

2017

# Causal Inference Using Variation In Treatment Over Time

Xinyao Ji
*University of Pennsylvania*, xinyaoji@wharton.upenn.edu

# Causal Inference Using Variation In Treatment Over Time

**Abstract**
This thesis and related research is motivated by my interest in understanding the use of time-varying treatments in causal inference from complex longitudinal data, which play a prominent role in public health, economics, and epidemiology, as well as in biological and medical sciences. Longitudinal data allow the direct study of temporal changes within individuals and across populations, therefore give us the edge to utilize time this important factor to explore causal relationships than static data. There are also a variety challenges that arise in analyzing longitudinal data. By the very nature of repeated measurements, longitudinal data are multivariate in various dimensions and have completed random-error structures, which make many conventional causal assumptions and related statistical methods are not directly applicable. Therefore, new methodologies, most likely data-driven, are always encouraged and sometimes necessary in longitudinal causal inference, as will be seen throughout this thesis
As a result of the various topics explored, this thesis is split into four parts corresponding to three dierent patterns of variation in treatment. The rst pattern is the one-directional change of a binary treatment assignment, meaning that each study participant is only allowed to experience the change from untreated to treated at the staggered time. Such pattern is observed in a novel cluster-randomized design called the stepped-wedge. The second pattern is the arbitrary switching of a binary treatment caused by changes in person-specic characteristics and general time trend. The patterns is the most common thing one would observe in longitudinal data and we develop a method utilizing trends in treatment to account for unmeasured confounding. The third pattern is that the underlying treatment, outcome, covariates are time-continuous, yet are only observed at discrete time points. Instead of modeling cross-sectional and pooled longitudinal data, we take a mechanistic view by modeling reactions among variables using stochastic dierential equations and investigate whether it is possible to draw sensible causal conclusions from discrete

measurements.

**Degree Type**
Dissertation

**Degree Name**
Doctor of Philosophy (PhD)

**Graduate Group**
Statistics

**First Advisor**
Dylan S. Small

**Keywords**
Causal inference, Longitudinal Data, Mendelian Randomization, Randomization Inference, Unmeasured Confounding

**Subject Categories**
Statistics and Probability

# CAUSAL INFERENCE USING VARIATION IN TREATMENT OVER TIME

Xinyao Ji

A DISSERTATION

in

Statistics

For the Graduate Group in
Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

**Supervisor of Dissertation**

Dylan S. Small
Professor of Statistics

**Graduate Group Chairperson**

Catherine Schrand
Celia Z. Moh Professor, Professor of Accounting

**Dissertation Committee**
Dylan S. Small, Professor
Sean Hennessy, Professor
Nancy R. Zhang, Associate Professor

CAUSAL INFERENCE USING VARIATION IN TREATMENT OVER TIME

# Acknowledgments

First and foremost, I want to thank my parents for the endless encouragement and love. Throughout all life of mine, they have always been there for me, showing unconditional understanding and support whenever I encounter difficulties.

I offer sincerest gratitude to my advisor, Dylan Small, who brought me to the world of causal inference and guided me in every aspect of research and professional pursuit ever since. I am also grateful to work with Sean Hennessy, who has broad knowledge and wonderful intuitions in epidemiological research, and inspired me to work on many interesting applications in my dissertation research. Furthermore, I want to thank Nancy Zhang for kindly agreeing to serve in my committee and sharing with me her insightful comments.

Last but not least, I would like to thank all other professors, my fellow student colleagues, and wonderful staff in the Wharton Statistics Department. They made my PHD life fruitful and enjoyable.

# ABSTRACT

# CAUSAL INFERENCE USING VARIATION IN TREATMENT OVER TIME

Xinyao Ji

Dylan S. Small

This thesis and related research is motivated by my interest in understanding the use of time-varying treatments in causal inference from complex longitudinal data, which play a prominent role in public health, economics, and epidemiology, as well as in biological and medical sciences. Longitudinal data allow the direct study of temporal changes within individuals and across populations, therefore give us the edge to utilize time this important factor to explore causal relationships than static data. There are also a variety challenges that arise in analyzing longitudinal data. By the very nature of repeated measurements, longitudinal data are multivariate in various dimensions and have completed random-error structures, which make many conventional causal assumptions and related statistical methods are not directly applicable. Therefore, new methodologies, most likely data-driven, are always encouraged and sometimes necessary in longitudinal causal inference, as will be seen throughout this thesis

As a result of the various topics explored, this thesis is split into four parts corresponding to three different patterns of variation in treatment. The first pattern is the one-directional change of a binary treatment assignment, meaning that each study participant is only allowed to experience the change from untreated to treated at the staggered time. Such pattern is observed in a novel cluster-randomized design called the stepped-wedge. The second pattern is the arbitrary switching of a binary treatment caused by changes in person-specific characteristics and general time trend. The patterns is the most common thing one would observe in longitudinal data and we develop a method utilizing trends in treatment to account for unmeasured confounding. The third pattern is that the underlying treatment, outcome, covariates are time-continuous, yet are only observed at discrete time points. Instead of modeling cross-sectional and pooled longitudinal data, we take a mechanistic view by modeling reactions among variables using stochastic differential equations and investigate whether it is possible to draw sensible causal conclusions from discrete measurements.

# Contents

# List of Tables

# List of Figures

# 1 Introduction

This thesis and related research is motivated by my interest in understanding the use of time-varying treatments in causal inference from complex longitudinal data, which play a prominent role in public health, economics, and epidemiology, as well as in biological and medical sciences. By measuring study participants over time, longitudinal data allow the direct study of temporal changes within individuals and across populations, therefore facilitate estimating the causal effect of certain treatment on the outcome. There are also a variety challenges that arise in analyzing longitudinal data. By their very nature, the repeated measures are multivariate and have a complex random-error structure that must be appropriated accounted for in the analysis. Hence, conventional causal assumptions and related statistical methods are not directly applicable to the topics that I explored in the thesis. They are either extended or modified and coupled with new methodologies depending on each data-driven problem.

As a result of the various topics explored, this thesis is split into four parts corresponding to three different patterns of variation in treatment. The first pattern is the one-directional change of a binary treatment assignment, meaning that each study participant is only allowed to experience the change from untreated to treated

at staggered time. Such pattern is observed in a novel cluster-randomized design called the stepped-wedge. The second pattern is the arbitrary switching of a binary treatment caused by changes in person-specific characteristics and general time trend. The third pattern is that the underlying treatment, outcome, and other covariates are time-continuous, and that they are only observed at discrete time points.

The majority of the work in this document is joint with my adviser Dyan S. Small and each chapter's contributors are appropriately marked at the beginning of the chapter and project-specific acknowledgments are appropriately marked at the close of each chapter.

## Chapter II

Here, I write about the work I did in *stepped-wedge cluster-randomized experiments* in which all clusters start out in the control and then clusters are randomized to cross over to the treatment at staggered time. I investigated statistical properties of the stepped-wedge design following the parametric mixed model approach proposed by Hussey and Hughes in 2007. I found that testing for the treatment effect is generally sensitive to specification of the parametric model. For instance, if one fails to account for cluster-by-time interactions present in the data, the Type I error rate is severely inflated. My collaborators and I developed a more robust and efficient strategy-randomization inference as a unified framework to test for constant treatment effects with guaranteed Type I error rate and satisfying power. The proposed method was applied to the Burkina Faso CBHI dataset to study the financial benefits of community-based health insurance (CBHI) schemes. We concluded that the insurance had limited effects on reducing the likelihood of low to moderate levels of catastrophic health expenditure, but substantially reduced the likelihood of extremely high health

expenditure that exceeds half of a person's monthly income.

## Chapter III

One-directional change of treatment status is a very strict rule for treatment assignment and randomized experiments are not always ethical and realistic to be implemented. For the majority of existing observational data sets in the fields of epidemiology, treatments, like prescriptions of a drug, are observed to switch statuses in unequal-spaced time intervals. Epidemiologic designs such as the cohort study have been extensively studied to examine causal effects of such treatment using individual-level data, yet can be biased if there are unmeasured competing risk factors. As a response, my collaborators and I proposed a hybrid ecologic-epidemiologic design called the trend-in-trend (TT) design utilizing a strong time-trend in exposure. Rather than comparing exposed vs. unexposed persons, the TT design derives causal estimates by examining time-trends in outcome as a function of time-trends in exposure across strata with different time-trends in exposure. We gave a mathematical derivation of using aggregated data that are covariates-free to infer individual-level causality. We reported a simulation study illustrating that the odds ratio estimated using the TT method is much less biased than that estimated using cohort methods when there is unmeasured confounding. Finally, we applied the TT method to healthcare data to reproduce the known positive association between rofecoxib and acute myocardial infarction (AMI), and two presumably null associations: rofecoxib and severe hypoglycemia, and rofecoxib and bone fracture.

## Chapter IV

The inference part of the TT design inspired me and my collaborators to think about developing corresponding a sequential testing methodology. Sequential testing methods are powerful tools that facilitate early termination of a newly introduced drug when the drug exceeds the pre-assumed adverse event rate. However, all the existing sequential testing methods in observational settings rely heavily on the unstable assumption of no unmeasured confounding. Because electronic health records are not, in general, collected for scientific purposes, the no unmeasured confounding assumption is unlikely to hold. We generalized the standard sequential likelihood ratio test to trend-in-trend design settings that utilizes time trends in exposure prevalence and accounts for both measured and unmeasured confounding under certain assumptions. As the log likelihood ratio of the TT design does not have known asymptotic distribution, we approximated the critical value using a Monte Carlo simulation method involving $\alpha$-spending approaches. The performance of the proposed approach is examined and compared to other approaches using simulation studies. We showed that the results obtained by the existing methods may be misleading with an inflated Type I error rate in the presence of unmeasured confounding while the proposed method provides valid results.

## Chapter V

While working on projects involving exploring discrete time-varying treatments, I realized that in many real world data sets, it is more reasonable to assume that the underlying processes are time-continuous processes, and that are only observed at discrete time points. In particular, most cross-sectional and pooled panel data that the Mendelian randomization is applied to are actually discrete snapshots of dynamic processes, in which the outcome, the exposure, and covariates exhibit non-

negligible serial correlations, and the outcome is an important determinant of future exposure levels either directly or indirectly through other unmeasured factors. When observations are inherently dynamic, conclusions from the Mendelian randomization are hard to be interpreted. Even worse is that the conventional exclusion restriction assumption, which is crucial for the validity of causal estimates is ungrounded so that estimators derived based on static analysis could be severely biased from what they were intended to estimate. We are therefore motivated to take a mechanistic view by modeling reactions among variables using stochastic differential equations. We show that discrete observations of a s time-continuous process generally obscure the underlying local independence between the outcome and the instrument. Hence, applying the Mendelian Randomization to discrete observational data without explicit time justification, could give insensible conclusions.

# 2

# Randomization Inference for Stepped-wedge Cluster-Randomized Trials: an Application to Community-based Health Insurance


## Abstract

National health insurance schemes are generally impractical in low-income countries due to limited resources and low organizational capacity. In response to such obstacles, community-based health insurance (CBHI) schemes have emerged over the past 20 years. CBHIs are designed to reduce the financial burden generated by unanticipated treatment cost among individuals falling sick, and thus are expected to make health care more affordable. In this paper, we investigate whether CBHI schemes effectively protect individuals against large financial shocks using a stepped-wedge cluster-randomized design on data from a CBHI program rolled out in rural Burkina Faso. We investigate statistical properties of the stepped-wedge design following the parametric mixed model approach proposed by Hussey and Hughes in 2007. We find that testing for the treatment effect is generally sensitive to specification of the


*Joint work with Gunther Fink, Paul Jacob Robyn, Dylan Small

parametric model. For instance, if we fail to account for cluster-by-time interactions present in the data, the Type I error rate is severely inflated. We develop a more robust and efficient strategy – randomization inference. We demonstrate how to apply randomization inference to test for constant treatment effects and discuss test statistics suitable for the stepped-wedge design. Randomization inference guarantees a valid Type I error rate; simulation studies show that randomization inference test statistics also have power that is comparable to the currently used procedures that do not guarantee a valid Type I error rate. Finally, we apply our proposed method to the Burkina Faso CBHI dataset. We conclude that the insurance had limited effects on reducing the likelihood of low to moderate levels of catastrophic health expenditure, but substantially reduced the likelihood of extremely high health expenditure that exceeds half of a person's monthly income.

## 2.1 Introduction and Motivation

### Community-based health insurance

The design of adequate health financing systems in low-income countries is a subject of significant debate. Due to low or modest economic growth, limited public tax resources, and low organizational capacity, national health insurance schemes are generally impractical. In response to such obstacles, community-based health insurance (CBHI) schemes, which are comparatively easier to set up, have emerged over the past 20 years (Asenso-Okyere et al., 1997; De Allegri et al., 2006; Devadasan et al., 2006; Ekman, 2004; Wang et al., 2009).

CBHI schemes are micro-insurance schemes that are voluntary, not-for-profit health insurance schemes organized at the community level. Under CBHI schemes, members

of a community, often defined by geographical proximity or through employment-based relationships, pool resources in order to provide support for covering health expenditure (Robyn et al., 2012). CBHI schemes seek to reduce the financial burden generated by unanticipated treatment cost among individuals falling sick and are thus expected to make health care more affordable. A natural question that emerges then is: do CBHI schemes work as intended and in fact enhance universal financial protection?

We consider a study of a CBHI program in rural Burkina Faso that was implemented by the Ministry of Health and Nouna Health Research Center in the Nouna District using a stepped-wedge cluster-randomized trial (Fink et al., 2013). We discuss properties of stepped-wedge cluster-randomized trials, problems with the currently used analysis methods for stepped-wedge cluster randomized trials, present solutions to these problems and analyze the study of the CBHI program in Burkina Faso.

## Stepped-wedge cluster-randomized trials

A stepped-wedge cluster-randomized trial is a one-way crossover trial in which all clusters start out in the control and then clusters are randomized to cross over to the treatment at staggered times (Hall et al., 1987; Hussey and Hughes, 2007). Figure 2.1 illustrates the treatment schedule for a stepped-wedge trial; the name "stepped-wedge" refers to the series of steps of the treatment schedule, which results in a wedge shape.

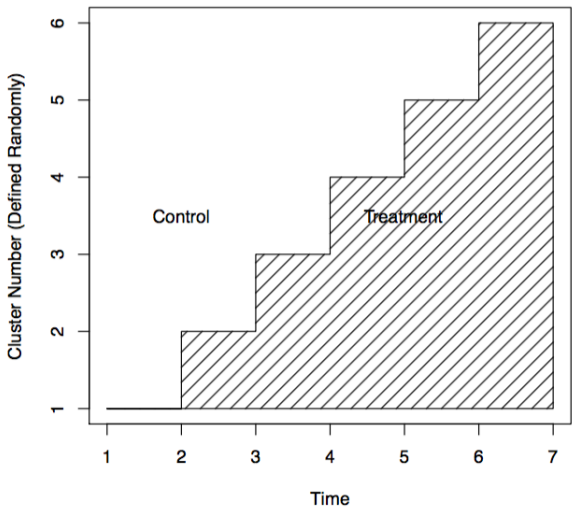



Figure 2.1: Illustration of a stepped-wedge design where different groups of clusters switch from control to treatment during different calendar periods.

