ir_{ik}}}} \right]^{Y_{ijk}}$$

$$* \left[1 - \mathbb{I}_{[l_{ik} \leq j \leq r_{ik}]} \frac{e^{-\nu_i \Lambda_{ij-1} - \nu_i^2 \Lambda'_{ij-1}} - e^{-\nu_i \Lambda_{ij} - \nu_i^2 \Lambda'_{ij}}}{e^{-\nu_i \Lambda_{il_{ik}-1} - \nu_i^2 \Lambda'_{il_{ik}-1}} - e^{-\nu_i \Lambda_{ir_{ik}} - \nu_i^2 \Lambda'_{ir_{ik}}}} \right]^{1-Y_{ijk}}$$

$$* \prod_{m=1}^{K_i} \left(e^{-\nu_i \Lambda_{il_{im}-1} - \nu_i^2 \Lambda'_{il_{im}-1}} - e^{-\nu_i \Lambda_{ir_{im}} - \nu_i^2 \Lambda'_{ir_{im}}} \right)$$

$$* \frac{\nu_i^{\frac{1}{\tau^2}-1} e^{-\frac{\nu_i}{\tau^2}}}{\mathrm{P}\left(l_i, r_i | \eta, X_{i1}, \ldots, X_{ir_i}\right) (\tau^2)^{\frac{1}{\tau^2}} \Gamma(\tau^2)},$$

where

$$\mathrm{P}\left(l_i, r_i | \eta, X_{i1}, \ldots, X_{ir_i}\right) = \int_{\nu_i=0}^{\infty} \mathrm{P}\left(l_i, r_i | \nu_i, \eta, X_{i1}, \ldots, X_{ir_i}\right) \mathrm{P}\left(\nu_i | \eta, X_{i1}, \ldots, X_{ir_i}\right)$$

$$\approx \int_{\nu_i=0}^{\infty} \prod_{m=1}^{K_i} \left(e^{-\nu_i \Lambda_{il_{im}-1} - \nu_i^2 \Lambda'_{il_{im}-1}} - e^{-\nu_i \Lambda_{ir_{im}} - \nu_i^2 \Lambda'_{ir_{im}}} \right) \frac{\nu_i^{\frac{1}{\tau^2}-1} e^{-\frac{\nu_i}{\tau^2}}}{(\tau^2)^{\frac{1}{\tau^2}} \Gamma(\tau^2)} \tag{4.9}$$

and the approximation holds for small hazards. To make the integral above analytically tractable would require re-expressing the product as an expanded sum. However, the product term in the above distribution functions evaluates to a sum of 2^{K_i} terms, which is difficult to compute for even moderate cluster sizes K_i. Therefore, the E-Step will be evaluated numerically. The expected values to be evaluated in the E-step are detailed in Appendix A4.4.

Since the EM Algorithm may be slow to converge, we may take the estimates from d) that

are obtained after a few iterations as starting values to maximize the observed data likelihood or the probability of the observed intervals given the parameters and the intervention status

$$\prod_{i=1}^{I} \mathrm{P}\left(l_i, r_i | \eta, X_{i1}, \ldots, X_{ir_i}\right). \tag{4.10}$$

Maximization of the observed data likelihood may not converge with arbitrary starting values but, if the estimates from the EM Algorithm are used as starting values, the process should quickly yield the MLEs.

The number of independent units of data in a cluster-randomized design is the number of clusters, therefore we will be evaluating the estimator of the log(HR) in this approach for finite sample bias and precision for varying number of clusters (I). The number of discrete time steps (J) and the number of individuals within each cluster (K_i) will be kept constant. All computations were performed in `R` version 2.15.1.

4.5 Simulations of SWD Trials with Interval Censoring

We simulated data from 1000 SWD trials measuring time to an event in each scenario when $I = 8$, 12, 16, and 24; $J = 5$; $K_i = 100$; and for coefficient of variation (CV - a measure of the variation between clusters defined as the ratio of the between-cluster standard deviation over the expected hazard) ranging from 0 to 0.5 (no variation in hazards to high variation in hazards between clusters). We can define the variance of the random effect (τ^2) by the CV. Assuming CV is the same in the treated and untreated groups, on the HR scale of the model in (4.1), CV $= \frac{\sqrt{\mathrm{Var}(\lambda_i(j|X_{ij}=0,\nu_i))}}{E(\lambda_i(j|X_{ij}=0,\nu_i))} = \frac{\sqrt{\lambda_0^2(j)\mathrm{Var}(\nu_i)}}{\lambda_0(j)E(\nu_i)} = \frac{\sqrt{\lambda_0^2(j)\tau^2}}{\lambda_0(j)} = \tau$. Therefore we generated

between-cluster variance τ^2 by squaring chosen CVs. We note that when CV $= \tau$ is truly 0, the true likelihood function is not conditional on ν_i and, hence, is not a function of τ. Additionally, the likelihood function in (4.6) is undefined when $\tau = 0$ since the variance of the random effect, τ^2, is required to be greater than 0. Therefore, when CV is specified to be 0, we expect that the log likelihood function in (4.7) will not be as efficient at estimating the intervention effect and the baseline hazards as an unconditional log likelihood function would be and it will tend to overestimate τ^2 in this situation (since τ^2 will never be estimated to be zero under (4.7)). In this case, we are evaluating the performance of the approach when there is no correlation of outcomes within clusters. When CV > 0, we expect that τ^2 will be underestimated, as MLEs of variance components are known to be underestimated because they do not account for the degrees of freedom lost in estimating fixed effects.

It was assumed that the intervention halved the hazard ratio, i.e., $\log(\text{HR}) = \log(0.5)$ with baseline hazards at time steps 1 through 5 of 0.055, 0.050, 0.045, 0.040, and 0.035, respectively. All individuals were observed for an event at time step 1 and then again at randomly determined time steps thereafter, creating censoring intervals overlapping no more than two time steps. We have assumed for the sake of simplicity that there are no dropouts (no right censoring) during the study period, i.e., that $C_{ij} = 0$ for all i and j. We examine the finite sample bias and standard error of the estimator of the log(HR) as I increases. We also examine the estimators of the baseline hazards and between-cluster variance. Our approach to estimation of the log(HR), baseline hazards, and between-cluster variance did not yield MLEs in a few simulations. The following are results from approximately 99.6% of all simulations.

The results of about 0.4% of the simulations were invalidated because the estimated variance-covariance matrices of the estimates were not positive-definite, which we assume occurred because of the approximation of $\prod_{s=1}^{j-1}\left[1-\lambda_i\left(s|X_{is},\nu_i\right)\right]$ with $e^{-\sum_{s=1}^{j-1}\left(\lambda_i(s|X_{is},\nu_i)+\frac{\lambda_i(s|X_{is},\nu_i)^2}{2}\right)}$.

4.5.1 Finite Sample Bias of the Estimate of the log(HR)

The overall estimate of the log(HR) was computed as the average log(HR) estimate from the valid simulations at each I and each CV. This approach to the estimation of the log(HR) resulted in estimates that tend to be less biased as I increased, i.e., estimation of the intervention effect improved with increasing number of clusters, particularly at CVs less than 0.3.

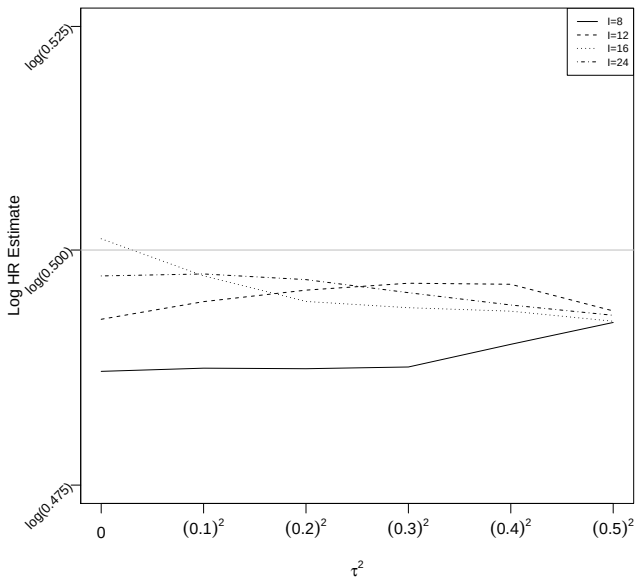

Figure 4.2 Intervention effect estimates by CV and number of clusters (I).

However, all estimates of the log(HR) are within the approximate Monte Carlo error ($\frac{1}{\sqrt{1000}}$)

of the true log (HR) of log(0.5) and each other at each CV, with the exception of the estimate when I=8 and CV=0.3. See Figure 4.2. At the smallest $I = 8$, the percent bias in the estimation of the log(HR) was at most 6.25%, which decreased to about 0.5% when $I = 24$. We note a tendency for the log(HR) to be underestimated in these simulations.