The stepped-wedge design has been gaining popularity in recent years because of a number of attractive features (Mdege et al., 2011). First, the stepped-wedge design is useful for settings in which limited resources or geographical constraints make it financially or logistically difficult to start the intervention in many clusters at once (e.g., (Brown and Lilford, 2006; Hall et al., 1987; Mdege et al., 2011; Moulton et al., 2007)). For example, in a parallel design (randomize half the clusters to treatment during during single calendar period) or a traditional crossover design (randomize half the clusters to treatment at baseline and then switch these clusters to control and the other clusters to treatment midway through the trial), the intervention must be implemented in half of clusters simultaneously, while the stepped-wedge design allows researchers to implement the intervention in a smaller fraction of clusters during each calendar period (Hussey and Hughes, 2007). Second, the stepped-wedge design (as with a traditional crossover design) allows clusters to serve as their own controls, which increases power when there are substantial cluster effects (Woertman et al., 2013). The stepped-wedge design differs from a traditional crossover design, however,

in that the crossovers are only in one direction; in particular, the intervention is never removed once it has been implemented (at least over the course of the trial). Third, because all clusters receive the treatment by the end of the trial and a cluster is never withdrawn from receiving the treatment, the stepped-wedge design is particularly useful for settings in which it would not be ethical, healthy or practical to withdraw the treatment or in which it would be difficult for participants to quickly revert to their pretreatment condition quickly after the withdrawal (Rhoda et al., 2011). The stepped-wedge design is also useful for evaluating the population-level impact of an intervention that has been shown to be effective in an individually randomized trial or for which there is a majority opinion that the intervention will be effective so that equipoise does not exist (Hussey and Hughes, 2007).

All these features made the stepped-wedge design ideal for studying the benefits of the CBHI program in Burkina Faso. Because the CBHI program was expected to confer benefits, every village in the study area wanted to be enrolled in the program at the early stage. However, it takes time to scale up the program, so the CBHI management team and the health district had no option but to rollout the program in a progressive manner. The stepped-wedge design allowed the program to be rolled out in a fair manner and the effect of the program to be studied through a randomized trial. The stepped-wedge nature of the trial also helped to alleviate the spillover effect, as the incentive to migrate to a different area just to benefit from the intervention was counterbalanced by the fact that this very same intervention was going to be implemented in the entire study area within the next few years.

## Analysis Methods

In line with the increasing interest in employing and implementing the stepped-wedge design, a handful of pivotal articles on testing intervention effects, sample size calcula-

tions, and analytical methods for continuous or dichotomous outcomes have emerged in the literature (e.g.,(Dimairo et al., 2011; Hussey and Hughes, 2007; Moulton et al., 2007; Woertman et al., 2013)). Most of them have adopted the linear mixed model approach proposed by Hussey and Hughes (2007).

Hussey and Hughes (2007) considered the linear mixed model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + Z_{ij}\theta + e_{ijk}, \tag{2.1}$$

where $Y_{ijk}$ is the observed response corresponding to individual $k$ during calendar period $j$ from cluster $i$ and $Z_{ij}$ denotes whether cluster $i$ has been assigned the treatment by calendar period $j$. $\alpha_i$ is a random effect for cluster $i$ such that $\alpha_i$ are iid $N(0, \tau^2)$, $\beta_j$ is a fixed effect corresponding to time interval $j$ ($j$ in $1, \ldots, T-1$, $\beta_T = 0$ for identifiability), $\theta$ is the treatment effect and $e_{ijk}$ are individual, time period specific effects that are assumed to be iid $N(0, \sigma_e^2)$ and independent of $\alpha_i$.

One possible violation of assumptions in the linear mixed model (2.1) is the existence of cluster-by-time interactions, which are prevalent in a number of settings. For example, cluster-by-time interactions were a concern in a recent proposal for using the stepped-wedge design to study a vaccine for Ebola while the Ebola epidemic was going on, because the Ebola epidemic, like other pandemics, was spreading from place to place over time (Bellan et al., 2015; van der Tweel and van der Graaf, 2013). In the CBHI study we are considering, cluster-by-time interactions are a concern because the clusters are communities that are affected by different local economic and political developments.

From a statistical point of view, adding an interaction term would increase the correlation among observations within the same cluster and time period, therefore changing the structure of the covariance matrix. Consequently, a larger sample size would be required to achieve sufficient statistical power and, even worse, lead to

invalid levels if the testing procedure is not chosen carefully.

Including all cluster-by-time interactions into the model as fixed effects would make the treatment effect unidentifiable. Hussey and Hughes (2007) proposed one strategy to deal with cluster-by-time interactions and still be able to estimate the treatment effect: create strata of clusters with similar expected time trends and then include stratum-by-time interaction as a factor in the model. This strategy requires some knowledge of the expected time trends before the trial and runs the risk of overfitting if interactions do not exist or are negligible. Without strong a priori knowledge of the pattern of cluster-by-time interactions, a better approach is needed to gauge the treatment effect than either excluding cluster-by-time interactions or including a specific pattern of them.

## Randomization inference

In this paper, we develop another approach for the analysis of stepped-wedge cluster-randomized trials that accounts for potential cluster-by-time interactions – randomization inference. In randomization inference as developed by Fisher (1935), hypotheses are tested using only the assumption that the randomization has been properly carried out. Fisher said that randomization inference is "reasoned basis for inference" because it uses only the physical act of randomization as a basis for inference, and is exact and distribution-free. Tukey (1993) said that randomization inference is the "platinum standard" inference. For discussion and examples of randomization inference, see Welch (1937); Raz (1990); Gail et al. (1996); Braun and Feng (2001); Rosenbaum (a,b); Greevy et al. (2004); Ho and Imai (2006); Small et al. (2008); Hansen and Bowers (2009).

Randomization inference can be applied to any test statistic of treatment effects. Here we consider Wald test statistics based on the model (2.1) or other generalized

linear mixed models. Because the randomization procedure adds an extra layer of security to the inference, the Type I error rate is valid even if parametric models for responses are misspecified such as failing to account for cluster-by-time interactions.

We contribute to the literature by applying randomization inference to stepped-wedge cluster-randomized trials. We build a unified framework to develop the randomization distribution for any test statistic, which can be used to calculate p-values and construct confidence intervals. Regarding our specific question, to what extent do CBHI schemes enhance universal financial protection, we use the data from the Burkina Faso study (Fink et al., 2013) to examine whether CBHI schemes help to reduce the likelihood of catastrophic health expenditure. Our final results show that the insurance had limited effects on reducing the likelihood of low to moderate levels of catastrophic health expenditure, but substantially reduced the likelihood of extremely high health expenditure that exceeds half of a person's monthly income.

The outline of our paper is as follows. In Section 2, we introduce the potential outcomes notation and set up that will be used throughout the paper. In Section 3, we discuss consequences of failing to consider cluster-by-time interactions. In Section 4, we develop our randomization inference approach for the stepped-wedge design. In Section 5, we conduct simulation studies comparing the randomization inference approach to other analytical approaches for the stepped-wedge design. In Section 6, we apply randomization inference for stepped-wedge trials to a study of a community-based insurance program in rural Burkina Faso(Fink et al., 2013; Robyn et al., 2012). In Section 7, we provide a summary.

## 2.2 Notation and Set Up

There are $I$ clusters, $T$ calendar periods, and $n_{ij}$ individuals sampled from cluster $i$ during calendar period $j$. $N = \sum_i^I \sum_j^T n_{ij}$ is the total number of observations in the study design. Let $ijk$ index individual $k$ in cluster $i$ during calendar period $j$. An individual might be sampled at multiple time points; the indices $k = 1, \ldots, n_{ij}$ are time specific so that the same individual might have index $k$ and $k' \neq k$ at different times. During calendar period $j$, $m_j$ clusters are randomized to start treatment, where $m_1 + \cdots + m_T = I$, so that each cluster eventually starts treatment. $m_1, m_2, \ldots, m_T$ are prespecified before the start of the trial. Let $Z_{ij}$ be the treatment corresponding to cluster $i$ during calendar period $j$, where $Z_{ij} = 1$ for the active treatment and 0 for the control. Since the trial is cluster-randomized, we index the treatment status for clusters rather than individuals. Let $\mathbf{Z}$ be the vector of all treatment assignments, $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{IT})$. Write $\Omega$ for the set containing $|\Omega| = \binom{I}{m_1,\ldots,m_T}$ possible values $\mathbf{z}$ of $\mathbf{Z}$. Let $Y_{ijk}$ be the observed response and $\mathbf{Y}$ be the vector of all observed responses, $\mathbf{Y} = (Y_{111}, Y_{112}, \ldots, Y_{ITn_{IT}})$. In case of a possible lag between the time of treatment assignments and the time that responses are observed, we assume that if individual $k$ in cluster $i$ enters the trial during calendar period $j$, so is assigned treatment $Z_{ij}$, then that individual will continue to receive treatment $Z_{ij}$ until response $Y_{ijk}$ is recorded. Each individual has a (row) vector of pretreatment covariates $\mathbf{X}_{ijk}$. $\mathbf{X}$ is the matrix whose rows are $\mathbf{X}_{ijk}$.

We define the causal effect of interest under the potential outcomes framework. We extend the notation of Neyman (1990) and Rubin (1974a) by representing each potential outcome as a function of the vector of all treatment assignments $\mathbf{z}$ (Rosenbaum, 2007). Write $Y_{ijk}^{(\mathbf{z})}$ the response that the $k$th individual in cluster $i$ during calendar period $j$ would have if the treatment assignment $\mathbf{Z} = \mathbf{z}$ for $\mathbf{z} \in \Omega$. $Y_{ijk}^{(\mathbf{z})}$ indicates that each individual has $|\Omega|$ possible outcomes, only one of which is observed,

namely $Y_{ijk}^{(\mathbf{Z})}$. Fisher's sharp null hypothesis of no-treatment effect says that every unit would exhibit the same response under all treatment assignments, $Y_{ijk}^{(\mathbf{z})} = Y_{ijk}^{(\mathbf{z}')}$ for all $\mathbf{z}, \mathbf{z}' \in \Omega$. Under the alternative hypothesis, observed outcomes may exhibit arbitrary dependence.

We let $\mathcal{F} = \langle \mathcal{Y}, \mathbf{X} \rangle$, where $\mathcal{Y}$ is the unobserved array with $N$ rows and $|\Omega|$ columns having entries $Y_{ijk}^{(\mathbf{z})}$. $\mathcal{F}$ does not change as the treatment assignments, $\mathbf{Z}$, change, whereas $\mathbf{Y}$ is a function of $\mathcal{F}$ and $\mathbf{Z}$ so may change with $\mathbf{Z}$. To employ the cluster-randomized inference, as shown in Section 4, we assume the following assumptions hold for $\mathcal{F}$:

Assumption I: (a) there are no hidden variations of treatments and (b) $Y_{ijk}^{(\mathbf{z})} = Y_{ijk}^{(\mathbf{z}')}$ whenever $z_{ij} = z'_{ij}$. Assumption I(a) is part of the Stable Unit Treatment Value Assumption (Rubin, 1980; Imbens and Rubin, 2015) and says that an individual receiving level $\mathbf{z}$ of the treatment cannot receive different forms of the treatment which have different effects. The assumption is implicit in the notation $Y_{ijk}^{(\mathbf{z})}$ which says that there is a single potential outcome for level $\mathbf{z}$ of the treatment. Assumption I(b) asserts that the potential outcomes would not be affected by treatment assignments in other clusters or subjects in different clusters do not interfere. Note that this assumption still allows for the possibility that units within a cluster at a given time interfere with each other. Assumption I(b) can be seen as a relaxation of the usual no interference part of the stable unit treatment value assumption (SUTVA) in the sense that a group of concentrated individuals are allowed to interfere with each other at a given time but interference is not allowed between groups or time points. This assumption also implies no carry-over effect, that is, a previous treatment for one subject does not affect later responses of this same subject and also treatments for other subjects in the same cluster at previous times do not affect the response of the given subject at this time.

Assumption II: $Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}) = \frac{1}{|\Omega|} = \frac{1}{\binom{I}{m_1,\ldots,m_T}}$. This assumption says that the clusters are randomly assigned as to when to start treatment according to the stepped-wedge design and the conditional distribution of treatment assignments given the potential responses and covariates is a fixed known constant.This assumption guarantees that tests derived solely from the randomization have the correct level whether or not potential responses within the same cluster are subject to interference (Fisher, 1935; Welch, 1937).

Assumption III: If $\mathbf{z}$ and $\mathbf{z}'$ are the same except that $z_{ij} = 1$ while $z'_{ij} = 0$, then $Y_{ijk}^{(\mathbf{z})} - Y_{ijk}^{(\mathbf{z}')} = \theta$. This assumption implies that the treatment effect is constant across population and over time. By removing the treatment effect from the whole cluster during a calendar period, the observed responses would be the same as if there were no treatments assigned. This constant effect $\theta$ is the causal effect of interest.

## 2.3 The Importance of Cluster-by-Time Interactions

To motivate the need for accounting for cluster-by-time interactions, we assume that $Y_{ijk}$ is generated by the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + Z_{ij}\theta + e_{ijk} \tag{2.2}$$

For simplicity, we assume the $e_{ijk}$ are independent but correlation among the $e_{ijk}$ (as might arise if individuals are observed multiple times) can be accommodated.

Both models (1.1) and (3.1) are observed data models that are consistent with Assumptions I and II. Compared to the model (2.1), the model (2.2) has an additional term $\gamma_{ij}$ that accounts for cluster-by-time interactions. $\gamma_{ij}$'s are assumed to be iid $N(0, \eta^2)$ and independent of $\alpha$ and $e$. Using matrix notation, the model (2.2) can be

rewritten as

$$Y \sim N(M\Gamma, \Sigma = \sigma^2 I + \tau^2 A + \eta^2 B) \tag{2.3}$$

where $\mathbf{Y} = (Y_{111}, \ldots, Y_{112}, \ldots, Y_{ITn_{IT}})$, $\Gamma = (\mu, \beta_1, \ldots, \beta_T, \theta)^T$, and $M$ is the $N \times (T+2)$ design matrix. Let $Y_p$ denote the $p$th element in the vector $\mathbf{Y}$ which corresponds to a value of $ijk$, then $M_{pq} = 1$ if (1) $q = 1$ or (2) $2 \leq j \leq T+1$ and $Y_p$ is observed during calendar period $q-1$ or (3) $q = T+2$ and $Y_p$ is both observed and treated. $M_{pq} = 0$ otherwise. $A$ and $B$ are symmetric positive definite matrices corresponding to cluster and cluster-by-time interactions, respectively:

$$A = diag(\mathbf{1}_{n_{1\cdot}}\mathbf{1}_{n_{1\cdot}}^T, \ldots, \mathbf{1}_{n_{I\cdot}}\mathbf{1}_{n_{I\cdot}}^T) \tag{2.4}$$

$$B = diag(\mathbf{1}_{n_{11}}\mathbf{1}_{n_{11}}^T, \mathbf{1}_{n_{12}}\mathbf{1}_{n_{12}}^T, \ldots, \mathbf{1}_{n_{IT}}\mathbf{1}_{n_{IT}}^T) \tag{2.5}$$

where $\mathbf{1}_{n1}$ denote a column vector of 1's with length $n_1$ and $n_{i\cdot} = \sum_{j=1}^{T} n_{ij}$ is the size of cluster $i$.

Given $\sigma^2, \tau^2, \eta^2$, the covariance matrix $\Sigma$ is known. The best linear unbiased estimator of $\Gamma$ is the Generalized Least Squares (GLS) estimator, which asymptotically has a normal distribution.

$$\hat{\Gamma}_{GLS} = (M'\Sigma^{-1}M)^{-1}M'\Sigma^{-1}Y \tag{2.6}$$

$$\hat{\Gamma}_{GLS} \xrightarrow{d} N(\Gamma, (M'\Sigma^{-1}M)^{-1}) \tag{2.7}$$

If $\sigma^2, \tau^2, \eta^2$ are not known, an implementable version of the GLS estimator is the Feasible Generalized Least Squares (FGLS) estimator, which requires a consistent estimate of $\Sigma$, say $\hat{\Sigma}$.

$$\hat{\Gamma}_{FGLS} = (M'\hat{\Sigma}^{-1}M)^{-1}M'\hat{\Sigma}^{-1}Y \tag{2.8}$$

One common strategy to find a consistent estimate $\hat{\Sigma}$ is to start by finding $\hat{\Gamma}_{OLS}$ or

another consistent (but inefficient) estimator, take the residuals from OLS to build a consistent estimator of the error covariance matrix $\Sigma$, update the FGLS estimation, and then apply the same idea iteratively until the estimators vary less than some tolerance. Under regularity conditions, such a FGLS estimator has the same asymptotic distribution as a GLS estimator.

$$\hat{\Gamma}_{FGLS} \xrightarrow{d} N(\Gamma, (M'\Sigma^{-1}M)^{-1}) \tag{2.9}$$

For finite samples, the estimated covariance matrix of $\hat{\Gamma}_{FGLS}$ is

$$\widehat{Var}[\hat{\Gamma}] = (M'\hat{\Sigma}^{-1}M)^{-1} \tag{2.10}$$

which converges to the asymptotic covariance matrix $(M'\Sigma^{-1}M)^{-1}$ given that $\hat{\Sigma}$ converges to $\Sigma$ (Greene, 2003).

However, it is not always the case that we can find a consistent estimator of the covariance matrix $\Sigma$. The convergence of $\hat{\Sigma}$ to $\Sigma$ relies on the correct specification of matrix structure and normality assumptions (Jacqmin-Gadda et al., 2007). In the process of iteratively computing $\hat{\Sigma}$, any deviation from the correct model would lead to an inconsistent version of $\hat{\Sigma}$. In particular, if we failed to account for cluster-by-time interactions in the case of stepped-wedge cluster-randomized trials, we would specify the structure of covariance matrix in a different form from the actual covariance matrix, i.e., we would assume the consistent estimate of $\Sigma$ to be $\hat{\Sigma} = \hat{\sigma}^2 I + \hat{\tau}^2 A$ while the actual covariance matrix is in the form of $\Sigma = \sigma^2 I + \tau^2 A + \eta^2 B$. Since $B$ is a positive definite matrix as defined in 2.5, no values of $\hat{\sigma}^2$ and $\hat{\tau}^2$ would satisfy the equation $\hat{\sigma}^2 I + \hat{\tau}^2 A = \sigma^2 I + \tau^2 A + \eta^2 B$. Consequently, any computed $\hat{\Sigma}$ would be inconsistent, even if it maximizes the likelihood. Therefore, inferences based on $\hat{\Sigma}$ using the asymptotic distribution would be invalid.

We use a simulation study to examine this difference between the estimated variance of the treatment effect, which is the last diagonal element of $\hat{\Sigma}$ and the Monte Carlo simulation of the true variance, which is the last diagonal element of $\Sigma$. R code for the simulation is available in supplementary materials.

In the simulation, $I$ and $T$ are set to be 30 and 4, respectively. All clusters start with control at T = 1 and during each calendar period starting from T =2, 10 clusters in the control group are randomly selected to be assigned to treatment. All clusters have equal size 100 and the true treatment effect $\theta = 0$. The magnitude of clustering is calibrated by the intracluster correlation coefficient (ICC), which is the proportion of the total variation explained by the respective blocking factor. In particular, the correlation between two randomly selected observations in the same cluster is:

$$ICC_I = \frac{\tau^2}{\tau^2 + \eta^2 + \sigma^2} \tag{2.11}$$

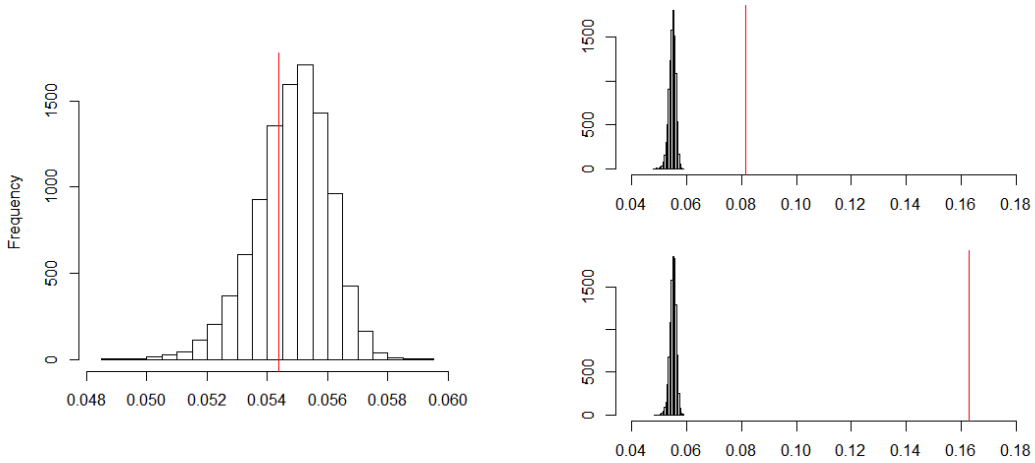
The correlation between two randomly selected observations in the same cluster and during the same calendar period is:

$$ICC_{IT} = \frac{\tau^2 + \eta^2}{\tau^2 + \eta^2 + \sigma^2} \tag{2.12}$$

As a result, the magnitude of interaction can be calibrated by $ICC_{IT} - ICC_I = \frac{\eta^2}{\tau^2+\eta^2+\sigma^2}$, which is the extra correlation from the same cluster and calendar period compared to just the cluster.

In Figure 2.2, we compare the distribution of estimated variances $\widehat{Var}[\hat{\theta}]$ over 10000 simulations with the Monte Carlo simulation of the true variance. When there are no cluster-by-time interactions, i.e., the model (2.1) is correctly specified, the left plot in Figure 2.2 indicates that the distribution of $\widehat{Var}[\hat{\theta}]$ is centered around the true variance, marked by the red vertical line. However, when interactions do exist,

the estimated variances are far off the true variance. The right plot describes two scenarios with different magnitudes of interactions. Neither of the distributions is close to the true variance.




(a) No cluster-by-time interactions

(b) With cluster-by-time Interactions

Figure 2.2: Comparison of estimated and true variances of the treatment effect in different settings of cluster-by-time interactions. In (a), there are no cluster-by-time interactions, $ICC_I := \frac{\tau^2}{\tau^2+\sigma^2} = 0.02$. In (b), there are cluster-by-time interactions, $ICC_I := \frac{\tau^2}{\tau^2+\sigma^2} = 0.02$, $ICC_{IT} := \frac{\tau^2+\eta^2}{\tau^2+\eta^2+\sigma^2} = 0.025$(upper) and 0.04 (lower).

Table 2.1 gives a more detailed summary of the estimated treatment effect $\hat{\theta}$ given by the FGLS estimator when the cluster-by-time interactions are not included in the model. As shown by column $E[\hat{\theta}]$, $\hat{\theta}$ is consistent in all settings. When there are no cluster-by-time interactions as shown by the first two rows, the average of the estimated variances $E(\widehat{Var}[\hat{\theta}])$ is almost the same as the Monte Carlo simulation of the true variance $Var[\hat{\theta}]$. But this is not the case when the interaction term $\gamma$ is nonzero. The last column $SD(\widehat{Var}[\hat{\theta}])$ describes the dispersion of the estimated variances, which is of a much smaller order than its average.

| dim(M) | $\alpha$ | $\gamma$ | $\epsilon$ | $E[\hat{\theta}]$ | $Var[\hat{\theta}]$ | $E(\widehat{Var}[\hat{\theta}])$ | $SD(\widehat{Var}[\hat{\theta}])$ |
|---|---|---|---|---|---|---|---|
| | N(0,1) | Zero | N(0,49) | -0.0031 | 0.0544 | 0.0548 | 0.0012 |
| $I$=30 | N(0,1) | Zero | $7/\sqrt{3}t(3)$ | -3.4$e^{-5}$ | 0.0544 | 0.0545 | 0.0080 |
| | N(0,1) | N(0,.25) | N(0,48.75) | -0.0008 | 0.0816 | 0.0549 | 0.0012 |
| $T$=4 | N(0,1) | N(0,.5) | N(0,48.5) | -0.0008 | 0.1083 | 0.0550 | 0.0012 |
| | N(0,1) | N(0,1) | N(0,48) | -0.0008 | 0.1626 | 0.0552 | 0.0011 |

Table 2.1: Properties of the estimated treatment effect given by the Feasible Generalized Least Square estimator

The above simulation results show that fitting a linear mixed model for the stepped-wedge design while ignoring cluster-by-time interactions can lead to severely wrong standard errors, and this leads to poor control of Type I error rate, as shown by Table 2.2.

| dim(M) | $\alpha$ | $\gamma$ | $\epsilon$ | $ICC_I$ | $ICC_{IT}$ | Type I error |
|---|---|---|---|---|---|---|
| | N(0,1) | Zero | N(0,49) | 0.02 | 0.02 | 0.052 |
| $I$= 30 | N(0,1) | Zero | $7/\sqrt{3}t(3)$ | 0.02 | 0.02 | 0.054 |
| | N(0,1) | N(0,.25) | N(0,48.75) | 0.02 | 0.025 | 0.511 |
| $T$=4 | N(0,1) | N(0,.5) | N(0,48.5) | 0.02 | 0.03 | 0.658 |
| | N(0,1) | N(0,1) | N(0,48) | 0.02 | 0.04 | 0.756 |

Table 2.2: Type I error of linear mixed models not accounting for cluster-by-time interactions

## 2.4 Randomization Inference for Stepped-Wedge Cluster-Randomized Trials

We would like to develop a strategy that accounts for cluster-by-time interactions if they exist. We will consider randomization inference. In randomization inference as developed by Fisher (1935), hypotheses are tested using only the assumption that the randomization has been properly carried out and randomization inference provides exact, distribution-free inferences. The significance level is always guaranteed

regardless of the underlying mechanism that generates the data.

## A General Setup

There are $I$ clusters and $T$ calendar periods. At time $t$, $m_t$ clusters are randomized to start treatment, where $m_1 + \cdots + m_T = I$, so that each cluster eventually starts treatment. Collect all possible values $\mathbf{z}$ of the treatment assignments $\mathbf{Z}$ in a set $\Omega$, $|\Omega| = \binom{I}{m_1,\ldots,m_T}$. Because random numbers are used to assign which clusters start treatment at which times, $P(\mathbf{Z} = \mathbf{z}) = 1/|\Omega|$ for each $\mathbf{z} \in \Omega$.

Let $\mathbf{e}$ be a function of $\mathcal{F} = \langle \mathcal{Y}, \mathbf{X} \rangle$ and let $t(\mathbf{Z}, \mathbf{e}, \mathbf{X})$ be any function of $\mathbf{Z}, \mathbf{e}, \mathbf{X}$. Because $\mathbf{e}$ and $\mathbf{X}$ are functions of $\mathcal{F}$ and randomization ensures $P(\mathbf{Z}|\mathcal{F}) = 1/|\Omega|$, it follows that for all $v$,

$$P(t(\mathbf{Z}, \mathbf{e}, \mathbf{X}) \geq v|\mathcal{F}) = \frac{|\{\mathbf{z} \in \Omega : t(\mathbf{z}, \mathbf{e}, \mathbf{X}) \geq v\}|}{|\Omega|}, \tag{2.13}$$

which is the randomization distribution of $t(\mathbf{Z}, \mathbf{e}, \mathbf{X})$. In words, given $\mathcal{F}$, the chance that $t(\mathbf{Z}, \mathbf{e}, \mathbf{X}) \geq v$ is simply the proportion of treatment assignments $\mathbf{z} \in \Omega$ such that $t(\mathbf{Z}, \mathbf{e}, \mathbf{X}) \geq v$. Moreover, (2.13) is the distribution of $t(\mathbf{Z}, \mathbf{e}, \mathbf{X})$ given $\mathcal{F}$ no matter what process produced $\mathcal{F}$. Fisher's (1935) description of randomization inference as the "reasoned basis for inference" refers to the fact that randomization creates the distribution (2.13) for every function $\mathbf{e}$ of $\mathcal{F}$ without further assumptions.

## Test of No Effect

The sharp null hypothesis of no effect asserts that the response of each individual is unchanged by receiving the treatment, $H_0 : \forall \mathbf{z}, \mathbf{z}' \in \Omega, Y_{ijk}^{(\mathbf{z})} = Y_{ijk}^{(\mathbf{z}')}$ for $i = 1, \ldots, I$, $t = 1, \ldots, T$, $k = 1, \ldots, n_{it}$, i.e., $\mathbf{Y}^{(\mathbf{z})} = \mathbf{Y}^{(\mathbf{z}')}$. If $H_0$ were true, randomization would label clusters treated or control but the observed outcomes would be unchanged. If

$H_0$ were true, the observed response $\mathbf{Y^{(Z)}}$ would equal $\mathbf{Y^{(0)}}$, a special case where all clusters are under control. Thus, under the null hypothesis of no treatment effect, the randomization distribution of $t(\mathbf{Z}, \mathbf{Y}, \mathbf{X}) = t(\mathbf{Z}, \mathbf{Y^{(0)}}, \mathbf{X})$ would be given by (2.13) with $\mathbf{e} = \mathbf{Y^{(0)}}$, where both $t(\mathbf{Z}, \mathbf{Y^{(0)}}, \mathbf{X})$ and its null distribution (2.13) would be calculated from the observed data when $H_0$ were true. For instance, in completely randomized experiments, Welch (1937) tested the null hypothesis of no effect using the randomization distribution of a test statistic suggested by analysis of variance and Raz (1990) used the randomization distribution of a test statistic that adjusted for $\mathbf{X}$ using a data smoother.

## Test of Constant Treatment Effect

The above method can be directly extended to test for constant treatment effect

$$H_0 : \theta = \theta_0 \quad v.s. \quad H_1 : \theta = \theta_1$$

Under the null hypothesis of $\theta = \theta_0$, $\mathbf{Y^{(0)}} = \mathbf{Y} - \mathbf{Z}\theta_0$. If $\mathbf{e} = \mathbf{Y^{(0)}}$, $t(\mathbf{Z}, \mathbf{e}, \mathbf{X}) = t(\mathbf{Z}, \mathbf{Y} - \mathbf{Z}\theta_0, \mathbf{X}) = t'(\mathbf{Z}, \mathbf{Y}, \mathbf{X})$, where $t'$ is a function on $\mathbf{Z}, \mathbf{Y}, \mathbf{X}$. This is to say $t'(\mathbf{Z}, \mathbf{Y}, \mathbf{X})$ would also have the randomization distribution given by (2.13).

Because of the randomization procedure, any function $t$ on $\mathbf{Z}, \mathbf{Y^{(0)}}, \mathbf{X}$ is a valid test statistic with the Type I error controlled by the prespecified significance level. However, it does not mean that any arbitrarily chosen $t$ is a good test statistic. We need to consider power. In the next section, we will discuss test statistics suitable for stepped-wedge cluster-randomized experiments.

## Wald Randomization Test

A natural choice of $t$ is the Wald statistic based on the maximum likelihood estimation of the treatment effect under the model (2.1) or (2.2). Under the null hypothesis $H_0$ : $\theta = \theta_0$, $\mathbf{Y} - \mathbf{Z}\theta_0 = \mathbf{Y^{(0)}} = \mathbf{e}$. The maximum likelihood estimator of $L(\theta|\mathbf{Z}, \mathbf{Y}, \mathbf{X}) = L(\theta|\mathbf{Z}, \mathbf{e} + \mathbf{Z}\theta_0, \mathbf{X})$ is a function on $\mathbf{Z}$, $\mathbf{e}$, and $\mathbf{X}$. $t(\mathbf{Z}, \mathbf{e}, \mathbf{X})$ can be chosen as the Wald statistic of the null hypothesis $H_0 : \theta = \theta_0$ over the alternate hypothesis $H_1 : \theta \neq \theta_0$

$$t(\mathbf{Z}, \mathbf{e}, \mathbf{X}) = \frac{(\hat{\theta} - \theta_0)^2}{\widehat{Var}(\hat{\theta})} \tag{2.14}$$

Instead of using its asymptotic distribution, which is a $\chi^2$ distribution under the null hypothesis, the level is calculated using the randomization distribution given by (2.13). We can also investigate the power by randomly generating numerous data sets under a pre-specific alternative hypothesis. For each of these data sets, randomization inference is carried out and the evidence for or against the null hypothesis is recorded.