4.5.2 Coverage and Standard Error of the Estimate of the log(HR)

In each simulated dataset, the observed variance of the log(HR) estimate was computed as the appropriate element of the negated and inverted expected matrix of second derivatives of the log of the observed data likelihood (4.10), evaluated at the final estimates of η. That is, the `R` hessian() function was applied to the log of (4.10), the result of which was negated and inverted to obtain an estimate of the observed variance of the parameter estimates. The estimated observed variance for the estimate of the log(HR) was used to derive a 95% confidence interval for the log(HR). Coverage is defined as the percentage of times, over all simulations, that the true log(HR) fell within the interval

$$\left[\hat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{\hat{\text{Var}}\left(\hat{\theta}\right)},\ \hat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{\text{Var}}\left(\hat{\theta}\right)}\right]$$

where $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)^{\text{th}}$ quantile of the standard normal distribution. Ideally, for significance level $\alpha = 0.05$, $\text{P}\left(\hat{\theta} - 1.96\sqrt{\hat{\text{Var}}\left(\hat{\theta}\right)} \leq \theta \leq \hat{\theta} + 1.96\sqrt{\hat{\text{Var}}\left(\hat{\theta}\right)}\right) = 0.95$. Figure 4.3 shows the actual coverage probabilities by CV and I, which ranged from approximately 92% to just above 96%. At the lowest I of 8, coverage varies more with increasing CV than when the number of clusters is higher, suggesting better accuracy of this approach to estimating

the log(HR) when outcomes vary across clusters with increasing number of clusters.

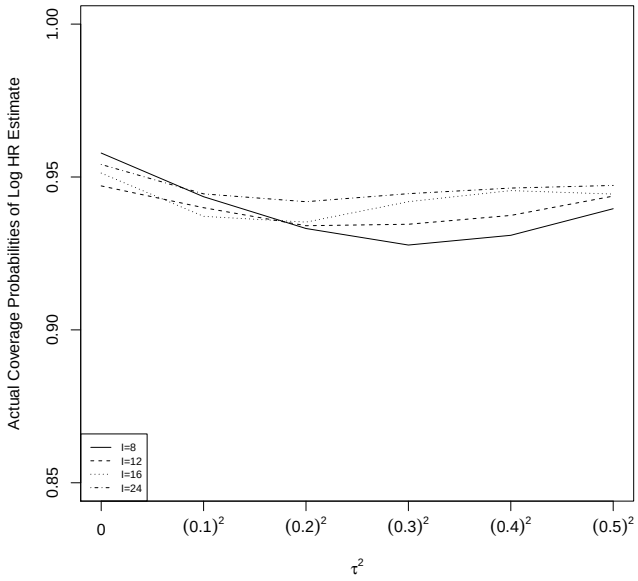

Figure 4.3 Coverage probabilities by CV and number of clusters (I).

We compared the mean of the square root of the observed variance in each simulation to the standard error of the log(HR) estimates (standard deviation of the estimates over all simulations) to evaluate the consistency of the estimate of standard error. Figure 4.4 shows that these two measures of the accuracy of the log(HR) estimate get closer as I increases regardless of CV.

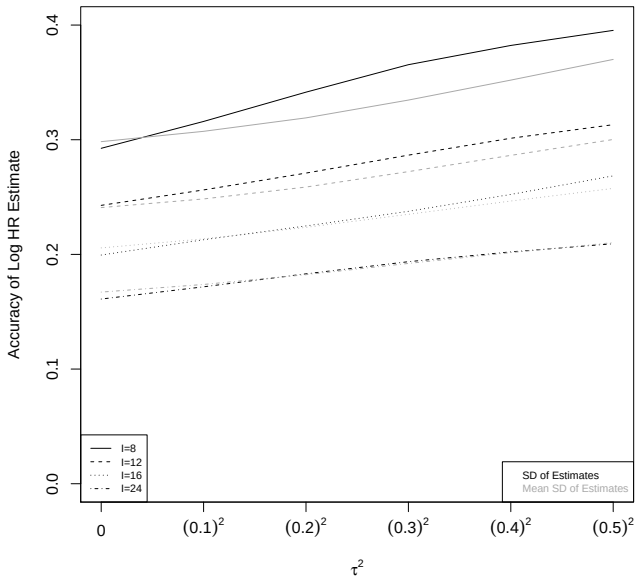

Figure 4.4 Comparison of the mean observed standard deviations with the standard deviation of the log(HR) estimates over all simulations by CV and number of clusters (I).

According to maximum likelihood estimation theory, estimates of the log(HR) from this approach have a normal distribution asymptotically. Efron and Hinkley [32] suggest that an appropriate measure of the asymptotic variance of these estimators is the observed variance computed from the observed likelihood in (4.10), the square root of which is depicted in Figure 4.4.

4.5.3 Baseline Hazards and Between-Cluster Variance

In Figure 4.5, we see that the estimates of the baseline hazards show no or very little bias at earlier time steps, are less biased with increasing I, and are only slightly influenced by CV. Though only showing a percent bias of at most 15% when $I = 8$ at the last time step, the baseline hazards estimates for later time steps are more biased than those for earlier time

steps. The bias seen in the later baseline hazards estimates can be attributed to the fact that an approximation in the computation of $\mathrm{P}\left(T = j | X_{i1}, \ldots, X_{ij}, \nu_i\right)$ was used to derive the likelihood function in (4.5), which yields a maximum likelihood estimate of $\lambda_0\left(j\right)$ greater than what we would have gotten without the approximation. The difference in the two estimates (with and without approximation) of $\lambda_0\left(j\right)$ is greater at later time steps, which, along with fewer events at later time steps, explains the greater bias seen with estimates of $\lambda_0\left(4\right)$ and $\lambda_0\left(5\right)$.

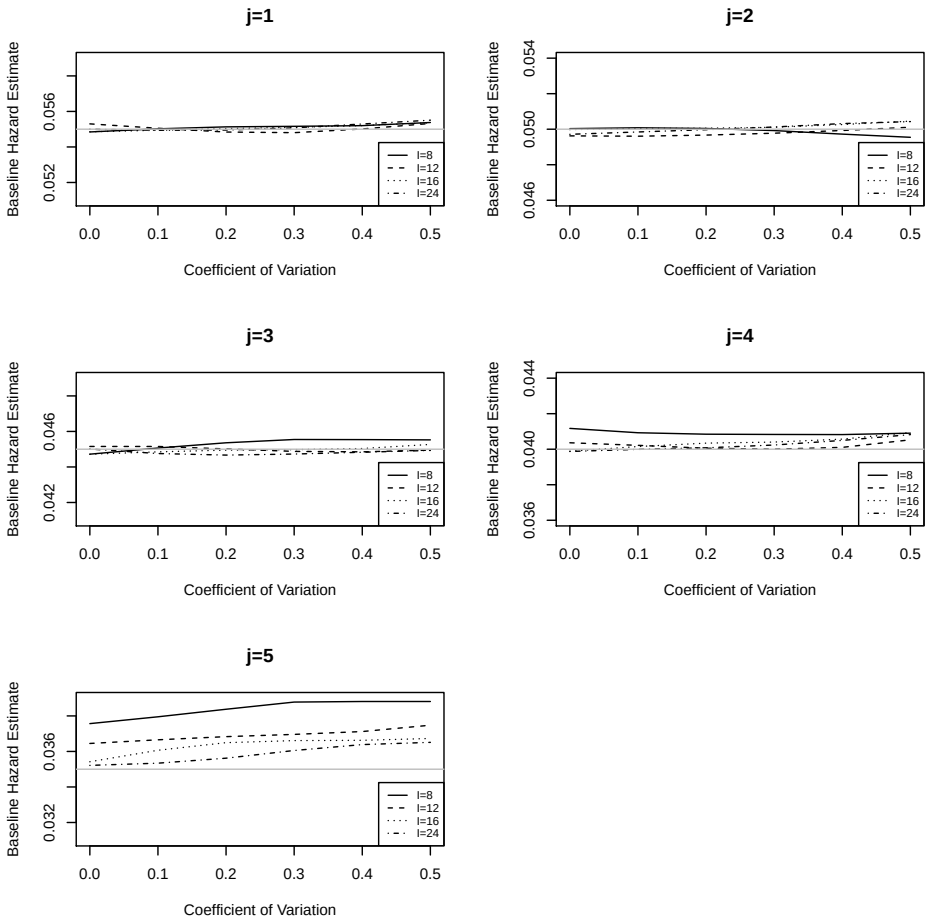

Figure 4.5 Baseline hazard estimates by CV and number of clusters (I).

The between-cluster variance is a nuisance parameter in this process. However, it is also estimated as a byproduct of this approach. Shown in Figure 4.6, the estimates of τ^2 when CV = 0 are slightly overestimated for reasons mentioned before. When CV = 0.1, τ^2 is also slightly overestimated and estimation gets slightly worse with increasing CV with a tendency for underestimation. Regardless of CV, estimates of τ^2 are less biased with increasing I with a huge improvement in going from 8 clusters to just 12 clusters. At the worst level of between-cluster variability (CV = 0.5), the bias in estimating τ^2 is only about 14% when I = 8 and decreases to about 3% when I = 24. The underestimation of τ^2 is not surprising since these are maximum likelihood estimates, which ignore the degrees of freedom lost in estimating the other parameters.

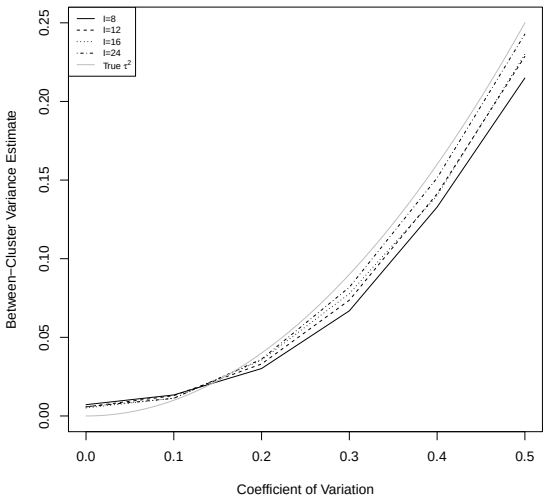

Figure 4.6 τ^2 estimates by CV and number of clusters (I).

4.6 Power and the Effect of Interval Censoring

In the absence of interval censoring, we can approximate the power to detect a particular log(HR) assuming specific baseline hazards and between-cluster variance, as outlined below. We also report how the standard error of the log(HR) estimate changes with varying levels of interval censoring as a means to illustrate the effect that interval censoring may have on power in this design.

The model for the log(HR) corresponding to (4.1) is

$$\log\left(\lambda_i\left(t|X_{it},\nu_i\right)\right) = \log\left(\lambda_0(t)\right) + \log\left(\nu_i\right) + \beta X_{it}; \quad \nu_i \sim \Gamma\left(\frac{1}{\tau^2}, \tau^2\right); \quad i = 1,\ldots,I; \quad j = 1,\ldots,J, \tag{4.11}$$

We wish to determine the power to test the hypothesis

$$H_0 : \beta = 0 \text{ vs. } H_a : \beta = \beta_a. \tag{4.12}$$

The power for this test could be approximated using a two-tailed, size α Wald test of (4.11) with

$$\Phi\left(\frac{|\beta_a|}{\sqrt{\text{Var}\left(\tilde{\beta}_a\right)}} - z_{1-\frac{\alpha}{2}}\right), \tag{4.13}$$

where Φ is the cumulative distribution function of the standard normal distribution and $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)^{\text{th}}$ quantile of the standard normal distribution. Note that the variance in the denominator of (4.12) and, hence, the power, depends not only on the true value of (β_a),

but also the specific baseline hazards $\lambda_0(j),\ j = 1, \ldots, J$ and between-cluster variance. The model in (4.10) is in the form of a generalized linear mixed model and, using weighted least squares formulation, we can approximate Var $\left(\tilde{\beta}_a\right)$ as follows:

i. Let $\mathbf{Z}$ represent the IJx(J+1) design matrix corresponding to the parameter vector $\eta = (\lambda_0(1), \ldots, \lambda_0(J), \beta_a)$ for the model in (4.10) for particular log(HR) β_a.

ii. Let $\mathbf{V}$ represent the IJxIJ block diagonal variance-covariance matrix for cluster mean log hazards across the J time steps. The diagonal elements of each JxJ block of $\mathbf{V}$ are sums of the within-cluster variance (a function of the expected hazard) and the between-cluster variance of the outcomes, both on the log scale. The off-diagonal elements are the between-cluster variance. Given the specific log(HR) and baseline hazards, the expected hazard in cluster i at time step j when there is no expected cluster effect (a regular Cox proportional hazards model as in (4.2)) is $p_j = \lambda(j|X_j) = \lambda_0(j)\, e^{\beta X_j}$. If approximating power when there is a cluster effect, the expected hazard is an integral of the hazard function in (4.1) over the cluster effect, i.e., $p_j = \int_0^\infty \lambda_0(j)\, \nu_i e^{\beta X_{ij}} d\nu_i$, which is also $\lambda_0(j)\, e^{\beta X_j}$. On the hazard scale (4.1), the within-cluster variance at the cluster level is $\frac{p_j(1-p_j)}{n_{ij}}$, where n_{ij} is the number of individuals expected to be at risk in cluster i at time j. However, we can use the Delta Method to approximate the within-cluster variance on the log scale as $\frac{1}{p_j^2}\frac{p_j(1-p_j)}{n_{ij}} = \frac{1-p_j}{p_j n_{ij}}$. The between-cluster variance will need to be similarly transformed. Under (4.1), Var $(\lambda_i(j|X_{ij}, \nu_i)) = \tau^2$ which implies that Var $[\log(\lambda_i(j|X_{ij}, \nu_i))] = \psi'\left(\frac{1}{\tau^2}\right)$, where ψ is the derivative of the log gamma function. See Appendix A4.5 for details. Each

diagonal JxJ block of $\mathbf{V}$ then looks like

$$\begin{bmatrix} \frac{1-p_1}{p_1 n_{i1}} + \psi'\left(\frac{1}{\tau^2}\right) & \psi'\left(\frac{1}{\tau^2}\right) & \dots & \psi'\left(\frac{1}{\tau^2}\right) \\ \psi'\left(\frac{1}{\tau^2}\right) & \frac{1-p_2}{p_2 n_{i2}} + \tau^2 & \psi'\left(\frac{1}{\tau^2}\right) & \vdots \\ \vdots & \psi'\left(\frac{1}{\tau^2}\right) & \ddots & \psi'\left(\frac{1}{\tau^2}\right) \\ \psi'\left(\frac{1}{\tau^2}\right) & \dots & \psi'\left(\frac{1}{\tau^2}\right) & \frac{1-p_J}{p_J n_{iJ}} + \psi'\left(\frac{1}{\tau^2}\right) \end{bmatrix}$$

with zeros elsewhere.

iii. Then $\text{Var}\left(\tilde{\beta}_a\right)$ is the appropriate element of $\text{Var}(\eta) = \left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\right)^{-1}$, i.e., the element in the $(J+1)^{\text{th}}$ row and the $(J+1)^{\text{th}}$ column and can be used to approximate power using (4.12).

Figure 4.7 shows how power varies with the number of clusters (I) and τ^2 for a log(HR) of log(0.5) in a SWD trial with $J = 5$, starting out with $K_i = 80$, and baseline hazards 0.055, 0.050, 0.045, 0.040, and 0.035. As expected, power increases with I and is only slightly affected by an increase in between-cluster variance because of the contribution of both within- and between-cluster comparisons to effect estimation. The effect of the magnitude of the between-cluster variance on power lessens as I increases.

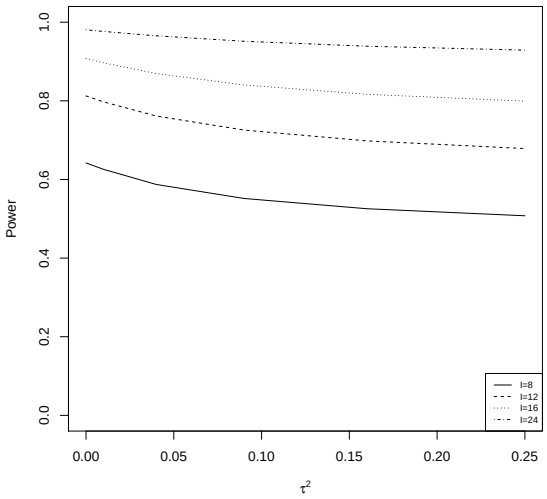

Figure 4.7 Power by number of clusters (I) and between-cluster variance (τ^2).

We simulated data from 1000 SWD time-to-event trials in each scenario with $I = 12$, $J = 9$, $K_i = 100$, $\log(\text{HR}) = \log(0.6)$, $(\lambda_0(1), \ldots, \lambda_0(9)) = (0.055, 0.053, 0.051, 0.049, 0.047, 0.045, 0.043, 0.041, 0.039)$, and $\tau^2 = 0$, 0.04, and 0.16, to evaluate the effect of varying outcome measurement interval lengths on the standard error of the log(HR) estimate with the method of estimation proposed in Section 4.4. As before, outcomes were measured at the 1^{st} time step then at random time steps thereafter creating censoring intervals overlapping 0 (failure time exactly known), 1, 2, and 3 time steps. We used the variance of the log(HR) estimate over the 1000 simulations as an estimate of variance for computing power as in (4.12). Figure 4.8 shows how power varies by between-cluster variance and by the length of the intervals.