The Wald randomization test is applicable to a wide range of parametric models corresponding to different distributions of observed outcomes and can be implemented using standard functions in R, such as lmer() in the *lme4* package for linear mixed models, glm() for generalized linear models, and censReg() in the *censReg* package for censored regression models.

## Other Randomization Tests

Instead of calculating the maximum likelihood estimate and its standard deviation, other test statistics are available for stepped-wedge cluster-randomized trials. For example, because the design is essentially a two-way layout, we can first eliminate row and column effects by estimating their values or using the median polish method if robustness is a concern (Hoaglin et al., 1983). We then carry out the aligned rank test

to compare the adjusted responses between clusters with different interventions (Sen, 1968). If responses have heavy-tailed distributions, we may consider test statistics involving ranks to avoid bias caused by extreme values.

### Covariates Adjustment

The discussion in Sections 4.2 and 4.3 make no use of the covariates, $X$, but it is straightforward to incorporate them, with no change in the logic, see Rosenbaum (a). Instead of letting $\mathbf{e} = \mathbf{Y^{(0)}}$, $\mathbf{e}$ could also be a function on $X$. For example, $\mathbf{e}$ could be residuals when $\mathbf{Y^{(0)}}$is regressed on $X$ by any method of regression. The randomization distribution of $t(\mathbf{Z}, \mathbf{e}, \mathbf{X})$ would still be given by (2.13).

## 2.5 Simulation Study

We use a simulation study to investigate the level and the power of the Wald test statistic with usual asymptotic inference and with randomization inference in the stepped-wedge design. For demonstration purpose, we assume responses are normal and continuous. In all simulation settings, $I = 30$ and $T = 4$. When $t = 0$, all clusters are in the control group. When $t = 1$, 10 out of 30 clusters are randomly selected to receive treatment. When $t = 2$, 10 out of the remaining 20 untreated clusters are randomly selected to receive treatment. When $t = 3$, all clusters are assigned to treatment. Cluster sizes are randomly sampled between 1000 and 2000 and fixed over time. The true treatment effect $\theta_0$ is set to be 0 and the power is calculated under the alternative $\theta_1 = .25, .5, 1, 1.5, 2$. $ICC_I = \frac{\tau^2}{\tau^2+\eta^2+\sigma^2}$ is the intracluster correlation coefficient corresponding to clusters. $ICC_{IT} = \frac{\tau^2+\eta^2}{\tau^2+\eta^2+\sigma^2}$ is the intracluster correlation coefficient corresponding to both clusters and interactions. All numbers reported are the average over 1000 sets of randomly simulated data set.

We first examine the Type I error rate in several scenarios. $Wald_{asy}$ and $Wald_{rand}$ are obtained under the model (2.1) with usual asymptotic inference and with randomization inference. $Wald^*_{asy}$ and $Wald^*_{rand}$ are obtained under the model (2.2) with usual asymptotic inference and with randomization inference.

| $\alpha$ | $\gamma$ | $e$ | $Wald_{asy}$ | $Wald_{rand}$ | $Wald^*_{asy}$ | $Wald^*_{rand}$ |
|---|---|---|---|---|---|---|
| N(0,1) | Zero | N(0,49) | .045 | .061 | .044 | .061 |
| N(0,1) | Zero | $7/\sqrt{3}t(3)$ | .042 | .055 | .042 | .054 |
| N(0,1) | Zero | Cauchy | .055 | .050 | .051 | .053 |
| N(0,1) | N(0,1) | N(0,48) | .315 | .055 | .069 | .059 |
| N(0,1) | N(0,1) | $4\sqrt{3}t(3)$ | .342 | .051 | .069 | .054 |
| N(0,1) | N(0,1) | Cauchy | .056 | .054 | .063 | .060 |

Table 2.3: Type I error rate of the Wald test statistic based on the asymptotic distribution and the randomization distribution

It can be seen from Table 2.3 that both randomization procedures $Wald_{rand}$ and $Wald^*_{rand}$ guarantee the correct Type I error rate in all settings. When the interaction $\gamma$ is zero, the Type I error rate is well-controlled by both tests with usual asymptotic inference. However, when $\gamma$ has a stand normal distribution, which leads to a small intracluster correlation coefficient $ICC_{IT} = 0.04$, the Type I error rate given by $Wald_{asy}$ is inflated to .315 and .342 when $e$ follows a normal and a t distributions, respectively. The $Wald^*_{asy}$ test performs better than $Wald_{asy}$ as it incorporates cluster-by-time interactions, but its Type I error rate is still slightly higher than its randomized version. Such phenomenon disappears when $e$ follows a Cauchy distribution. This might be explained by the fact that the Cauchy distribution is so heavy-tailed that it dominates the small interaction term $\gamma$.

We next examine power. According to results in Table 2.4, when there are no cluster-by-time interactions, the randomization tests have comparable power with the tests using the asymptotic distribution. When there are cluster-by-time interactions, we ignore the power calculated from $Wald_{asy}$ and $Wald^*_{asy}$ as the level is no longer

valid, but only focus on their randomized versions, which give sufficient power to detect wrong values of the treatment effect.

| $\theta_1$ | $\alpha$ | $\gamma$ | $e$ | $ICC_I$ | $ICC_{IT}$ | $Wald_{asy}$ | $Wald_{rand}$ | $Wald^*_{asy}$ | $Wald^*_{rand}$ |
|---|---|---|---|---|---|---|---|---|---|
| .25 | N(0,1) | Zero | N(0,49) | .02 | .02 | . 254 | .275 | .254 | .276 |
| .5 | N(0,1) | Zero | N(0,49) | .02 | .02 | .723 | .715 | .721 | .725 |
| 1 | N(0,1) | Zero | N(0,49) | .02 | .02 | .999 | .999 | .999 | .999 |
| 1.5 | N(0,1) | Zero | N(0,49) | .02 | .02 | 1 | 1 | 1 | 1 |
| 2 | N(0,1) | Zero | N(0,49) | .02 | .02 | 1 | 1 | 1 | 1 |
| .25 | N(0,1) | N(0,1) | N(0,48) | .02 | .04 | . 427 | .096 | .136 | .108 |
| .5 | N(0,1) | N(0,1) | N(0,48) | .02 | .04 | .636 | .253 | .335 | .277 |
| 1 | N(0,1) | N(0,1) | N(0,48) | .02 | .04 | .941 | .726 | .798 | .752 |
| 1.5 | N(0,1) | N(0,1) | N(0,48) | .02 | .04 | .999 | .969 | .982 | .975 |
| 2 | N(0,1) | N(0,1) | N(0,48) | .02 | .04 | 1 | 1 | 1 | 1 |
| .25 | N(0,1) | Zero | $7/\sqrt{3}t(3)$ | .02 | .02 | .266 | .272 | .261 | .280 |
| .5 | N(0,1) | Zero | $7/\sqrt{3}t(3)$ | .02 | .02 | .751 | .734 | .740 | .744 |
| 1 | N(0,1) | Zero | $7/\sqrt{3}t(3)$ | .02 | .02 | .998 | .999 | .998 | .999 |
| 1.5 | N(0,1) | Zero | $7/\sqrt{3}t(3)$ | .02 | .02 | 1 | 1 | 1 | 1 |
| 2 | N(0,1) | Zero | $7/\sqrt{3}t(3)$ | .02 | .02 | 1 | 1 | 1 | 1 |
| .25 | N(0,1) | N(0,1) | $4\sqrt{3}t(3)$ | .02 | .04 | .416 | .107 | .124 | .115 |
| .5 | N(0,1) | N(0,1) | $4\sqrt{3}t(3)$ | .02 | .04 | .630 | .240 | .310 | .272 |
| 1 | N(0,1) | N(0,1) | $4\sqrt{3}t(3)$ | .02 | .04 | .942 | .718 | .786 | .786 |
| 1.5 | N(0,1) | N(0,1) | $4\sqrt{3}t(3)$ | .02 | .04 | .999 | .971 | .999 | .992 |
| 2 | N(0,1) | N(0,1) | $4\sqrt{3}t(3)$ | .02 | .04 | 1 | 1 | 1 | 1 |

Table 2.4: Power of the Wald test statistic for linear mixed models based on the asymptotic distribution and the randomization distribution

We also carry out a set of simulations for dichotomous outcomes, according to the model

$$logit(E(Y_{ijk})) = \mu + \alpha_i + \beta_j + \gamma_{ij} + Z_{ij}\theta \tag{2.15}$$

Results are summarized in Table 2.5, showing similar advantages of using randomization inference for the stepped-wedge cluster-randomized trials.

| $\theta_1$ | $\alpha$ | $\gamma$ | $e$ | $ICC_I$ | $ICC_{IT}$ | $Wald_{asy}$ | $Wald_{rand}$ | $Wald^*_{asy}$ | $Wald^*_{rand}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | N(0,1) | Zero | N(0,49) | .02 | .02 | .043 | .051 | .044 | .051 |
| .5 | N(0,1) | Zero | N(0,49) | .02 | .02 | .223 | .208 | .216 | .195 |
| 1 | N(0,1) | Zero | N(0,49) | .02 | .02 | .791 | .747 | .773 | .740 |
| 1.5 | N(0,1) | Zero | N(0,49) | .02 | .02 | 1 | .999 | .998 | .998 |
| | | | | | | | | | |
| 0 | N(0,1) | N(0,1) | N(0,48) | .02 | .04 | .217 | .060 | .091 | .048 |
| .5 | N(0,1) | N(0,1) | N(0,48) | .02 | .04 | .412 | .172 | .318 | .159 |
| 1 | N(0,1) | N(0,1) | N(0,48) | .02 | .04 | .726 | .448 | .544 | .377 |
| 1.5 | N(0,1) | N(0,1) | N(0,48) | .02 | .04 | .922 | .681 | .837 | .572 |

Table 2.5: Power of the Wald test statistic for generalized linear mixed models based on the asymptotic distribution and the randomization distribution

## 2.6 Application to Study of Community-Based Health Insurance Program

### Background

The Ministry of Health and Nouna Health Research Center in Nouna District, Burkina Faso implemented a CBHI scheme from 2004 to 2006 that aimed to make health care more affordable and to protect local communities from large health expenditure shocks (Fink et al., 2013; Robyn et al., 2012). To allow for a proper evaluation, the rollout of the program followed a stepped-wedge cluster-randomized design, enrolling randomly selected communities in three phases. In order to investigate the effect of CBHI schemes on household welfare, we follow Fink et al. (2013) to analyze the effect of CBHI schemes on catastrophic expenditure.

### 2.6.1 Data

The data we use is the Nouna Health and Demographic Surveillance Site (HDSS) survey data collected from 2003 to 2008. Data from year 2003 are the baseline prior to the intervention and data from years 2007 and 2008 are controls after the final rollout phase. There are 48 areas in the health district and each of them is considered a cluster. Due to residential mobility and migration, the study population is dynamic with an attrition rate of 59% from 2003 to 2008. There are 59,905 records in total and the number of individuals targeted by the insurance program in phase I, II, and III are 27,696, 14,292, and 17,917, respectively. Equal mean test indicates that these three rollout groups have balanced covariates of age, gender, years of education, literacy, religion, marital status, household size, and wealth index, see Table 4 in Fink et al. (2013).

Since the primary objective of CBHI schemes is to protect individuals against large financial shocks, we investigate the probabilities of facing health expenditure greater than 5%, 10%, 15%, 25% and 50% of monthly income. The catastrophic expenditure is a dichotomous outcome, which is coded as one if the total health expenditure is greater than a certain percentage of the monthly income. For example, in the 2003 data suggest that about 10.4% of individuals faced health expenditure larger than 5% of their monthly income in the sample, and 2.7% of individuals had to cover health expenditure of more than half their monthly income. See Table 2.6 for a detailed year-by-year summary of the data.

| Year | Population size | Expenditure Cutoff | | | | |
|---|---|---|---|---|---|---|
| | | $\geq$ 5% | $\geq$10% | $\geq$15% | $\geq$25% | $\geq$50% |
| 2003 | 7796 | 814 | 610 | 460 | 347 | 207 |
| 2004 | 8619 | 1037 | 716 | 577 | 361 | 191 |
| 2005 | 6875 | 1402 | 977 | 742 | 519 | 311 |
| 2006 | 10712 | 925 | 576 | 481 | 306 | 224 |
| 2007 | 13784 | 1316 | 939 | 690 | 377 | 211 |
| 2008 | 12118 | 950 | 663 | 452 | 291 | 141 |

Table 2.6: Distribution of catastrophic expenditure over time, Nouna HDSS Household Survey, 2003 – 2008.

## Model

The models (2.1) and (2.2) assume the continuity of observed responses and the normality of random components. In our data, catastrophic health expenditure is binary so we use the generalized linear mixed model and then apply the Wald randomization test. In particular, we use $P_{ijk}$ to denote the probability of facing catastrophic expenditure for individual $k$ during calendar period $j$ from cluster $i$, the observed response $Y_{ijk}$ follows the model

$$\begin{aligned} Y_{ijk} &\sim \text{Bernoulli}(P_{ijk}) \\ logit(P_{ijk}) &= \mu + \alpha_i + \beta_j + Z_{ij}\theta + X_{ijk}^T\gamma + e_{ijk} \end{aligned} \tag{2.16}$$

where $\alpha_i, \beta_j, Z_{ij}$, and $\theta$ are defined the same as in the model (2.1) and $X_{ijk}$ is a vector of covariates that we adjust for, which are age, gender, years of education, literacy, religion, marital status, household size, and wealth index. Because we have repeated observations on people and there might be unmeasured covariates not included in $X_{ijk}$, $e_{ijk}$ could be correlated for $j \in \{1, 2, \ldots, T\}$. As a result, we include person-

level random effects to allow for correlation between $e_{ijk}$ and $e_{ij'k}$.

## Results

We first investigate catastrophic expenditure that is greater than 5% of monthly income. We use the function lmer() from the package *lme4* to solve for the maximum likelihood estimate of $\theta$ in (2.16), which has mean value -0.3966 and standard deviation 0.0554. Hence, the Wald test statistic for the actual insurance rollout is 51.093 with p-value $< 0.001$, indicating that there is significant evidence that the CBHI insurance program helped to reduce the likelihood of facing health expenditure greater than 5% of monthly income. We then carry out the Wald randomization test by assuming that there was no such effect. The p-value given by (2.13) is 0.117, indicating that there is no strong evidence that insurance had an effect on the catastrophic expenditure. We also consider an expanded version of the model (2.16) that includes cluster-by-time interactions:

$$\begin{aligned} Y_{ijk} &\sim \text{Bernoulli}(P_{ijk}) \\ logit(P_{ijk}) &= \mu + \alpha_i + \beta_j + \gamma_{ij} + Z_{ij}\theta + X_{ijk}^T\gamma + e_{ijk} \end{aligned} \tag{2.17}$$

The Wald statistic based on this model for the actual insurance rollout is 2.229 with p-value 0.135 and the Wald randomization test gives p-value 0.115.

We repeat the same analysis for expenditure cutoffs 10%, 15%, 25%, and 50% and summarize results in Table 2.7. P-values in columns $Wald_{asy}$ and $Wald_{rand}$ are obtained under the model (2.16) with usual asymptotic inference and with randomization inference. P-values in columns $Wald^*_{asy}$ and $Wald^*_{rand}$ are obtained under the model (2.17) with usual asymptotic inference and with randomization inference.

| Expenditure cutoff | p-value | | | |
|---|---|---|---|---|
| | $Wald_{asy}$ | $Wald_{rand}$ | $Wald^{*}_{asy}$ | $Wald^{*}_{rand}$ |
| ≥5% | < .001 | .117 | .135 | .115 |
| ≥10% | < .001 | .339 | .351 | .331 |
| ≥15% | < .001 | .431 | .463 | .427 |
| ≥25% | < .001 | .442 | .422 | .410 |
| ≥50% | .009 | .041 | .014 | .038 |

Table 2.7: CBHI's impact on catastrophic health expenditure based on generalized linear mixed models

## Conclusion

Based on randomization inference that controls the Type I error rate properly, there is no strong evidence that the CBHI program carried out in Nouna District, Burkina Faso affected catastrophic expenditure that are defined to be greater than 5%, 10%, 15%, and 25% of monthly income. The CBHI program, however, conferred a large benefit to people facing extremely high health expenditure that exceeds half of their monthly income. We see discrepancy between results from the model (2.16) and the model (2.17) using asymptotic inference. The model (2.16) would conclude that the CBHI program substantially reduced the likelihood of all levels of catastrophic health expenditure, but the model (2.17) would conclude so only for the 50% cutoff.