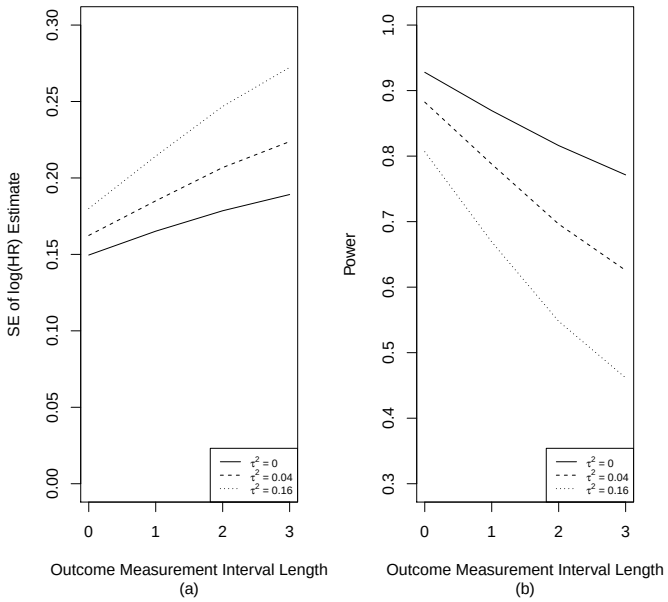

Figure 4.8 Standard error (a) and power (b) by outcome measurement interval length and between-cluster variance (τ^2). $I = 12$, $J = 9$, and $K = K_i = 100$. $(\lambda_0(1), \ldots, \lambda_0(9)) = (0.055, 0.053, 0.051, 0.049, 0.047, 0.045, 0.043, 0.041, 0.039)$.

Power decreases with increased measurement interval length and increased between-cluster variance. The decrease with increased measurement interval length is sharper at higher between-cluster variance, i.e., losses in power due to interval censoring are more pronounced when outcomes are more varied across clusters. At the highest τ^2 of 0.16, power goes from approximately 81% when there is no interval censoring to approximately 47% when outcome measurement intervals span 3 time steps. When there is no variability in outcome across clusters, power decreases from about 93% to just over 77% in the same scenario.

4.7 Discussion

We showed analytically that, similar to observations by Henderson and Oman [25], ignoring frailty in the SWD will lead to misspecified hazard models and HR estimates that are biased and that approach 1 as time increases. Additionally, if the frailty has a Gamma distribution with variance τ^2, the amount of bias is dependent on the magnitude of the variance, which is typically unknown. Consequently, we recommend frailty models (or models accounting for correlation) for the analysis of failure time data in cluster-randomized SWD trials or SWD studies with correlated outcomes.

We used a frailty model in simulations of cluster-randomized SWD trials measuring interval censored discrete failure times to estimate the log(HR) with an EM Algorithm having the cluster effect and true failure times as joint latent variables. The performance of the estimator of the log(HR) was evaluated in terms of its finite sample bias and predicted variation as demonstrated by the actual coverage probability. We also examined the behavior of estimates of the discrete baseline hazards and the between-cluster variance, a nuisance parameter. We suggested a way of approximating power in the absence of interval censoring and illustrated the effect different lengths of observation intervals may have on power when outcomes are interval censored.

The estimator of the log(HR) was shown to have very little finite sample bias, which decreased as the number of clusters used in the trial increased. The percent bias was at most 6.25%, which we saw at the lowest number of clusters of size 8. The bias varied by CV but we saw reduced variability in bias with CV at higher I. Therefore, bias is reduced

and remains consistent regardless of between-cluster variability when a large number of clusters is used in the trial. Coverage ranged from about 92% to 96% when I=8. At a higher number of clusters (I=24), coverage ranged from about 94% to 96%. The conclusion is that this approach provides accurate estimates of the log(HR) regardless of the between-cluster variation, which get better with increasing number of clusters.

Estimators of the baseline hazards at early time steps had little or no bias regardless of the between-cluster variability for even the smallest number of clusters. At later time steps, when the approximation of $\prod_{s=1}^{j-1}\left[1-\lambda_i\left(s|X_{is},\nu_i\right)\right]$ with $e^{-\sum_{s=1}^{j-1}\left(\lambda_i(s|X_{is},\nu_i)+\frac{\lambda_i(s|X_{is},\nu_i)^2}{2}\right)}$ in $\mathrm{P}\left(T=j|X_{i1},\ldots,X_{ij},\nu_i\right)$ produces overestimates of $\lambda_0\left(j\right)$, percent bias was at most about 15% when $I=8$ and 5% when $I=24$. Consequently, the approximation, used as a mathematical convenience, might not be practical if primary interest is in the baseline hazards and/or the baseline hazards are not small, though a large number of clusters reduces the percent bias. Estimators of the between-cluster variance showed percent biases that ranged between 3% and 34%. The underestimation of the between-cluster variance expected from maximum likelihood (compared to restricted maximum likelihood) estimates was seen at higher CVs. The slight overestimation seen at CV = 0 is as a result of using a likelihood function that is conditional τ^2 on when the true likelihood is unconditional on τ^2. However, as previously mentioned, τ^2 is considered a nuisance parameter in this approach to estimating the log(HR).

We have suggested an approach to approximating power in this context in the absence of interval censoring using the Wald test. Power was seen to increase with increasing number

of clusters and decreasing between-cluster variability as expected. Power was only slightly affected by changes in between-cluster variability and the effect of CV was lower at higher number of clusters. We explored the effect of varying outcome measuring interval lengths on the variance of the log(HR) estimate and, hence, on the power to detect intervention effects. We saw that, as expected, power decreases as the level of interval censoring increases, i.e., as outcome measuring interval increased. When there truly was no correlation between outcomes within clusters, the decrease in power was not as severe as it was when there was correlation between outcomes and the severity increased with the amount of correlation. In terms of percentage increase in the standard error of the log(HR) estimate, when there was no between-cluster variability, we saw about a 11%, 21%, and 26% increase in standard error when outcome measuring intervals overlapped 1, 2, and 3 time steps, respectively, compared to when there was no interval censoring. When CV = 0.4 ($\tau^2 = 0.16$), the percentage increase was about 18%, 40%, and 50% respectively. If CV is thought to be high in planned studies, care should be taken to reduce the outcome measuring intervals as much as is feasible.

In a cluster-randomized discrete failure time setting with interval censoring, we have shown that an EM Algorithm approach to a conditional random effects hazard model can work well in some situations to estimate the intervention effect, the discrete baseline hazard function, and the variance of the cluster effect. The power to detect intervention effects when there is no interval censoring can be approximated using the methods presented in this paper but it is difficult to evaluate in the presence of interval censoring. However, losses in power when there is interval censoring may be less severe when there is no between-cluster

variability in outcomes but can be substantial when outcomes vary widely across clusters.

5 — CONCLUSION

We have outlined the characteristics that define the SWD and summarized several applications of the design in trials/studies. We saw that a unique feature of the design is that it allows for both within- and between-unit comparisons to estimate intervention effects. Both comparisons are important, as an analysis that relies only on within-unit comparisons ignores potential time trends in the outcomes and, hence, is susceptible to bias. An analysis that relies only on between-unit comparisons, while accounting for time trends, is inefficient and results in relatively much larger standard error estimates of the effect estimate. In a cluster-randomized trial, whether the outcomes are mean/prevalence or incidence outcomes, two concerns are to be addressed in the estimation of intervention effects: dependence of outcomes within clusters (which can inflate the variance of the intervention effect estimate) and a trend in outcomes over time (which can result in a biased intervention effect estimate). Generalized linear mixed models or cross-sectional models accounting for correlated outcomes have been presented as valid approaches to analyzing data from a SWD trial. Both address cluster correlation and time trends, including that trends induced by healthy survivors.

Two previously unresolved issues with the SWD were described. The first concerns the

lack of analytical methods for SWD data to address delays in the effect of the intervention or delays in the rollout of the intervention, both of which can result in a truly effective intervention appearing ineffective or an ineffective intervention appearing harmful. We proposed a two-step method of estimating the duration of any lag in the intervention effect and, hence, the true long-term intervention effect. In simulations of SWD trials measuring a prevalence outcome, we found that a large number of clusters, time steps, and individuals sampled at each time step in each cluster are required to get good estimates ($< 5\%$ bias) of the lag duration and the long-term intervention effect. The full intervention effect has to be attained in some clusters by the end of the study to be able get good estimates of the long-term intervention effect, i.e., the trial has to be run long enough. We noted that relatively small increases in the number of times steps used in the trial does more to reduce bias in estimation of the intervention effect than does large increases in either the number of clusters or the number of individuals sampled. We acknowledge that small increases in the number of time steps used in a SWD trial might result in much greater financial costs for the collection of additional data points. However, the improvement in effect estimation brings an important argument to the cost-benefit discussion. We applied this method to data obtained from the Washington State EPT trial with inconclusive results. We suspect this is due to an assumption we have made in our approach that is not demonstrated in the data, one which we were unable to identify. With respect to rollout delays, we refer to Hussey and Hughes [1] for an approach to intervention effect estimation when a coverage delay is anticipated. Reasonably straight-forward methods for computing power when anticipating

either a lag in effect or a coverage delay were presented when there is some idea about the lag duration or the percentage of each cluster in which the intervention will be implemented at each time step.

The second issue concerns the lack of analytical methods for estimation of effects in time-to-event SWD trials when the outcomes are interval censored. We proposed a Cox proportional hazard model where the parameters are estimated by the EM Algorithm, followed by a subsequent optimization of the observed likelihood to ensure achievement of the global maximum. What is unique about this EM Algorithm approach is the presence of joint latent effects of the actual event times and the random cluster effect. This approach worked vey well for estimating the log(HR) in simulations when the number of clusters were small (8) regardless of the level of correlation that exists between events in a cluster. A method for computing power when there is no interval censoring was proposed. While it is difficult to anticipate the kind of interval censoring that will be experienced in a trial and, hence, estimate power for such a SWD trial, we were able to demonstrate the effect of increasing lengths of censoring intervals on the standard error of the intervention effect estimate and power in simulations.

The stepped wedge design (SWD) of cluster-randomized trials is increasingly being utilized in public health studies due to its ability to to efficiently evaluate the rollout of interventions in a community setting. Unanticipated delays in intervention effects or intervention rollout, which could lead to interval censoring in a time-to-event study is a reality of even the most well-planned study. We have presented methods that can address lags in intervention

effects and intervention rollout and interval censoring in the context of the SWD, which make the SWD a more viable option for future public health studies.

Bibliography

[1] Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Clinical Trials.* 2007;28:182-191.

[2] Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology.* 2006;6:54-62.

[3] Mdege ND, Man M, Taylor CA, et al. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology.* 2011;64:936-948.

[4] Kotz D, Spigt M, Arts ICW, et al. Use of the stepped wedge design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. *Journal of Clinical Epidemiology.* 2012;65:1249-1252.

[5] Mdege ND, Man M, Taylor CA, et al. There are some circumstances where the stepped-wedge cluster randomized trial is preferable to the alternative: no randomized trial at all. Response to the commentary by Kotz and colleagues. *Journal of Clinical Epidemiology.* 2012;65:1253-1254.

[6] Kotz D, Spigt M, Arts ICW, et al. Researchers should convince policy makers to perform a classic cluster randomized controlled trial instead of a stepped wedge design when an intervention is rolled out. *Journal of Clinical Epidemiology.* 2012;65:1255-1256.

[7] The Gambia Hepatitis Study Group. The Gambia Hepatitis Intervention Study. *Cancer Research.* 1987;47:5782-5787.

[8] Nielson J, Benn CS, Balé C, et al. Vitamin A supplementation during war-emergency in Guinea-Bissau 1998-1999. *Acta Tropica.* 2005;93:275-282.

[9] Grant AD, Charalambous S, et al. Effect of routine isoniazid preventive therapy on tuberculosis incidence among HIV-infected men in South Africa: a novel randomized incremental recruitment study. *JAMA.* 2005;293:2719-2725.

[10] Moulton LH, Golub JE, et al. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clinical Trials.* 2007;4:190-199.

[11] Killiam WP, Tambatamba BC, Chintu N, et al. Antiretroviral therapy in antenatal care to increase treatment initiation in HIV-infected pregnant women: a stepped-wedge evaluation. *AIDS.* 2010;24:85-91.

[12] Mhurchu CN, Gorton D, Turley M, et al. Effects of a free school breakfast programme on children's attendance, academic achievement and short-term hunger: results from a stepped-wedge cluster randomized controlled trial. *Journal of Epidemiology & Community Health.* 2013;67:257-264.

[13] Golden MR, Whittington WLH, et al. Impact of expedited sex partner treatment on recurrent or persistent gonorrhea or chlamydial infection: a randomized controlled trial. *NEJM.* 2005;352:676-685.

[14] Scott J. Vaccine Efficacy Trials Using Stepped Wedge Design (Unpublished Dissertation) 2008.

[15] Golden MR, Hughes JP, et al. Evaluation of a population-based program of expedited partner therapy for gonorrhea and chlamydial infection. *Sexually Transmitted Diseases.* 2007;34:598-603.

[16] Laird N, Ware J. Random-effects models for longitudinal data. *Biometrics.* 1982;38:963-974.

[17] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association.* 1993;88:9-25.

[18] Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986;73:13-22.

[19] Mancl AM, DeRouen AD. A covariance estimator for GEE with improved small-sample properties. *Biometrics.* 2001;57:126-134.

[20] Coates TJ, Richter L, Caceres C. Behavioural strategies to reduce HIV transmission: how to make them work better *Lancet.* 2008;372:669-684.

[21] Manji A, Pena R, Dubrow R. Sex, condoms, gender roles, and HIV transmission knowledge among adolescents in Leon, Nicaragua: implications for HIV prevention *AIDS Care: Psychological and Socio-Medical Aspects of AIDS/HIV.* 2007;19:989-995.

[22] Hong Y, Li XM. HIV/AIDS behavioral interventions in China: a literature review and recommendation for future research *AIDS and Behavior.* 2009;13:603-613.

[23] Yelland LN, Salter AB, et al.. Adjusted intraclass correlation coefficients for binary data: methods and estimates from a cluster-randomized trial in primary care. *Clinical Trials.* 2011;8:48-58.

[24] Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological).* 1972;34:187-220.

[25] Henderson R, Oman P. Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society, Series B (Statistical Methodology).* 1999;61:367-379.

[26] Dempster AP, Laird NM, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological).* 1977;39:1-38.

[27] Broyden CG. A new double-rank minimization algorithm *Notices of the American Mathematical Society.* 1969;16:670.

[28] Broyden CG. The convergence of single-rank quasi-Newton methods *Mathematics of Computation.* 1970;24:365-382.

[29] Fletcher R. A new approach to variable metric algorithms *Computer Journal.* 1970;13:317-322.

[30] Goldfarb D. A family of variable-metric methods derived by variational means *Mathematics of Computation.* 1970;24:23-26.

[31] Shanno DF. Conditioning of quasi-Newton methods for function minimization *Mathematics of Computation.* 1970;24:647-656.

[32] Efron B, Hinkley DV. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information *Mathematics of Computation.* 1970;24:647-656.

A — APPENDICES

A2.1 Derivation of the Variance of the Cross-Sectional Estimate

Notations:

$\bar{Y}_{R_j}$ is the average outcome in the clusters averaged over all clusters in treatment at time j,

$\bar{Y}_{C_j}$ is the average outcome in the clusters averaged over all clusters in control at time j,

n_{R_j} is the number of clusters in treatment at time j, and

n_{C_j} is the number of clusters in control at time j.

$$\begin{aligned}
\text{Var}\left(\tilde{\beta}_1\right) &= \text{Var}\left(\sum_{j=2}^{T-1}\frac{w_j}{w}\tilde{\beta}_{1j}\right) = \frac{1}{w^2}\left[\sum_{j=2}^{T-1} w_j^2 \text{Var}\left(\tilde{\beta}_{1j}\right) + \sum_{j\neq k} w_j w_k \text{Cov}\left(\tilde{\beta}_{1j}, \tilde{\beta}_{1k}\right)\right] \\
&= \frac{1}{w^2}\left[\sum_{j=2}^{T-1} w_j^2 \text{Var}\left(\bar{Y}_{R_j} - \bar{Y}_{C_j}\right) + \sum_{j\neq k} w_j w_k \text{Cov}\left(\bar{Y}_{R_j} - \bar{Y}_{C_j}, \bar{Y}_{R_k} - \bar{Y}_{C_k}\right)\right] \\
&= \frac{1}{w^2}\left[\sum_{j=2}^{T-1} w_j^2 \left(\frac{\sigma^2}{N} + \tau^2\right)\left(\frac{1}{n_{R_j}} + \frac{1}{n_{C_j}}\right) + 2\sum_{j<k} w_j w_k \text{Cov}\left(\bar{Y}_{R_j} - \bar{Y}_{C_j}, \bar{Y}_{R_k} - \bar{Y}_{C_k}\right)\right] \\
&= \frac{1}{w^2}\left[\sum_{j=2}^{T-1} w_j + 2\sum_{j<k} w_j w_k \left\{\text{Cov}\left(\bar{Y}_{R_j}, \bar{Y}_{R_k}\right) - \text{Cov}\left(\bar{Y}_{R_j}, \bar{Y}_{C_k}\right) - \text{Cov}\left(\bar{Y}_{C_j}, \bar{Y}_{R_k}\right) + \text{Cov}\left(\bar{Y}_{C_j}, \bar{Y}_{C_k}\right)\right\}\right] \\
&= \frac{1}{w^2}\left[w + 2\sum_{j<k} w_j w_k \left\{\frac{n_{R_j}\tau^2}{n_{R_j} n_{R_k}} - \frac{0 * \tau^2}{n_{R_j} n_{C_k}} - \frac{\left(n_{C_j} - n_{C_k}\right)\tau^2}{n_{C_j} n_{R_k}} + \frac{n_{C_k}\tau^2}{n_{C_j} n_{C_k}}\right\}\right]
\end{aligned}$$

$$= \frac{1}{w^2}\left[w + 2\sum_{j<k} w_j w_k \tau^2 \left\{\frac{1}{n_{R_k}} - \frac{(n_{C_j} - n_{C_k})}{n_{C_j} n_{R_k}} + \frac{1}{n_{C_j}}\right\}\right]$$

$$= \frac{1}{w^2}\left[w + 2\sum_{j<k} w_j w_k \tau^2 \left\{\frac{(n_{R_k} + n_{C_k})}{n_{C_j} n_{R_k}}\right\}\right]$$