Table 2.7 suggests that conclusions given by the asymptotic inference and the randomization inference are consistent only for the model (2.17), which is an indication of the presence of cluster-by-time interactions. If we failed to consider the cluster-by-time interactions, the standard asymptotic inference is likely to greatly overestimate the protective effects of the insurance program

## 2.7 Summary

There is a lack of literature on the theoretical aspects of analyzing the stepped-wedge cluster-randomized trials. We focus on statistical properties of the stepped-wedge design following the linear mixed model approach proposed by Hussey and Hughes (Hussey and Hughes, 2007). Our simulations raise a red flag about using model-based inference for stepped-wedge trials. Specifically, the results can be very sensitive to model misspecification. As a result, bias can be introduced by cluster-by-time interactions and any other violations of assumptions.

We thus propose a new approach to the analysis of stepped-wedge cluster-randomized trials – using randomization inference to test for constant interventions. We introduce a unified framework to develop the randomization distribution for any test statistic, which can be used to calculate p-values and construct confidence intervals. Simulations based on linear mixed models show that randomization inference always guarantees the valid Type I error rate and has power comparable to the usual asymptotic inference.

We demonstrate our method using the Burkina Faso CBHI dataset to investigate whether CBHI schemes protect individuals against large financial shocks. We conclude that the insurance had limited effects on reducing the likelihood of low to moderate levels of catastrophic health expenditure in the target areas, but substantially benefited people facing extremely high health expenditure that exceeds half of their monthly income.

We hope that this paper serves as a valuable contribution to the literature on statistical properties of stepped-wedge cluster-randomized trials and its practical implementation in health economics, education, public health and other fields in which cluster-randomized trials are of interest. Our goal in this paper is to emphasize the value of randomization inference for stepped-wedge cluster-randomized trials and pro-

vide methods for implementing such randomization inference. With a strong belief in a parametric model, one can make inferences and calculate power and sample size based on asymptotic distributions but these inferences can be sensitive to the model; randomization inference can deliver similar power while the inferences remain valid regardless of whether the parametric model holds or not.

# 3

# The Trend-in-trend Design for Causal Inference


## Abstract

Cohort studies can be biased by unmeasured confounding. We propose a hybrid ecologic-epidemiologic design called the trend-in-trend design, which requires a strong time-trend in exposure, but is unbiased unless there are unmeasured factors affecting outcome for which there are time-trends in prevalence that are correlated with time-trends in exposure across strata with different exposure trends. Thus, the conditions under which the trend-in-trend study is biased are a subset of those under which a cohort study is biased. The trend-in-trend design first divides the study population into strata based on the cumulative probability of exposure (CPE) given covariates, which effectively stratifies on time-trend in exposure, provided there is a trend. Next, a covariates-free maximum likelihood model estimates the odds ratio (OR) using data on exposure prevalence and outcome frequency within CPE strata, across multiple periods. In simulations, the trend-in-trend design produced ORs with negligible bias in the presence of unmeasured confounding. In empiric applications, trend-in-trend reproduced the known positive association between rofecoxib and myocardial infarction


*Joint work with Sean Hennessy, Charles Leonard, Dylan Small

(observed OR: 1.23, 95% confidence interval: 1.05, 1.44), and known null associations between rofecoxib and severe hypoglycemia [OR = 1.09 (0.92, 1.28)] and non-vertebral fracture [OR = 0.84 (0.64, 1.09)]. The trend-in-trend method may be useful in settings where there is a strong time-trend in exposure, such as a newly-approved drug or other medical intervention.

## 3.1 Introduction

Many important causal questions cannot be addressed through randomized trials because of ethical or practical reasons. Ecologic studies address causal questions by examining time trends in exposure and outcome, but can be biased by co-occurring trends in other factors affecting outcome (Cook et al., 1979). Epidemiologic designs such as the cohort study can be biased if there are unmeasured determinants of exposure that are associated with outcome (i.e., unmeasured confounders). In this paper, we introduce a novel hybrid ecologic-epidemiologic design called the trend-in-trend design. Rather than comparing exposed vs. unexposed persons, the trend-in-trend design examines time-trends in outcome as a function of time-trends in exposure across strata with different time-trends in exposure. Intuitively, in a population stratified on time-trends in exposure, an association between exposure time-trends and outcome time-trends across strata should provide evidence for causation unless there are unmeasured factors affecting outcome for which there are time-trends in prevalence that are correlated with time-trends in exposure across strata. Thus, the scenarios under which a trend-in-trend study is susceptible to unmeasured confounding should be a subset of those under which a cohort study is susceptible, making the trend-in-trend design more resistant to unmeasured confounding. The trade-offs are that a trend-in-trend study is feasible only when there is a strong time-trend in exposure, and a

trend-in-trend study should have less statistical precision than a cohort study.

While novel, the trend-in-trend design is related to two established econometric approaches. One is the difference-in-difference (DID) method (Lechner et al., 2011; Meirik, 2008), as both address unmeasured confounding by examining within-group changes and time-trends in outcome. However, unlike the DID method, the trend-in-trend design estimates an individual-level causal parameter. In particular, the trend-in-trend design yields the odds ratio (OR), which approximates the risk ratio when the outcome is rare (Viera, 2008). The trend-in-trend method is also related to the use of calendar time as an instrumental variable (IV) (Cain et al., 2009; Johnston et al., 2008), and in fact the two are equivalent if only a single stratum is used in the trend-in-trend design. However, use of calendar time as an IV can be biased by any time-trend in the prevalence of an unmeasured factor that affects outcome. In contrast, the trend-in-trend design is biased by such a trend only if the time-trend in the unmeasured factor is correlated with the time-trends in exposure across strata defined by factors associated with exposure. The trend-in-trend design therefore relaxes the assumptions under which a calendar time IV study is valid.

In this paper, we first introduce the cumulative probability of exposure (CPE), which is used to divide the population into strata with different exposure prevalences and thus different time-trends in exposure, provided that an overall time-trend exists. We then propose two reasonable models for individuals and subgroups respectively. Under the assumptions that the outcome is rare, covariates are either time-invariant or change randomly over time within person, and there are no time-trends in unmeasured causal factors that are associated with time-trends in exposure across strata, we give a mathematical derivation of the connection between individuals and subgroups and a method to estimate the OR using group-level data. We then show mathematically that this estimate is unbiased by both measured and unmeasured confounders. We

report a simulation study illustrating that the OR estimated using the trend-in-trend method is much less biased than that estimated using cohort methods when there is unmeasured confounding by factor with no trend in prevalence. Finally, we apply the trend-in-trend method to healthcare data to reproduce the known positive association between rofecoxib and acute myocardial infarction (AMI) (Jüni et al., 2004), and two presumably null associations: rofecoxib and severe hypoglycemia, and rofecoxib and bone fractur (Vestergaard et al., 2006).

## 3.2 Method and Models

### 3.2.1 Stratification based on the Cumulative Probability of Exposure (CPE)

The analysis of a trend-in-trend study involves two stages. In the first stage, we estimate the CPE, which is the predicted probability of exposure over the entire study period, based on variables other than exposure, outcome, and their potential effects. In particular, suppose we observe a population in which each individual's binary exposure status over the study period is observed. We also observe a set of variables that affect but are known from subject-area knowledge not to be affected by exposure, such as age, sex, geographic residence, diagnoses, etc. We fit a logistic regression model using these variables as independent variables, with the dependent variable being exposure. The fitted value is the estimated CPE. Since the unit of analysis for the CPE model is the individual, and covariates are treated as invariant, each subject will be in the same CPE stratum for all observation periods. If, analogously to a new user cohort study, subjects are required to be present for a baseline period prior to the first opportunity for exposure, then the values for all variables in the CPE model can be fixed at the first opportunity for exposure (e.g., drug approval). However,

many healthcare databases have high turnover rates, and restricting the study to persons with sufficient baseline period prior to the first opportunity for exposure may drastically reduce available sample size. In such a situation, one can allow the value of CPE variables that require time to ascertain (e.g., appearance of diagnoses) to be determined by data observed during the study period, provided that subject-area knowledge can rule out the possibility that exposure status affected any CPE variable. For an exposure with an overall time-trend in prevalence, intuition tells us that the magnitude of the trend should vary across strata defined by the CPE. The CPE is similar to the propensity score (Rosenbaum and Rubin, 1983), since both predict exposure, but differs from it in that the propensity score is used to balance observed covariates across exposure groups, while the CPE is used to identify strata with different time-trends in exposure. It may also be possible to directly model the trend itself rather than the CPE. The second stage analysis, described below, applies to any population stratified on time-trend in exposure prevalence.

### 3.2.2 Models in the Trend-in-Trend Design

To derive a quantitative estimate of a causal effect, we propose two models of outcomes. One model is defined for each subject at each time point to account for covariates heterogeneity across population and time trends of outcome. The other one is specified at the population level at each time point, which depicts the mean outcome among those subjects within the same subgroup. We assume that the study population consists of $N$ individuals and there are $T$ time periods. Let $X_i^t$ denote the vector of covariates associated $i$ with individual $t$ at time period , which represents intrinsic characteristics that might influence a particular exposure and/or outcome. $X_i^t$ can be either observed, unobserved, or partially observed. $X_i^t$ is assumed to follow a distribution $F$ across the population. $Z_i^t$ and $Y_i^t$ are exposure and outcome variables

for individual $i$ at time period $t$. $G$ is the index for CPE strata.

#### 3.2.2.1 Subject-Specific Model

The conditional expected outcomes are assumed to satisfy

$$h(\mu_i^t) = \beta_0 + \beta_1 Z_i^t + \beta_2 t + \gamma^T X_i^t \tag{3.1}$$

where $h$ is the link function. The subject-specific model is a special case of the generalized linear mixed model with exposure and time period being the fixed effects and the covariates for an individual (some of which may be unobserved) represented as random effects (Zeger et al., 1988). Because the trend-in-trend design is intended to estimate the instantaneous risk of an exposure, only $Z_i^t$ instead of the past treatment history $Z_i^{1:T}$ is considered as a predictor of the the conditional expected outcome. The coefficient $\beta_1$ for exposure has a causal interpretation at the individual level. It is also the logarithm of the OR when both exposure and outcome are binary, and the function $h$ is logit.

When unmeasured confounding does not exist, i.e., $X_i^t$ can be fully observed, it is valid to estimate all coefficients in equation (1) using individual-level data. For example, the cohort design utilizes information about every unit in a group to examine associations with exposures (Benjamin et al., 1994). However, in observational studies, we cannot rule out the existence of unmeasured confounding, which may distort estimates of the fixed effects coefficients. In addition, the subject-specific model can be computationally challenging for the study of rare diseases because a large number of subjects is required.

#### 3.2.2.2 Population-Averaged Model

We assume the marginal expectation

$$h^*(\nu_i^t) = \beta_0^* + \beta_1^* Z_i^t + \beta_2^* t + C(Z_i^t, G) \tag{3.2}$$

where $h^*$ is the link function. $C(Z_i^t, G)$ is a function on exposure and group, which represents the heterogeneity across exposed and unexposed subgroups. The population-averaged model is the marginal expectation of the subject-specific model. It does not require knowledge of covariates or assumptions of the heterogeneity across individuals. Its coefficients are directly estimable from the aggregated data on exposure and outcome, but do not have individual causal interpretation.

#### 3.2.2.3 Connection between the Subject-Specific Model and the Population-Averaged Model

In general, the two models can be related by integrating out $X_i^t$. In Zeger et al. (1988),10 the cases of identity, log, probit, and logit link functions are discussed and the corresponding mathematical relations between $(\beta_0, \beta_1, \beta_2)$ and $(\beta_0^*, \beta_1^*, \beta_2^*)$ are listed in detail. The trend-in-trend method will be built on the population-averaged model. With the purpose of making causal inferences on individuals with a binary outcome, we require the link function $h$ to be logistic such that deriving OR $e^{\beta_1}$ is possible and the estimated quantity approximates the risk ratio obtained from a cohort study of a rare outcome.

We next provide a mathematical derivation of the connection between the two models and of how to estimate the causal OR using only data on trends in the prevalence of both exposure and outcome in strata. We further show that under plausible assumptions, the trend-in-trend method is unconfounded by measured and unmeasured factors, provided that there are no trends in the prevalence of covariates

that are correlated with the prevalence of the exposure over time. As the scenarios that will lead to a confounded estimate in a trend-in-trend study are a subset of those that will lead to a confounded estimate in a cohort study, the trend-in-trend design is more resistant to potential confounding. However, unlike the cohort design, the trend-in-trend design requires a strong time-trend in exposure, so is available in fewer scenarios.

### 3.2.3 Estimation of the Odds Ratio

We first stratify the entire population into $K$ strata according to the quintiles of the estimated CPE. For each subgroup $G$ and each time period $t$, we aggregate the individual-level data to obtain quantities in the following table.

| | Outcome $Y_i^t = 1$ | Outcome $Y_i^t = 0$ | Total |
|---|---|---|---|
| Exposure $Z_i^t = 1$ | $n_{11}^t$ | $n_{10}^t$ | $n_1^t$ |
| Exposure $Z_i^t = 0$ | $n_{01}^t$ | $n_{00}^t$ | $n_0^t$ |

Because $h$ is the logit function, we have

$$\begin{aligned} E(Y_i^t|Z_i^t, G) &= E(E(Y_i^t|Z_i^t, G, X_i^t)) \\ &= \int \frac{exp(\beta_0 + \beta_1 Z_i^t + \beta_2 t + \gamma^T X_i^t)}{1 + exp(\beta_0 + \beta_1 Z_i^t + \beta_2 t + \gamma^T X_i^t)} dF(X_i^t|Z_i^t, G) \end{aligned} \tag{3.3}$$

In general, there is no closed-form for the marginal mean as a function of the fixed effects and $\beta_1$ cannot be identified. However, an approximate form is available when we impose the following assumptions:

(1) Covariates and time period have multiplicative effects on being exposed. i.e., $P(Z_i^t|X_i^t) = h_1(X_1^t)h_2(t)$. $h_1$ and $h_2$ are two deterministic functions but unknown.

(2) Covariates for all individuals in any subgroup $G$ are either time-invariant or change randomly over time. They are random variables from an unknown distribution, i.e., $P(X_i^t|G) = f_G$.

(3) The outcome is rare, and therefore we can omit the denominator of the integrand in equation.