$$= \frac{1}{w^2}\left[w + 2\tau^2 \sum_{j<k} w_j w_k \left\{\frac{I}{n_{C_j} n_{R_k}}\right\}\right]$$

$$= \frac{1}{w^2}\left[w + 2\tau^2 \sum_{j<k} \left[\left(\frac{\sigma^2}{N} + \tau^2\right)\left(\frac{I}{n_{R_j} n_{C_j}}\right)\right]^{-1} \left[\left(\frac{\sigma^2}{N} + \tau^2\right)\left(\frac{I}{n_{R_k} n_{C_k}}\right)\right]^{-1} \left\{\frac{I}{n_{C_j} n_{R_k}}\right\}\right]$$

$$= \frac{1}{w^2}\left[\sum_{j=2}^{T-1} \frac{n_{R_j} n_{C_j}}{I\left(\frac{\sigma^2}{N} + \tau^2\right)} + 2\frac{\tau^2}{I\left(\frac{\sigma^2}{N} + \tau^2\right)^2} \sum_{j<k} n_{R_j} n_{C_k}\right]$$

$$= \frac{I^2\left(\frac{\sigma^2}{N} + \tau^2\right)^2}{\left(\sum_{j=2}^{T-1} n_{R_j} n_{C_j}\right)^2}\left[\sum_{j=2}^{T-1} \frac{n_{R_j} n_{C_j}}{I\left(\frac{\sigma^2}{N} + \tau^2\right)} + 2\frac{\tau^2}{I\left(\frac{\sigma^2}{N} + \tau^2\right)^2} \sum_{j<k} n_{R_j} n_{C_k}\right]$$

$$= \frac{I\left(\frac{\sigma^2}{N} + \tau^2\right)\left(\sum_{j=2}^{T-1} n_{R_j} n_{C_j}\right) + 2\tau^2 I \sum_{j<k} n_{R_j} n_{C_k}}{\left(\sum_{j=2}^{T-1} n_{R_j} n_{C_j}\right)^2}.$$

Using the notation that $n_{R_j} = \sum_i X_{ij}$ (used in Hussey and Hughes [1]), $n_{C_j} = I - \sum_i X_{ij}$ and $\text{Var}\left(\tilde{\beta}_1\right)$, for comparison with $\text{Var}\left(\hat{\beta}_1\right)$, can be written as

$$\text{Var}\left(\tilde{\beta}_1\right) = \frac{I\left(\frac{\sigma^2}{N} + \tau^2\right)\left[\sum_{j=2}^{T-1}\left(\sum_i X_{ij}\left(I - \sum_i X_{ij}\right)\right)\right] + 2\tau^2 I \sum_{j<k}\left(\sum_i X_{ij}\left(I - \sum_i X_{ik}\right)\right)}{\left[\sum_{j=2}^{T-1}\left(\sum_i X_{ij}\left(I - \sum_i X_{ij}\right)\right)\right]^2}$$

$$= \frac{I\left(\frac{\sigma^2}{N} + \tau^2\right)\left(I\sum_{ij} X_{ij} - \sum_j\left(\sum_i X_{ij}\right)^2\right) + 2\tau^2 I \sum_{j<k}\left(I\sum_i X_{ij} - \left(\sum_i X_{ij}\right)\left(\sum_i X_{ik}\right)\right)}{\left(I\sum_{ij} X_{ij} - \sum_j\left(\sum_i X_{ij}\right)^2\right)^2}.$$

A4.1 True Marginal Hazard Function When There is a Cluster Effect

The true marginal hazard function when there is a cluster effect is derived from the cluster-specific or conditional hazard function as follows:

If $\lambda_i\left(t|X_{it},\nu_i\right)=\lambda_0(t)\nu_i e^{\beta X_{it}}$ then the conditional survival function,

$$S\left(j|X_{ij},\nu_i\right)=e^{-\int_0^\infty \lambda_i(s|X_{is},\nu_i)ds}=e^{-\int_0^\infty \lambda_0(s)\nu_i e^{\beta X_{is}}ds}=e^{-\nu_i\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds}.$$

The marginal survival function is then

$$\begin{aligned}S\left(j|X_{ij}\right)&=\int_0^\infty S\left(j|X_{ij},\nu_i\right)dF\left(\nu_i\right)=\int_0^\infty e^{-\nu_i\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds}dF\left(\nu_i\right)\\&=M_{\nu_i}\left(-\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds\right)=\frac{1}{\left[1+\tau^2\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds\right]^{\frac{1}{\tau^2}}},\end{aligned}$$

where $M_{\nu_i}()$ is the moment generating function of ν_i. The marginal distribution function is

$$f\left(j|X_{ij}\right)=\frac{d}{dj}F\left(j|X_{ij}\right)=\frac{d}{dj}\left[1-S\left(j|X_{ij}\right)\right]=\frac{d}{dj}\left[1-\frac{1}{\left[1+\tau^2\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds\right]^{\frac{1}{\tau^2}}}\right]$$

$$=\frac{\frac{1}{\tau^2}\frac{d}{dj}\left(1+\tau^2\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds\right)\left[1+\tau^2\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds\right]^{\frac{1}{\tau^2}-1}}{\left[1+\tau^2\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds\right]^{\frac{2}{\tau^2}}}$$

$$= \frac{\frac{1}{\tau^2}\tau^2\lambda_0(j)e^{\beta X_{ij}}}{\left[1+\tau^2\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds\right]^{\frac{1}{\tau^2}+1}} = \frac{\lambda_0(j)e^{\beta X_{ij}}}{\left[1+\tau^2\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds\right]^{\frac{1}{\tau^2}+1}}.$$

Therefore, the marginal hazard function,

$$\lambda\left(j|X_{ij}\right) = \frac{f\left(j|X_{ij}\right)}{S\left(j|X_{ij}\right)} = \frac{\lambda_0(j)e^{\beta X_{ij}}}{\left[1+\tau^2\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds\right]^{\frac{1}{\tau^2}+1}}\left[1+\tau^2\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds\right]^{\frac{1}{\tau^2}}$$

$$= \frac{f\left(j|X_{ij}\right)}{S\left(j|X_{ij}\right)} = \frac{\lambda_0(j)e^{\beta X_{ij}}}{1+\tau^2\int_0^\infty \lambda_0(s)e^{\beta X_{is}}ds} \neq \lambda_0(j)e^{\beta X_{ij}}$$

if $\tau^2 \neq 0$.

A4.2 Approximation of $\mathbf{P}\left(T = j|X_{i1},\ldots,X_{ij},\nu_i\right)$

$$\begin{aligned}
\mathrm{P}\left(T \geq j|X_{i1},\ldots,X_{ij},\nu_i\right) &= \mathrm{P}\left(T \geq j|T \geq j-1, X_{i1},\ldots,X_{ij},\nu_i\right)\mathrm{P}\left(T \geq j-1|X_{i1},\ldots,X_{i(j-1)},\nu_i\right) \\
&= \mathrm{P}\left(T \geq j|T \geq j-1, X_{i1},\ldots,X_{ij},\nu_i\right) \\
&\quad * \mathrm{P}\left(T \geq j-1|T \geq j-2, X_{i1},\ldots,X_{i(j-1)},\nu_i\right)\mathrm{P}\left(T \geq j-2|X_{i1},\ldots,X_{i(j-2)},\nu_i\right) \\
&= \mathrm{P}\left(T \geq j|T \geq j-1, X_{i1},\ldots,X_{ij},\nu_i\right) \\
&\quad * \mathrm{P}\left(T \geq j-1|T \geq j-2, X_{i1},\ldots,X_{i(j-1)},\nu_i\right) \\
&\qquad \vdots \\
&\quad * \mathrm{P}\left(T \geq 2|T \geq 1, X_{i1}, X_{i2},\nu_i\right)\mathrm{P}\left(T \geq 1|, X_{i1},\nu_i\right) \\
&= \left[1-\mathrm{P}\left(T = j-1|T \geq j-1, X_{i1},\ldots,X_{i(j-1)},\nu_i\right)\right] \\
&\quad * \left[1-\mathrm{P}\left(T = j-2|T \geq j-2, X_{i1},\ldots,X_{i(j-2)},\nu_i\right)\right] \\
&\qquad \vdots \\
&\quad * \left[1-\mathrm{P}\left(T = 1|T \geq 1, X_{i1},\nu_i\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \left[1 - \lambda_i\left(j-1|X_{i1},\ldots,X_{i(j-1)},\nu_i\right)\right] \\
&\quad * \left[1 - \lambda_i\left(j-2|X_{i1},\ldots,X_{i(j-2)},\nu_i\right)\right] \\
&\qquad\qquad \vdots \\
&\quad * \left[1 - \lambda_i\left(1|X_{i1},\nu_i\right)\right] \\
&= \prod_{s=1}^{j-1}\left[1 - \lambda_i\left(s|X_{i1},\ldots,X_{is},\nu_i\right)\right]
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{P}\left(T=j|X_{i1},\ldots,X_{ij},\nu_i\right) &= \mathrm{P}\left(T=j|T\geq j,X_{i1},\ldots,X_{ij},\nu_i\right)\mathrm{P}\left(T\geq j|X_{i1},\ldots,X_{ij},\nu_i\right) \\
&= \lambda_i\left(j|X_{ij},\nu_i\right)\prod_{s=1}^{j-1}\left[1-\lambda_i\left(s|X_{i1},\ldots,X_{is},\nu_i\right)\right] \\
&= \lambda_i\left(j|X_{ij},\nu_i\right)e^{\log\left(\prod_{s=1}^{j-1}\left[1-\lambda_i\left(s|X_{i1},\ldots,X_{is},\nu_i\right)\right]\right)} \\
&= \lambda_i\left(j|X_{ij},\nu_i\right)e^{\sum_{s=1}^{j-1}\log\left[1-\lambda_i\left(s|X_{i1},\ldots,X_{is},\nu_i\right)\right]} \\
&= \lambda_i\left(j|X_{ij},\nu_i\right)e^{\sum_{s=1}^{j-1}-\sum_{n=1}^{\infty}\frac{\lambda_i\left(s|X_{i1},\ldots,X_{is},\nu_i\right)^n}{n}} \qquad \left[\text{for } \left|\lambda_i\left(s|X_{i1},\ldots,X_{is},\nu_i\right)\right|<1\right] \\
&\approx \lambda_i\left(j|X_{ij},\nu_i\right)e^{-\sum_{s=1}^{j-1}\left(\lambda_i\left(s|X_{i1},\ldots,X_{is},\nu_i\right)+\frac{\lambda_i\left(s|X_{i1},\ldots,X_{is},\nu_i\right)^2}{2}\right)}.
\end{aligned}
$$