With these assumptions, we have:

$$\begin{aligned} E(Y_i^t|Z_i^t, G) &\approx \int exp(\beta_0 + \beta_1 Z_i^T + \beta_2 t + \gamma^T X_i^t) dF(X_i^t|Z_i^t, G) \\ &= exp(\beta_0 + \beta_1 Z_i^T + \beta_2 t) E(\gamma^T X_i^t|Z_i^t, G) \end{aligned} \tag{3.4}$$

In order to expand $E(\gamma^T X_i^t|Z_i^t, G)$, we compute the conditional distribution of covariates $X_i^t$ given $Z_i^T$ and $G$ using the Bayes rule:

$$\begin{aligned} p(X_i^t|Z_i^t = 1, G) &= \frac{p(Z_i^t = 1, X_i^t|G)}{p(Z_i^t = 1|G)} = \frac{p(Z_i^t = 1|X_i^t, G)p(X_i^t|G)}{p(Z_i^t = 1|G)} \\ &= \frac{p(Z_i^t = 1|X_i^t)p(X_i^t|G)}{p(Z_i^t = 1|G)} = \frac{h_1(X_i^t)h_2(t)f_G}{p(Z_i^t = 1|G)} \end{aligned} \tag{3.5}$$

$$\begin{aligned} p(X_i^t|Z_i^t = 0, G) &= \frac{p(Z_i^t = 0, X_i^t|G)}{p(Z_i^t = 0|G)} = \frac{p(Z_i^t = 0|X_i^t, G)p(X_i|G)}{p(Z_i^t = 0|G)} \\ &= \frac{p(Z_i^t = 0|X_i^t)p(X_i^t|G)}{p(Z_i^t = 0|G)} = \frac{f_G - h_1(X_i^t)h_2(t)f_G}{p(Z_i^t = 0|G)} \end{aligned} \tag{3.6}$$

Therefore,

$$p(X_i^t|Z_i^t = 1, G) = \frac{h_1(X_i^t)h_(t)f_G}{p(Z_i^t = 1|G)} \tag{3.7}$$

$$p(X_i^t|Z_i^t = 0, G) = \frac{f_G - h_1(X_i^t)h_2(t)f_G}{p(Z_i^t = 0|G)} \tag{3.8}$$

Define the following constants which only depend on $G$

$$C_{1G} := \int exp(\gamma^T X_i^t) h_1(X_i^t) f_G dX_i^t \tag{3.9}$$

$$C_{2G} := \int exp(\gamma^T X_i^t) f_G dX_i^t \tag{3.10}$$

$$C_{3G} := \int h_1(X_i^t) f_G dX_i^t \tag{3.11}$$

The marginal expectation $E(Y_i^t | Z_i^t, G)$ now becomes:

$$E(Y_i^t | Z_i^t = 1, G) = exp(\beta_0 + \beta_1 + \beta_2 t) \frac{C_{1G}}{C_{3G}} \tag{3.12}$$

$$E(Y_i^t | Z_i^t = 0, G) = exp(\beta_0 + \beta_2 t) \frac{C_{2G} - C_{1G} h_2(t)}{1 - C_{3G} h_2(t)} \tag{3.13}$$

where $C_{1G}, C_{2G}, C_{3G}$ are unknown constants that depend on group.

Equations (3.12) and (3.13) are covariate-free. In other words, the marginal expectation of outcome is the same across treated/control individuals within the same subgroup. Because each $Y_i^t$ is binary, aggregating outcomes for the treated and untreated yields two binomial distributions. Consequently, we can write the parametric likelihood for $(n_{11}^t, n_{10}^t, n_{01}^t, n_{00}^t)$:

$$n_{11} \sim Binomial(n_{11}^t + n_{10}^t, e^{\beta_0 + \beta_1 + \beta_2 t} \frac{C_{1G}}{C_{3G}}) \tag{3.14}$$

$$n_{01} \sim Binomial(n_{01}^t + n_{00}^t, e^{\beta_0 + \beta_2 t} \frac{C_{2G} - h_2(t) C_{1G}}{1 - h_2(t) C_{3G}}) \tag{3.15}$$

$(\beta_0, \beta_1, \beta_2 C_{1G}, C_{2G}, C_{3G})$ are unknown parameters and can be estimated by maximizing the above likelihood using an optimization algorithm. In particular, $e^{\beta_1}$ is the OR of interest. We have written a package for the R computing language called TrendInTrend that performs this maximization and calculates the OR with its 95% confidence interval given $(n_{11}^t, n_{10}^t, n_{01}^t, n_{00}^t)$, $t \in \{1, 2, \ldots, T\}$ .

## 3.3 Simulations: Comparing the Trend-in-trend Design with the Cohort Study Method

### Setup

We performed simulation studies to confirm that when unmeasured confounding is present, the OR produced by the trend-in-trend method is negligibly biased (albeit somewhat less precise) than that produced by a cohort study. We simulated a study population of size 250,000 with 20 calendar quarters as study periods. The data were generated according to the following procedure:

- Step 1: The covariates $X_i^t$ are a five-dimensional vector with three entries generated from a multivariate Gaussian distribution and two other entries generated from Bernoulli distributions with different success probabilities. Three scenarios are examined: 1) covariates are sampled only once and fixed over time 2) covariates are sampled independently for each calendar period 3) covariates are sampled repeatedly for each calendar period with autocorrelation coefficient of 0.5.

- Step 2: Assign $Z_i^t$ to 1 with the probability of $e^{a_0+a_1X_I^t+a_2t+a_3t^2}$.

- Step 3: Simulate $Y_i^t$ based on the subject-specific model and the choice of link function $h$.

We choose $(a_0, a_1, a_2, a_3)$ such that the simulated exposure prevalence has the up-and-down shape shown in Fig. 5.1, which mimics the exposure trend of a drug that becomes widely used after introduction, and is then withdrawn (e.g., rofecoxib). However, the method should work for a unidirectional trend as well.

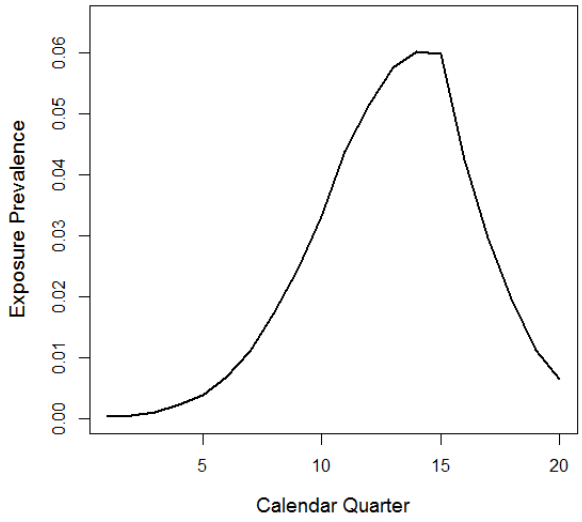



Figure 3.1: Simulated overall exposure prevalence over time

Based on the CPEs estimated via logistic regression, the study population was stratified into quintiles, i.e., $K = 5$. As expected, these strata, each with 50,000 individuals, had different trends in exposure prevalence. The CPE model included all five covariates, as shown in Fig.5.2.

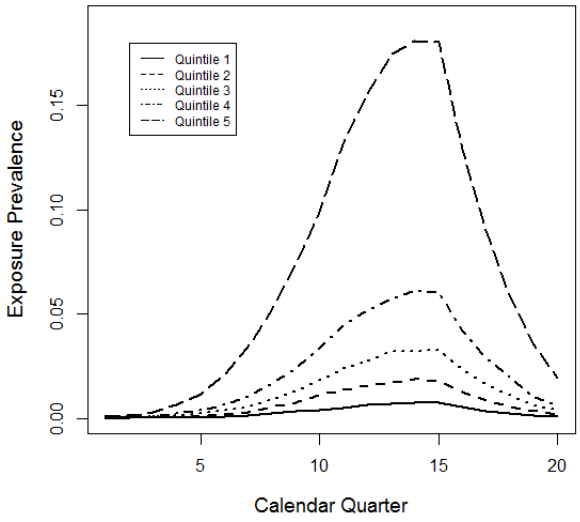



Figure 3.2: Simulated exposure prevalence across subgroups based on CPE quintiles over time.

We considered the following scenarios under the rare events assumption: (1), the OR takes values of 1.0, 1.5, 2.0, and 2.5; (2) the strength of the CPE model has three levels quantified by 0, 2, and 4 omitted confounders out of 5 confounders in total, and a c-statistic is calculated for each level to gauge unobserved heterogeneity in factors affecting outcome; (3) the number of CPE strata is either 5. We compare the estimated OR with those calculated using the cohort method. The results, which are the average values of 1000 simulations, are summarized in Tables 3.1, 3.2, and 3.3, corresponding three different scenarios of covariates sampling as described above.

## Results

| True Odds Ratio | Number of Unmeasured Confounders | C-statistic of the CPE Model | Trend-in-Trend Odds Ratio | | Cohort Study Odds Ratio | |
|---|---|---|---|---|---|---|
| | | | Mean (% bias) | SD | Mean (% bias) | SD |
| 2.5 | 0 | 0.68 | 2.47 (-1.2) | 0.0165 | 2.50 (0.0) | 0.0092 |
| 2.5 | 2 | 0.63 | 2.45 (-2.0) | 0.0170 | 4.75 (90.0) | 0.0091 |
| 2.5 | 4 | 0.51 | 2.43 (-2.8) | 0.0171 | 4.80 (92.0) | 0.0091 |
| 2.0 | 0 | 0.68 | 1.97 (-1.5) | 0.0147 | 2.01 (0.5) | 0.0087 |
| 2.0 | 2 | 0.63 | 1.95 (-2.5) | 0.0153 | 4.22 (111) | 0.0081 |
| 2.0 | 4 | 0.51 | 1.94 (-3.0) | 0.0131 | 4.25 (113) | 0.0078 |
| 1.5 | 0 | 0.68 | 1.52 (1.3) | 0.0101 | 1.50 (0.0) | 0.0083 |
| 1.5 | 2 | 0.63 | 1.49 (-0.7) | 0.0106 | 3.25 (117) | 0.0081 |
| 1.5 | 4 | 0.51 | 1.48 (-1.3) | 0.0108 | 3.30 (120) | 0.0082 |
| 1.0 | 0 | 0.68 | 1.02 (2.0) | 0.0082 | 0.99 (-1.0) | 0.0079 |
| 1.0 | 2 | 0.63 | 1.02 (2.0) | 0.0089 | 2.08 (108) | 0.0074 |
| 1.0 | 4 | 0.51 | 1.02 (2.0) | 0.0089 | 2.20 (120) | 0.0073 |

Table 3.1: Comparison of the estimated causal odds ratio using the time-in-trend design and the cohort study method. Confounders are sampled only once and fixed over time.

| True Odds Ratio | Number of Unmeasured Confounders | C-statistic of the CPE Model | Trend-in-Trend Odds Ratio | | Cohort Study Odds Ratio | |
|---|---|---|---|---|---|---|
| | | | Mean (% bias) | SD | Mean (% bias) | SD |
| 2.5 | 0 | 0.62 | 2.46 (-1.6) | 0.0198 | 2.50 (0.0) | 0.0101 |
| 2.5 | 2 | 0.57 | 2.45 (-2.0) | 0.0207 | 4.79 (91.6) | 0.0104 |
| 2.5 | 4 | 0.41 | 2.41 (-3.6) | 0.0212 | 4.91 (96.4) | 0.0104 |
| 2.0 | 0 | 0.62 | 2.03 (-1.5) | 0.0184 | 2.01 (0.5) | 0.0091 |
| 2.0 | 2 | 0.57 | 1.94 (-3.0) | 0.0191 | 4.28 (114) | 0.0089 |
| 2.0 | 4 | 0.41 | 1.93 (-3.5) | 0.0177 | 4.32 (116) | 0.0090 |
| 1.5 | 0 | 0.62 | 1.53 (2.0) | 0.0124 | 1.51 (0.7) | 0.0092 |
| 1.5 | 2 | 0.57 | 1.53 (2.0) | 0.0132 | 3.24 (116) | 0.0088 |
| 1.5 | 4 | 0.41 | 1.46 (-2.7) | 0.0129 | 3.38 (125) | 0.0094 |
| 1.0 | 0 | 0.62 | 1.02 (2.0) | 0.0098 | 0.99 (-1.0) | 0.0087 |
| 1.0 | 2 | 0.63 | 1.03 (3.0) | 0.0111 | 2.11 (111) | 0.0083 |
| 1.0 | 4 | 0.51 | 0.97 (-3.0) | 0.0112 | 2.27 (127) | 0.0084 |

Table 3.2: Comparison of the estimated causal odds ratio using the time-in-trend design and the cohort study method. The population is stratified into five subgroups for the time-in-trend algorithm. Confounders are sampled independently for each calendar period.

| True Odds Ratio | Number of Unmeasured Confounders | C-statistic of the CPE Model | Trend-in-Trend Odds Ratio | | Cohort Study Odds Ratio | |
|---|---|---|---|---|---|---|
| | | | Mean (% bias) | SD* | Mean (% bias) | SD* |
| 2.5 | 0 | 0.66 | 2.46 (-1.6) | 0.0195 | 2.50 (0.0) | 0.0097 |
| 2.5 | 2 | 0.60 | 2.45 (-2.0) | 0.0202 | 4.78 (89.2) | 0.0098 |
| 2.5 | 4 | 0.41 | 2.42 (-3.2) | 0.0207 | 4.87 (94.8) | 0.0098 |
| 2.0 | 0 | 0.66 | 1.98 (-1.0) | 0.0176 | 2.00 (0.0) | 0.0087 |
| 2.0 | 2 | 0.60 | 1.94 (-3.0) | 0.0185 | 4.23 (112) | 0.0086 |
| 2.0 | 4 | 0.46 | 1.94 (-3.0) | 0.0172 | 4.30 (115) | 0.0085 |
| 1.5 | 0 | 0.66 | 1.53 (2.0) | 0.0119 | 1.51 (0.7) | 0.0087 |
| 1.5 | 2 | 0.60 | 1.52 (1.3) | 0.0125 | 3.25 (117) | 0.0086 |
| 1.5 | 4 | 0.46 | 1.47 (-2.0) | 0.0122 | 3.35 (123) | 0.0091 |
| 1.0 | 0 | 0.66 | 1.02 (2.0) | 0.0094 | 0.99 (-1.0) | 0.0081 |
| 1.0 | 2 | 0.60 | 1.02 (2.0) | 0.0105 | 2.09 (109) | 0.0079 |
| 1.0 | 4 | 0.46 | 1.03 (3.0) | 0.0107 | 2.20 (120) | 0.0080 |

Table 3.3: Comparison of the estimated causal odds ratio using the time-in-trend design and the cohort study method. Confounders are sampled with autocorrelation 0.5 between any two consecutive calendar periods.

As expected, when there were no unmeasured confounders, both the trend-in-trend and cohort designs yielded ORs that were close to the truth. However, as the number of unmeasured confounders increased, the ORs produced by the cohort design became very biased, with biases ranging from 90% to 127%, while those from the trend-in-trend design remained close to the truth, with bias ranging from -3.5% to 3%. The standard deviations for the trend-in-trend method were one to two times as large as those for the cohort method, which is to be expected as individual-level information is partially lost when counts of outcomes are aggregated.

## 3.4 Application: Confirming THE Causal effect of Rofecoxib on AMI

We applied the trend-in-trend method to *Clinformatics*$^{TM}$ Data Mart Database (OptumInsight, Eden Prairie, MN) to examine association between rofecoxib and AMI, severe hypoglycemia, and non-vertebral bone fracture. We first identified all persons age 18 years or older in Optum who received one or more prescriptions for rofecoxib during the study period from April 1, 2000 through Dec 30, 2004. For each rofecoxib-exposed person episode, we ascertained the first month and the last month of their continuous enrollment episode (or episodes, for persons with multiple enrollment episodes) during the study period. Thus, the unit of observation was the enrollment episode, defined as a period of continuous enrollment for a person. A person could contribute multiple episodes. For each rofecoxib-exposed episode, we randomly sampled, without replacement, nine rofecoxib-unexposed enrollment episodes with an enrollment start date on or before no more than one year of the rofecoxib-exposed subjects enrollment start date, and with an enrollment end date on or after the rofecoxib-exposed subjects enrollment end date. The rationale for this criterion

was to ensure sufficient overlap in follow-up calendar time for exposed and unexposed subjects. Thus, the analysis set contained ten times as many total episodes as there were rofecoxib-exposed enrollment episodes. This was done to improve computational efficiency versus including the entire study population.