The approximation holds within four decimal places for $\lambda_i\left(s|X_{i1},\ldots,X_{is},\nu_i\right)<10\%$.

A4.3 Distribution Function of $Y_{ijk},\nu_i|\eta,X_{i1},\ldots,X_{ij},l_i,r_i$

$$
\begin{aligned}
\mathrm{P}\left(Y_{ijk},\nu_i|\eta,X_{i1},\ldots,X_{ij},l_i,r_i\right) &= \mathrm{P}\left(Y_{ijk}|\nu_i,\eta,X_{i1},\ldots,X_{ij},l_i,r_i\right)\mathrm{P}\left(\nu_i|\eta,X_{i1},\ldots,X_{ij},l_i,r_i\right) \\
&= \mathrm{P}\left(Y_{ijk}|\nu_i,\eta,X_{i1},\ldots,X_{ij},l_i,r_i\right)\mathrm{P}\left(\nu_i|\eta,X_{i1},\ldots,X_{ir_i},l_i,r_i\right) \\
&= \mathrm{P}\left(Y_{ijk}|\nu_i,\eta,X_{i1},\ldots,X_{ij},l_{ik},r_{ik}\right)\frac{\mathrm{P}\left(l_i,r_i|\nu_i,\eta,X_{i1},\ldots,X_{ir_i}\right)\mathrm{P}\left(\nu_i|\eta,X_{i1},\ldots,X_{ir_i}\right)}{\mathrm{P}\left(l_i,r_i|\eta,X_{i1},\ldots,X_{ir_i}\right)}
\end{aligned}
$$

$$= \mathrm{P}\left(Y_{ijk}|\nu_i,\eta,X_{i1},\ldots,X_{ij},l_{ik},r_{ik}\right)\left[\prod_{m=1}^{K_i}\mathrm{P}\left(l_{im},r_{im}|\nu_i,\eta,X_{i1},\ldots,X_{ir_i}\right)\right]$$

$$* \frac{\mathrm{P}\left(\nu_i|\eta,X_{i1},\ldots,X_{ir_i}\right)}{\mathrm{P}\left(l_i,r_i|\eta,X_{i1},\ldots,X_{ir_i}\right)}$$

$$\approx \left[\mathbb{I}_{[l_{ik}\leq j\leq r_{ik}]}\frac{e^{-\nu_i\Lambda_{ij-1}-\nu_i^2\Lambda'_{ij-1}}-e^{-\nu_i\Lambda_{ij}-\nu_i^2\Lambda'_{ij}}}{e^{-\nu_i\Lambda_{il_{ik}-1}-\nu_i^2\Lambda'_{il_{ik}-1}}-e^{-\nu_i\Lambda_{ir_{ik}}-\nu_i^2\Lambda'_{ir_{ik}}}}\right]^{Y_{ijk}}$$

$$* \left[1-\mathbb{I}_{[l_{ik}\leq j\leq r_{ik}]}\frac{e^{-\nu_i\Lambda_{ij-1}-\nu_i^2\Lambda'_{ij-1}}-e^{-\nu_i\Lambda_{ij}-\nu_i^2\Lambda'_{ij}}}{e^{-\nu_i\Lambda_{il_{ik}-1}-\nu_i^2\Lambda'_{il_{ik}-1}}-e^{-\nu_i\Lambda_{ir_{ik}}-\nu_i^2\Lambda'_{ir_{ik}}}}\right]^{1-Y_{ijk}}$$

$$* \prod_{m=1}^{K_i}\left(e^{-\nu_i\Lambda_{il_{im}-1}-\nu_i^2\Lambda'_{il_{im}-1}}-e^{-\nu_i\Lambda_{ir_{im}}-\nu_i^2\Lambda'_{ir_{im}}}\right)$$

$$* \frac{\nu_i^{\frac{1}{\tau^2}-1}e^{-\frac{\nu_i}{\tau^2}}}{\mathrm{P}\left(l_i,r_i|\eta,X_{i1},\ldots,X_{ir_i}\right)(\tau^2)^{\frac{1}{\tau^2}}\Gamma(\tau^2)}.$$

A4.4 Expected values of the E-step with Interval Censoring

The following expectations were evaluated numerically to execute the E-step in the presence of interval censoring.

1.

$$E\left[D_{ij}\log\left(\nu_i\right)|\eta,X_{i1},\ldots,X_{ij},l_i,r_i\right] = \sum_{\mathcal{R}_{ij}}E_{\nu_i}\left[E_{Y_{ijk}}\left\{Y_{ijk}\log\left(\nu_i\right)|\nu_i,\eta,X_{i1},\ldots,X_{ij},l_i,r_i\right\}\right]$$

$$= \sum_{\mathcal{R}_{ij}}\int_{\nu_i}\sum_{Y_{ijk}=\{0,1\}}Y_{ijk}\log\left(\nu_i\right)\mathrm{P}\left(Y_{ijk},\nu_i|\eta,X_{i1},\ldots,X_{ij},l_i,r_i\right)\,\mathrm{d}\nu_i$$

$$\approx \sum_{\mathcal{R}_{ij}}\int_{\nu_i}\log\left(\nu_i\right)\left[\mathbb{I}_{[l_{ik}\leq j\leq r_{ik}]}\frac{e^{-\nu_i\Lambda_{ij-1}-\nu_i^2\Lambda'_{ij-1}}-e^{-\nu_i\Lambda_{ij}-\nu_i^2\Lambda'_{ij}}}{e^{-\nu_i\Lambda_{il_{ik}-1}-\nu_i^2\Lambda'_{il_{ik}-1}}-e^{-\nu_i\Lambda_{ir_{ik}}-\nu_i^2\Lambda'_{ir_{ik}}}}\right]^{Y_{ijk}}$$

$$* \prod_{m=1}^{K_i} \left(e^{\text{-}\nu_i \Lambda_{il_{im}-1} - \nu_i^2 \Lambda'_{il_{im}-1}} - e^{\text{-}\nu_i \Lambda_{ir_{im}} - \nu_i^2 \Lambda'_{ir_{im}}} \right)$$

$$* \frac{\nu_i^{\frac{1}{\tau^2}\text{-}1} e^{\text{-}\frac{\nu_i}{\tau^2}}}{\mathrm{P}\left(l_i, r_i | \eta, X_{i1}, \ldots, X_{ir_i}\right) \left(\tau^2\right)^{\frac{1}{\tau^2}} \Gamma(\tau^2)} \, \mathrm{d}\nu_i.$$

2.

$$E\left[D_{ij} | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right] = \sum_{\mathcal{R}_{ij}} E_{\nu_i} \left[E_{Y_{ijk}} \left\{ Y_{ijk} \log\left(\nu_i\right) | \nu_i, \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i \right\} \right]$$

$$= \sum_{\mathcal{R}_{ij}} \int_{\nu_i} \sum_{Y_{ijk}=\{0,1\}} Y_{ijk} P\left(Y_{ijk}, \nu_i | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right) \, \mathrm{d}\nu_i$$

$$\approx \sum_{\mathcal{R}_{ij}} \int_{\nu_i} \left[\mathbb{I}_{[l_{ik} \leq j \leq r_{ik}]} \frac{e^{\text{-}\nu_i \Lambda_{ij-1} - \nu_i^2 \Lambda'_{ij-1}} - e^{\text{-}\nu_i \Lambda_{ij} - \nu_i^2 \Lambda'_{ij}}}{e^{\text{-}\nu_i \Lambda_{il_{ik}-1} - \nu_i^2 \Lambda'_{il_{ik}-1}} - e^{\text{-}\nu_i \Lambda_{ir_{ik}} - \nu_i^2 \Lambda'_{ir_{ik}}}} \right]^{Y_{ijk}}$$

$$* \prod_{m=1}^{K_i} \left(e^{\text{-}\nu_i \Lambda_{il_{im}-1} - \nu_i^2 \Lambda'_{il_{im}-1}} - e^{\text{-}\nu_i \Lambda_{ir_{im}} - \nu_i^2 \Lambda'_{ir_{im}}} \right)$$

$$* \frac{\nu_i^{\frac{1}{\tau^2}\text{-}1} e^{\text{-}\frac{\nu_i}{\tau^2}}}{\mathrm{P}\left(l_i, r_i | \eta, X_{i1}, \ldots, X_{ir_i}\right) \left(\tau^2\right)^{\frac{1}{\tau^2}} \Gamma(\tau^2)} \, \mathrm{d}\nu_i.$$