We then fit a logistic regression to estimate the CPE using age, sex, diagnosis of rheumatoid arthritis, and diagnosis of osteoarthritis as explanatory variables. For rofecoxib-exposed subjects, these covariates were measured at their first prescription date. For control subjects, these covariates were measured the same date as their corresponding exposed subjects. The c-statistic was 0.608, which produced good separation of exposure prevalence across quintiles, as shown in Figure 3. The estimated coefficients and standard deviations (as shown in parentheses) are 0.0228 (0.0001) for continuous age, 0.1458 (0.0027) for female sex, 2.4418 (0.0124) for rheumatoid arthritis, and -0.6444 (0.0191) for osteoarthritis.

The trend-in-trend method yielded an OR (95% confidence interval) for rofecoxib and AMI of 1.23 (1.05, 1.44), which is consistent with the results of prior epidemiologic studies: a 2005 meta-analysis yielded a pooled relative risk of 1.20 (1.10, 1.30) for cohort and nested case-control studies (Hernández-Díaz et al., 2006), and a more recent meta-analysis reported a pooled relative risk of 1.34 (1.22, 1.48) (Varas-Lorenzo et al., 2013). The ORs for the negative control outcomes, severe hypoglycemia and non-vertebral bone fracture (neither of which is thought to be affected by rofecoxib), were 1.09 (0.92, 1.28) and 0.84 (0.64, 1.09), which are both consistent with no effect (Solomon et al., 2010).

## 3.5 Discussion

We describe a novel hybrid ecologic-epidemiologic study design called the trend-in-trend design, provide a mathematical derivation of the resulting odds ratio, use simulation to confirm that the results are less biased (albeit somewhat less precise) than those of a cohort study when there is unmeasured confounding, and apply that method to reproduce one positive and two null associations using real-world data. The results of the empiric study using real-world data show that the design is readily applicable and produces expected results.

Importantly, the trend-in-trend design avoids the Achilles heel of most epidemiologic studies of healthcare interventions: conflation of receiving a treatment with needing that treatment. Unlike cohort studies, the trend-in-trend design does not assume no unmeasured confounders, but instead examines changes in outcome occurrence as a function of changes in exposure prevalence across strata with differential time-trends in exposure. Therefore, the results of a trend-in-trend study will be unconfounded unless there are unmeasured factors affecting outcome for which there are time-trends in prevalence that are correlated with time-trends in exposure across the strata defined by exposure trend. This could occur if there are co-interventions for which the trend in use is positively correlated with trends in use of the exposure, or alternatives for which the trend in use is negatively correlated with trends in use of the exposure. As the scenarios that would produce a confounded trend-in-trend estimate are a subset of those that would produce a confounded cohort estimate, the trend-in-trend design is more resistant to confounding. However, the trend-in-trend design is feasible only if there is a strong time-trend in exposure prevalence. Similarly, the effect estimates produced using calendar period as an IV will be biased if there is any time-trend in an unmeasured causal factor, whereas a trend-in-trend study will be biased only if changes in the prevalence of such a factor are correlated

with changes in exposure prevalence across CPE strata. The trend-in-trend design therefore relaxes the assumptions under which use of calendar time as an IV is valid.

The causal contrast examined by the trend-in-trend approach deserves discussion. It is the instantaneous effect of use of the exposure of interest rather than the exposure(s) (if any) that the increasing (or declining) trend in use of the exposure of interest displaced (or was displaced by). In the example of rofecoxib, this is likely to be some combination of nonselective nonsteroidal anti-inflammatory drugs, opioids, and no treatment. Thus, the trend-in-trend results may not mimic the results of placebo-controlled trials evaluating the study treatment. Nevertheless, the causal contrast with the alternatives that it displaces or is displaced by is arguably more relevant from a public health perspective.

The main limitations of the trend-in-trend method are the need for a strong trend in exposure prevalence and the reduced statistical precision that accompanies group-level rather than individual-level analyses. Limitations of the current study include the modest range of scenarios simulated and the fact that there is no empirical example with a causal effect known with complete certainty.

Additional work is needed to improve the utility of the trend-in-trend design. Such work should address control for measured factors for which there may be time-trends that are correlated with time-trends in exposure across CPE strata, examination of treatment effect heterogeneity, sequential analysis methods to allow multiple looks while limiting type-1 error, and estimation of statistical power and detectable alternative.

# 4

# Squential Testing for the Trend-in-trend Design: an Application to Drug Safety Surveillance In the Presence of Unmeasured Confounding


## Abstract

Post license drug safety surveillance is a critical step of the drug evaluation because rare but serious adverse events may not be detected in pre-license randomized trials. Sequential testing methods are powerful tools that facilitate early termination of the drug usage when the drug exceeds the pre-assumed adverse event rate. However, applying sequential tests on observational data can be misleading in the presence of unmeasured confounders. We generalize the standard sequential testing to trend-in-trend design settings that utilizes time trends in exposure prevalence and accounts for both measured and unmeasured confounding. The performance of the proposed approach is examined and compared to other approaches using simulation studies. We also apply the method to Clinformatics Data Mart Database (OptumInsight, Eden Prairie, MN) to test the risk of rofecoxib on acute acute myocardial infarction (AMI).


*Joint work with Ashkan Ertefaie, Sean Hennessy, Charles Leonard, Dylan Small

## 4.1 Introduction

Identifying all the adverse events (AE) of a drug may not be possible during the pre-license randomized trials because such trials are often powered for efficiency (DuMouchel, 1999; Davis et al., 2005). In fact, investigators often do not even have a comprehensive list of all the possible AEs because treatments can interact with patients' genotype, characteristics and other treatments. Thus there are always possibilities of unexpected AEs. This motivates investigators to use big databases such as electronic health records to study the risk of AEs among exposed and unexposed groups. Another appealing feature of electronic health records is that because they are updating frequently, investigators can monitor AEs in real time. For the public health safety reasons, the drug use must be terminated as soon as there are enough evidence of increased AEs rate which can only be done using sequential testing methods, see Ghosh et al. (2011); Mukhopadhyay and De Silva (2008); Govindarajulu (2004).

Wald (Wald, 1945, 1947) proposed a sequential probability ratio test (SPRT), where the null hypotheses of drug safety is rejected when the likelihood ratio exceeds a predetermined critical value. SPRT approves the safety of the drug if by the end of the study period, the likelihood ratio stays below the critical value. An important feature of the test is that it adjusts for the p-values without knowing the number of times that the test needs to be performed. One drawback of Wald test is that the result highly depends on the specified alternative hypothesis. In fact, Kulldorf et. al. Kulldorff et al. (2011) showed that for some alternative hypotheses, Wald test can significantly delay or completely miss the signal. The latter paper proposed a maximized sequential probability ration test (MaxSPRT) that handles the problem by considering a composite alternative rather than simple (Hoel et al., 1976; Lachin, 1981; Meeker Jr, 1981; Van der Tweel et al., 1996). Sequential testing has also been

extended to the Bayesian settings where a continuous-time Poisson process is assumed for AEs (Lechner, 1962; Peskir and Shiryaev, 2000).

There are some sequential testing methods that are designed specifically for the observation safety surveillance settings. Brown et. al. Brown et al. (2007) used a Poisson based MaxSPRT that adjusts for confounding by stratifying patients based on their baseline characteristics. However, this method requires reliable estimate of the expected number of AEs under the null hypothesis which mat not be available in many settings. Li (Li, 2009) addressed this shortcoming by proposing the conditional sequential sampling procedure (CSSP) that preserves the type-I error rate by implementing $\alpha$-spending approaches. Li's method also adjusts for confounding by stratifying patients based on their baseline characteristics (see also Lan and DeMets (1983) and Jennison and Turnbull (1997)). Group sequential generalized estimating equations (GS GEE) is another approach that adjusts for confounding using an estimating equation based method (**?**). Lan and DeMets (Lan and DeMets, 1983) also adopted the $\alpha$-spending approach that adjusts for confounding in a regression context (Jennison and Turnbull, 1997). For more detailed discussion on sequential safety monitoring using observational data see Stratton (2012).

All the existing sequential testing methods in observational settings rely heavily on an unstable assumption of no unmeasured confounding. Because electronic health records are not, in general, collected for scientific purposes, the no unmeasured confounding assumption is unlikely to hold. We propose a sequential likelihood ratio test with trend-in-trend design (SLR-TT) that is robust to unmeasured confounding under certain assumptions. Per the discussion in Chapter 3, the trend-in-trend design is a novel design that is used in observational settings to study causal effect of treatments for which there are strong time trends. We show that the results obtained by the existing methods may be misleading in the presence of unmeasured confounding

while the proposed SLR-TT provides valid results.

## 4.2 Sequential Testing for the Trend-in-trend Design

Inferences based on observational data are subject to bias due to unmeasured confounding. Ji et. al. Ji et al. (2016) proposed a new hybrid ecologic-epidemiologic design called the trend-in-trend design, that utilizes time trends in exposure prevalence and accounts for both measured and unmeasured confounding. In the trend-in-trend design we first estimate the cumulative probability of exposure (CPE; the predicted probability of cumulative exposure based on variables other than exposure and outcome) to stratify individuals and hence identify subgroups with different exposure trends over the period of study. For example, when we define the treated group as patients who have been exposed to treatment at least once during, then the CPE is a logistic regression that includes patients characteristics and possibly time as independent variable and the treatment indicator as dependent variable. In this particular example, the statistical model of CPE is similar to the propensity score. However, the CPE is only used to identify subgroups with different trends in exposure prevalence and not for balancing purposes. At the second stage, we aggregate the data for each subgroup and each time, and form a $2 \times 2$ table that shows the number of patients who did and didn't experience the AE and AEs among treated and untreated group. Then a likelihood is derived and the parameters of interest are estimated using MLE approach. In the sequel, we discuss the trend-in-trend design and the proposed sequential testing approach in more details.

We study the effect of two treatment groups on the rate of adverse events using a longitudinal data collected over a fixed period of time. Our dataset is composed of $n$

i.i.d. trajectories of length $T$. The $i$th trajectory is the sequence $(\mathbf{X}_i, D_{i1}, Y_{i1}, ..., D_{iT}, Y_{iT})$ where $\mathbf{X}_i$ is a vector of measured baseline characteristics that follows a distribution $F$. $D_{it}$ is the treatment status at time $t$ and $Y_{it}$ is the indicator of whether an AE has happened between time $t-1$ and $t$. Let $g = 1, 2, ..., G$ denote the index for subgroups after stratification based on CEP. The aggregated individual-level data based on each subgroup $g$ and time $t$ consists of $n^t_{11g}$ patients who were treated and experienced AEs and $n^t_{01g}$ patients who were untreated and experienced AEs. We also have the number of patients who were treated (untreated) and did not experience an AE denoted as $n^t_{10g}$ ($n^t_{00g}$).

### 4.2.1 Derivation of the Likelihood Ratio

Assuming a logit model for AEs, the conditional mean of $Y_{it}$ given $D_{it}$, and $g$ is

$$\begin{aligned}\boldsymbol{E}(Y_{it}|D_{it}, g) &= \boldsymbol{E}[\boldsymbol{E}(Y_{it}|D_{it}, g, \mathbf{X}_i)] \\ &= \int \frac{exp(\beta_0 + \beta_1 D_{it} + \beta_2 t + \gamma^\top \mathbf{X}_i)}{1 + exp(\beta_0 + \beta_1 D_{it} + \beta_2 t + \gamma^\top \mathbf{X}_i)} dF(\mathbf{X}_i|D_{it}, g)\end{aligned}$$

which does not have closed-form. However, assuming that the outcome is a rare event and the covariates and time have multiplicative effects on being treated, i.e. $P(D_{it}|\mathbf{X}_i) = h_1(\mathbf{X}_i)h_2(t)$, the conditional mean can be written as

$$\begin{aligned}\boldsymbol{E}(Y_{it}|D_{it}, g) &\approx \int exp(\beta_0 + \beta_1 D_{it} + \beta_2 t + \gamma^\top \mathbf{X}_i) dF(\mathbf{X}_i|D_{it}, g) \\ &= exp(\beta_0 + \beta_1 D_{it} + \beta_2 t)\boldsymbol{E}(\gamma^\top \mathbf{X}_i|D_{it}, g).\end{aligned}$$

Thus, $\beta_1$ is the parameter of interest and null hypothesis can be written as $H_0 : \beta_1 = 0$. Consequently, the likelihood in the trend-in-trend analysis is given by

$$L(\beta_0, \beta_1, \beta_2) = \prod_g \prod_t \left( exp(\beta_0 + \beta_1 + t\beta_2) \frac{C_{1g}}{C_{3g}} \right)^{n_{11g}^t} \left( 1 - exp(\beta_0 + \beta_1 + t\beta_2) \frac{C_{1g}}{C_{3g}} \right)^{n_{10g}^t}$$
$$\left( exp(\beta_0 + t\beta_2) \frac{C_{2g} - h_2(t)C_{1g}}{1 - h_2(t)C_{3g}} \right)^{n_{01g}^t} \left( 1 - exp(\beta_0 + t\beta_2) \frac{C_{2g} - h_2(t)C_{1g}}{1 - h_2(t)C_{3g}} \right)^{n_{00g}^t}$$

where

$$C_{1g} = \int exp(\gamma^\top \mathbf{X}_i) h_1(X_i) f_g d\mathbf{X}_i$$
$$C_{2g} = \int exp(\gamma^\top \mathbf{X}_i) f_g d\mathbf{X}_i$$
$$C_{3g} = \int h_1(\mathbf{X}_i) f_g d\mathbf{X}_i,$$

are unknown constants that depend on stratum $g$ and $f_g = f(\mathbf{X}_i|g)$ Ji et al. (2016). Following Shih et. al. Shih et al. (2010), the stopping rule is

$$\tau = \inf \left\{ t \geq 1 : \log \frac{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)}{L(\hat{\beta}_0^{H_0}, 0, \hat{\beta}_2^{H_0})} > c \right\},$$

where $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ are the maximum likelihood estimates of the parameter vector $(\beta_0, \beta_1, \beta_2)$ and $(\hat{\beta}_0^{H_0}, 0, \hat{\beta}_2^{H_0})$ are the maximum likelihood estimates of the corresponding parameters under the null hypothesis $H_0 : \beta_1 = 0$. The critical value $c$ is approximated using a Monte Carlo method that is discussed latter. The log-likelihood ratio is

$$LLR = \log \frac{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)}{L(\hat{\beta}_0^{H_0}, 0, \hat{\beta}_2^{H_0})}$$
$$= \sum_g \sum_t n_{11g}^t (\hat{\beta}_0 + \hat{\beta}_1 + t\hat{\beta}_2) + n_{11g}^t \log(\kappa_g) + n_{10g}^t \log(1 - exp(\hat{\beta}_0 + \hat{\beta}_1 + t\hat{\beta}_2)\kappa_g)$$

$$
\begin{aligned}
&+ n^t_{01g}(\hat{\beta}_0 + t\hat{\beta}_2) + n^t_{01g}\log(\kappa^t_g) + n^t_{00g}\log(1 - exp(\hat{\beta}_0 + t\hat{\beta}_2)\kappa^t_g) \\
&- n^t_{11g}(\hat{\beta}^{H_0}_0 + t\hat{\beta}^{H_0}_2) - n^t_{11g}\log(\kappa_g) - n^t_{10g}\log(1 - exp(\hat{\beta}^{H_0}_0 + t\hat{\beta}^{H_0}_2)\kappa_g) \\
&- n^t_{01g}(\hat{\beta}^{H_0}_0 + t\hat{\beta}^{H_0}_2) - n^t_{01g}\log(\kappa^t_g) - n^t_{00g}\log(1 - exp(\hat{\beta}^{H_0}_0 + t\hat{\beta}^{H_0}_2)\kappa^t_g) \\
&= \sum_g \sum_t n^t_{11g}(\hat{\beta}_0 - \hat{\beta}^{H_0}_0 + \hat{\beta}_1 + t\hat{\beta}_2 - t\hat{\beta}^{H_0}_2) + n^t_{01g}(\hat{\beta}_0 - \hat{\beta}^{H_0}_0 + t\hat{\beta}_2 - t\hat{\beta}^{H_0}_2) \\
&+ n^t_{10g}\log\frac{1 - exp(\hat{\beta}_0 + \hat{\beta}_1 + t\hat{\beta}_2)\kappa_g}{1 - exp(\hat{\beta}^{H_0}_0 + t\hat{\beta}^{H_0}_2)\kappa_g} + n^t_{00g}\log\frac{1 - exp(\hat{\beta}_0 + t\hat{\beta}_2)\kappa^t_g}{1 - exp(\hat{\beta}^{H_0}_0 + t\hat{\beta}^{H_0}_2)\kappa^t_g}
\end{aligned}
$$

where $\kappa_g = \frac{C_{1g}}{C_{3g}}$ and $\kappa^t_g = \frac{C_{2g} - h_2(t)C_{1g}}{1 - h_2(t)C_{3g}}$.