3.

$$E\left[D_{ij} \nu_i | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right] = \sum_{\mathcal{R}_{ij}} E_{\nu_i} \left[E_{Y_{ijk}} \left\{ Y_{ijk} \log\left(\nu_i\right) | \nu_i, \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i \right\} \right]$$

$$= \sum_{\mathcal{R}_{ij}} \int_{\nu_i} \sum_{Y_{ijk}=\{0,1\}} Y_{ijk} \nu_i P\left(Y_{ijk}, \nu_i | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right) \, \mathrm{d}\nu_i$$

$$\approx \sum_{\mathcal{R}_{ij}} \int_{\nu_i} \left[\mathbb{I}_{[l_{ik} \le j \le r_{ik}]} \frac{e^{-\nu_i \Lambda_{ij-1} - \nu_i^2 \Lambda'_{ij-1}} - e^{-\nu_i \Lambda_{ij} - \nu_i^2 \Lambda'_{ij}}}{e^{-\nu_i \Lambda_{il_{ik}-1} - \nu_i^2 \Lambda'_{il_{ik}-1}} - e^{-\nu_i \Lambda_{ir_{ik}} - \nu_i^2 \Lambda'_{ir_{ik}}}} \right]^{Y_{ijk}}$$

$$* \prod_{m=1}^{K_i} \left(e^{-\nu_i \Lambda_{il_{im}-1} - \nu_i^2 \Lambda'_{il_{im}-1}} - e^{-\nu_i \Lambda_{ir_{im}} - \nu_i^2 \Lambda'_{ir_{im}}} \right)$$

$$* \frac{\nu_i^{\frac{1}{\tau^2}} e^{-\frac{\nu_i}{\tau^2}}}{\mathrm{P}\left(l_i, r_i | \eta, X_{i1}, \ldots, X_{ir_i}\right) \left(\tau^2\right)^{\frac{1}{\tau^2}} \Gamma(\tau^2)} \, \mathrm{d}\nu_i.$$

4.

$$E\left[\bar{D}_i \nu_i | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right] = E\left[\left(K_i - \sum_j D_{ij} \right) \nu_i | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i \right]$$

$$= K_i E\left[\nu_i | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right] - \sum_j E\left[D_{ij} \nu_i | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right],$$

where $E\left[\nu_i | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right]$ is approximated below.

5.

$$E\left[\log\left(\nu_i\right) | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right] = \int_{\nu_i} \sum_{Y_{ijk} = \{0,1\}} \log\left(\nu_i\right) P\left(Y_{ijk}, \nu_i | \eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right) \, \mathrm{d}\nu_i$$

$$\approx \int_{\nu_i} \log\left(\nu_i\right) \frac{\prod_{m=1}^{K_i} \left(e^{-\nu_i \Lambda_{il_{im}-1} - \nu_i^2 \Lambda'_{il_{im}-1}} - e^{-\nu_i \Lambda_{ir_{im}} - \nu_i^2 \Lambda'_{ir_{im}}} \right) \nu_i^{\frac{1}{\tau^2}-1} e^{-\frac{\nu_i}{\tau^2}}}{P\left(l_i, r_i | \eta, X_{i1}, \ldots, X_{ij}\right) \left(\tau^2\right)^{\frac{1}{\tau^2}} \Gamma\left(\frac{1}{\tau^2}\right)} \, \mathrm{d}\nu_i.$$

6.

$$E\left[\nu_i|\eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right] = \int_{\nu_i} \sum_{Y_{ijk}=\{0,1\}} \nu_i P\left(Y_{ijk}, \nu_i|\eta, X_{i1}, \ldots, X_{ij}, l_i, r_i\right) \, \mathrm{d}\nu_i$$

$$\approx \int_{\nu_i} \nu_i \frac{\prod_{m=1}^{K_i} \left(e^{-\nu_i \Lambda_{il_{im}-1} - \nu_i^2 \Lambda'_{il_{im}-1}} - e^{-\nu_i \Lambda_{ir_{im}} - \nu_i^2 \Lambda'_{ir_{im}}} \right) \nu_i^{\frac{1}{\tau^2}-1} e^{-\frac{\nu_i}{\tau^2}}}{P\left(l_i, r_i|\eta, X_{i1}, \ldots, X_{ij}\right) (\tau^2)^{\frac{1}{\tau^2}} \Gamma\left(\frac{1}{\tau^2}\right)} \, \mathrm{d}\nu_i.$$

A4.5 Variance of $\mathbf{Y = \log(X)}$ when $\mathbf{X \sim \Gamma\left(a, \text{scale} = b\right)}$

If $X \sim \Gamma\left(a, \text{scale} = b\right)$ then $f_X(x) = \frac{x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a}$, $x > 0$. The moment generating function of Y is

$$M_Y(t) = E\left(e^{tY}\right) = E\left(e^{t\log(X)}\right) = E\left(X^t\right) = \int_0^\infty x^t \frac{x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a} dx$$

$$= \int_0^\infty \frac{x^{a+t-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a} dx = \frac{\Gamma(a+t) b^{a+t}}{\Gamma(a) b^a} \int_0^\infty \frac{x^{a+t-1} e^{-\frac{x}{b}}}{\Gamma(a+t) b^{a+t}} dx$$

$$= \frac{\Gamma(a+t) b^{a+t}}{\Gamma(a) b^a} = \frac{\Gamma(a+t) b^t}{\Gamma(a)}.$$

Therefore,

$$E(Y) = \left. \frac{d}{dt} M_Y(t) \right|_{t=0} = \left. \frac{d}{dt} \frac{\Gamma(a+t) b^t}{\Gamma(a)} \right|_{t=0}$$

$$= \left. \frac{\Gamma(a+t) \psi\left(a+t\right) b^t}{\Gamma(a)} \right|_{t=0} + \left. \frac{\Gamma(a+t) \log(b) b^t}{\Gamma(a)} \right|_{t=0}$$

$$= \frac{\Gamma(a)\psi\left(a\right)}{\Gamma(a)} + \frac{\Gamma(a)\log(b)}{\Gamma(a)} = \psi\left(a\right) + \log(b),$$

where $\psi()$ is the derivative of the log of the gamma function.

$$E\left(Y^2\right) = \frac{d^2}{dt^2} M_Y(t)\Big|_{t=0} = \frac{d}{dt}\frac{\Gamma(a+t)\psi\left(a+t\right)b^t}{\Gamma(a)}\Big|_{t=0} + \frac{d}{dt}\frac{\Gamma(a+t)\log(b)b^t}{\Gamma(a)}\Big|_{t=0}$$

$$= \frac{\Gamma(a+t)\psi^2\left(a+t\right)b^t}{\Gamma(a)}\Big|_{t=0} + \frac{\Gamma(a+t)\psi^{'}\left(a+t\right)b^t}{\Gamma(a)}\Big|_{t=0}$$

$$+ \frac{\Gamma(a+t)\psi\left(a+t\right)\log(b)b^t}{\Gamma(a)}\Big|_{t=0} + \frac{\Gamma(a+t)\psi\left(a+t\right)\log(b)b^t}{\Gamma(a)}\Big|_{t=0}$$

$$+ \frac{\Gamma(a+t)\log^2(b)b^t}{\Gamma(a)}\Big|_{t=0}$$

$$= \frac{\Gamma(a)\psi^2\left(a\right)}{\Gamma(a)} + \frac{\Gamma(a)\psi^{'}\left(a\right)}{\Gamma(a)} + \frac{\Gamma(a)\psi\left(a\right)\log(b)}{\Gamma(a)} + \frac{\Gamma(a)\psi\left(a\right)\log(b)}{\Gamma(a)} + \frac{\Gamma(a)\log^2(b)}{\Gamma(a)}$$

$$= \psi^2\left(a\right) + \psi^{'}\left(a\right) + \psi\left(a\right)\log(b) + \psi\left(a\right)\log(b) + \log^2(b)$$

$$= \left(\psi(a) + \log(b)\right)^2 + \psi^{'}(a).$$

Therefore, $\mathrm{Var}\left(\log(X)\right) = \mathrm{Var}(Y) = E\left(Y^2\right) - E^2(Y) = \psi^{'}(a)$.