## 4.3 The Sequential Likelihood Ratio Algorithm

1. With collected data up to $T \leq T_{max}$, fit a CPE model to stratify the population and tabulate AEs in treated and control groups.

2. Estimate the log odds ratio and all nuisance parameters using MLE. $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{C}_{1g}, \hat{C}_{2g}, \hat{C}_{3g})$.

3. Estimate nuisance parameters under the null hypothesis $H_0 : \beta_1 = 0$. $(\hat{\beta}^{H_0}_0, 0, \hat{\beta}^{H_0}_2, \hat{C}_{1g}, \hat{C}_{2g}, \hat{C}_{3g})$.

4. Calculate log likelihood ratios, $\overline{LLR} = (LLR(1), LLR(2), \ldots, LLR(T))$.

5. Reject the null at time $\tau$ using the following Stopping Rule

$$\tau = \inf\{t \leq T : LLR(t) > c_T\}$$

6. If the $\tau > T$, proceed to time $T + 1$ and repeat step 1.

### 4.3.1 Critical Value Approximation

The log likelihood ratio of the trend-in-trend design does not have known asymptotic distribution and we approximate the critical value $c$ using a Monte Carlo simulation method that is outlined below:

1. Consider a set of candidates for $c_T$, say $\mathcal{C}$.

2. Given the estimated parameters $(\hat{\beta}_0^{H_0}, 0, \hat{\beta}_2^{H_0}, \hat{C}_{1g}^{H_0}, \hat{C}_{2g}^{H_0}, \hat{C}_{3g}^{H_0})$ under the null hypothesis, generate $K$ simulated datasets and obtain $K$ realizations of $\overline{LLR}$, i.e., each realization is a vector $\overline{LLR}_k = (LLR_k^{sim}(1), ...LLR_k^{sim}(T))$.

5. For any $c \in \mathcal{C}$, calculate the proportion of times that a signal is detected by $T$ using the stopping rule. Let $\hat{p}(\overline{LLR} > c|H_0) = \frac{\sum_k I(\tau_K <= T)}{K}$.

$$\tau_k = \inf\left\{t \leq T : LLR_k^{sim}(t) > c\right\}$$

6. Pick the value of $c_T$ such that $\hat{p}(\overline{LLR} > c|H_0) \approx \alpha(T)$ where $\alpha(T)$ is the type-I error rate spent up to $T$.

## 4.4 Simulations: Comparing SLR-TT with CSSP

### Setup

We present a simulation study with population size $N$=50,000 and calendar periods $T = 10$. We stratify the population based on CEPs into five subgroups, hence each subgroup has 10,000 individuals. The vector of baseline covariates $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)$, where $X_1 \sim N(2, 1)$, $X_2 \sim N(2, 1)$, $X_3 \sim Bernoulli(0.8)$, $X_4 \sim Bernoulli(0.2)$, and $X_5 \sim Bernoulli(0.1)$.

We consider three sets of simulation settings in which there are no unmeasured confounders, 2 out of 5 unmeasured confounders, and 4 out of 5 unmeasured confounders. Within each set, the rejection rate and the average signal time of SLT-TT and CSSP are evaluated under various scenarios with different values of the odds ratio OR (which is approximately the risk ratio for rare events), and the exponent of the $\alpha$-spending function, where $\alpha(k) = \alpha(k/K)^{\gamma}$ and $\alpha = 0.05$, $\gamma = .5, 1, 2$. Different $\alpha$-spending functions allocate different Type-I error we want to spend at each interim test, as shown in Figure 4.1

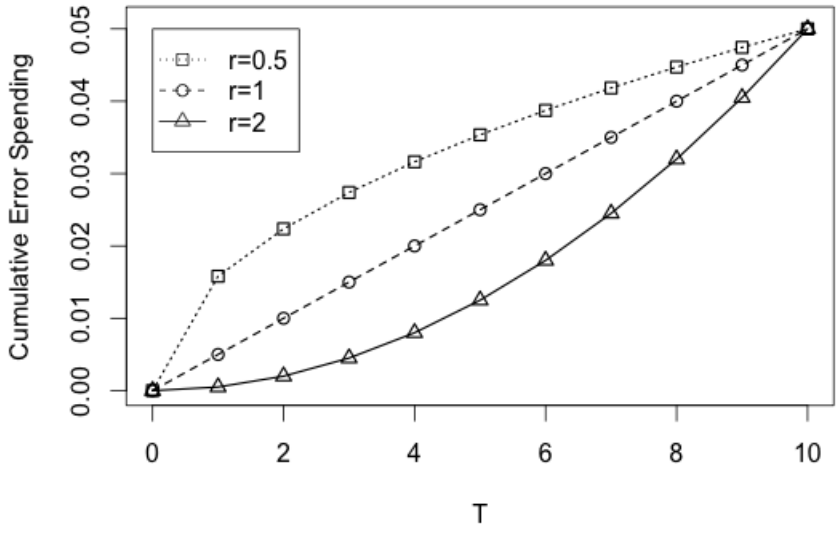



Figure 4.1: Allocation of Type-I error rates for $\alpha$-spending functions $\alpha(k) = \alpha(k/K)^{\gamma}$ with $\alpha = 0.05$ and $\gamma = .5, 1, 2$

At each calendar period $t \leq T$, the treatment and outcome variables are generated from

$$Z^t \sim Bernoulli(exp\{\alpha_0 + \alpha_1 \mathbf{X} + \alpha_2 t - \alpha_3 t^2\})$$

$$Y^t \sim Bernoulli(exp\{\beta_0 + \beta_1 Z^t + \beta_2 t + \beta_3 \mathbf{X}\})$$

where $\alpha_0 = -15, \alpha_1 = (2, 1, 0.1, 0.1, 0.1), \alpha_3 = 0.031$ and $\beta_0 = -4, \beta_1 = log(OR), \beta_2 = 0.001, \beta_3 = 0.1(2, 1, 0.1, 0.1, 0.1)$. The parameter $\alpha_2$ represents the incidence rate of exposure, which plays a critical role in the trend-in-trend analysis. Specifically, as the value of $\alpha_2$ increases the power of the trend-in-trend analysis increases. To examine the effect of the incidence rate of exposure in our sequential setting, we consider a high incidence rate of exposure ($\alpha_2 = 0.9$ ) and a low incidence rate of exposure ($\alpha_2 = 0.6$). Moreover, we vary the odds ratio (OR) by setting $\beta_1 = \log(2.5)$, $\log(2)$, $\log(1.5)$, and 0. We parametrized the treatment assignment model so that the exposure has an upward trend, which mimics the exposure trend of a newly approved drug that becomes widely used after market debut (e.g., rofecoxib). The results are based on 1,000 datasets simulated from the generative models.

## Results

Table 4.1 shows that when there is no unmeasured confounding, both SLR-TT and CSSP achieve the 0.05 nominal error rate and when there is a relatively strong exposure trend, i.e., $\alpha_2 = 0.9$, the proposed SLR-TT have more power than the CSSP. The signal detection period is also shorter in SLR-TT. When the exposure trend is weaker, i.e., $\alpha_2 = 0.6$, the SLR-TT has slightly less power than the CSSP and the signal detection times are slightly longer as well.

Table 4.2 shows that, in the presence of unmeasured confounding, the CSSP method results in an inflated type-I error rate that is roughly 3 time higher than the nominal 0.05 rate, while the SLR-TT maintain the 0.05 rate.

| True Odds Ratio | Rejection Rate | | | | | | Rejection Periods | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SLR-TT | | | CSSP | | | SLR-TT | | | CSSP | | |
| | r=.5 | r=1 | r=2 | r=.5 | r=1 | r=2 | r=.5 | r=1 | r=2 | r=.5 | r=1 | r=2 |
| strong trend of exposure prevalence | | | | | | | | | | | | |
| 1 | .047 | .046 | .046 | .047 | .047 | .047 | 4.144 | 4.921 | 6.890 | 4.681 | 5.462 | 7.002 |
| 1.5 | .485 | .481 | .476 | .337 | .328 | .327 | 3.591 | 4.845 | 5.689 | 4.026 | 5.373 | 6.155 |
| 2 | .734 | .717 | .709 | .581 | .536 | .522 | 2.680 | 3.750 | 4.717 | 3.441 | 4.335 | 5.702 |
| 2.5 | 1 | 1 | 1 | .838 | .815 | .809 | 2.122 | 2.910 | 3.559 | 2.524 | 3.248 | 4.113 |
| week trend of exposure prevalence | | | | | | | | | | | | |
| 1 | 048 | .046 | .046 | .048 | .048 | .047 | 4.566 | 6.226 | 7.021 | 4.130 | 5.966 | 6.771 |
| 1.5 | .309 | .293 | .274 | .372 | .313 | .290 | 4.788 | 5.871 | 6.513 | 4.210 | 5.502 | 6.193 |
| 2 | .437 | .426 | .420 | .538 | .502 | .485 | 4.012 | 4.755 | 5.787 | 3.745 | 4.586 | 5.228 |
| 2.5 | .687 | .671 | 661 | .803 | .770 | .748 | 3.270 | 3.942 | 4.596 | 2.689 | 3.551 | 4.207 |

Table 4.1: Comparing rejection rates and average signal detection periods for different odds ratio when there is no unmeasured confounding

| Number of Unmeasured Confounders | True Odds Ratio | Rejection Rate | | | | | | Rejection Periods | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SLR-TT | | | CSSP | | | SLR-TT | | | CSSP | | |
| | | r=.5 | r=1 | r=2 | r=.5 | r=1 | r=2 | r=.5 | r=1 | r=2 | r=.5 | r=1 | r=2 |
| Strong Trend of Exposure Prevalence | | | | | | | | | | | | | |
| 2 | 1 | .052 | .049 | .049 | **.127** | **.116** | **.113** | 4.529 | 5.210 | 6.223 | 4.487 | 5.209 | 6.138 |
| | 1.5 | .481 | .476 | .460 | .493 | .481 | .478 | 4.211 | 5.010 | 5.980 | 4.209 | 5.053 | 5.991 |
| | 2 | .728 | .711 | .685 | .657 | .612 | .601 | 2.993 | 3.985 | 5.102 | 3.401 | 4.355 | 5.366 |
| | 2.5 | 1 | 1 | .991 | .837 | .802 | .793 | 2.367 | 3.206 | 3.981 | 2.647 | 3.690 | 4.351 |
| 4 | 1 | .049 | .047 | .046 | **.189** | **.168** | **.150** | 4.561 | 5.412 | 6.311 | 4.424 | 5.210 | 6.126 |
| | 1.5 | .481 | .477 | .462 | .505 | .499 | .484 | 4.422 | 5.340 | 6.080 | 4.168 | 4.943 | 5.941 |
| | 2 | .724 | .709 | .701 | .733 | .725 | .699 | 3.302 | 3.994 | 5.137 | 3.329 | 3.954 | 5.114 |
| | 2.5 | 1 | 1 | .980 | .844 | .820 | .809 | 2.514 | 3.366 | 4.205 | 2.988 | 3.769 | 4.310 |
| Week Trend of Exposure Prevalence | | | | | | | | | | | | | |
| 2 | 1 | .049 | .049 | .048 | **.114** | **.092** | **.089** | 5.257 | 6.402 | 7.124 | 5.016 | 6.101 | 7.087 |
| | 1.5 | .298 | .290 | .277 | .390 | .377 | .363 | 4.810 | 5.991 | 6.796 | 4.625 | 5.730 | 6.419 |
| | 2 | .431 | .419 | .410 | .591 | .579 | .554 | 4.106 | 4.857 | 5.987 | 3.891 | 4.657 | 5.578 |
| | 2.5 | .788 | .779 | .765 | .804 | .791 | .782 | 3.354 | 4.056 | 4.818 | 3.102 | 3.924 | 4.927 |
| 4 | 1 | .048 | .047 | .047 | **.141** | **.123** | **.116** | 5.710 | 6.617 | 7.218 | 5.317 | 6.352 | 7.204 |
| | 1.5 | .298 | .287 | .282 | .397 | .388 | .370 | 5.340 | 6.125 | 6.981 | 4.890 | 5.896 | 6.715 |
| | 2 | .432 | .414 | .407 | .607 | .591 | .573 | 4.284 | 4.946 | 6.082 | 4.033 | 4.882 | 5.919 |
| | 2.5 | .772 | .766 | .752 | .820 | .803 | .794 | 3.509 | 4.277 | 5.161 | 3.216 | 4.049 | 5.002 |

Table 4.2: Rejection rates and average signal detection periods in the presence of unmeasured confounding for different odds ratios

## 4.5 Application: Detecting the risk of Rofecoxib using Sequential Data

We applied the SLR-TT method to Clinformatics Data Mart Database (OptumInsight, Eden Prairie, MN) to test the risk of rofecoxib on acute acute myocardial infarction(AMI). The data were sampled as follows. We first identified all persons age 18 years or older in Optum who received one or more prescriptions for rofecoxib during the study period from April 1, 2000 through Dec 30, 2004. For each rofecoxib-exposed person episode, we ascertained the first month and the last month of their continuous

enrollment episode (or episodes, for persons with multiple enrollment episodes) during the study period. Thus, the unit of observation was the enrollment episode, defined as a period of continuous enrollment for a person. A person could contribute multiple episodes. For each rofecoxib-exposed episode, we randomly sampled, without replacement, nine rofecoxib-unexposed enrollment episodes with an enrollment start date on or before no more than one year of the rofecoxib-exposed subjects enrollment start date, and with an enrollment end date on or after the rofecoxib-exposed subjects enrollment end date. The rationale for this criterion was to ensure sufficient overlap in follow-up calendar time for exposed and unexposed subjects. Thus, the analysis set contained ten times as many total episodes as there were rofecoxib-exposed enrollment episodes. This was done to improve computational efficiency versus including the entire study population.

We fit a logistic regression to estimate the CPE using age, sex, diagnosis of rheumatoid arthritis, and diagnosis of osteoarthritis as explanatory variables. For rofecoxib-exposed subjects, these covariates were measured at their first prescription date. For control subjects, these covariates were measured the same date as their corresponding exposed subjects. The c-statistic was 0.608 and the estimated coefficients and standard deviations (as shown in parentheses) are 0.0228 (0.0001) for continuous age, 0.1458 (0.0027) for female sex, 2.4418 (0.0124) for rheumatoid arthritis, and -0.6444 (0.0191) for osteoarthritis. We then stratified the study population into five strata based on the estimated CPE quintiles.

Table 4.3 indicates that the proposed SLR-TT rejects the null hypothesis that Rofecoxib has a potitive risk on AMI using sequentially available data at the third calendar periods. The results are consistent for all three $\alpha$-spending functions we are considering, even though the critical values obtained via the Monte Carlo approximation are slightly different.

However, the comparison method CSSP fails to detect the signal as the lowest conditional probability that occurs at the forth calendar period is 0.1867, larger than the overall Type-I error rate of 0.05.

| Calendar Quarter | 1 | 2 | 3 | 4 | 5 | ... |
|---|---|---|---|---|---|---|
| TT Likelihood Ratio | 0.6534 | 1.5946 | 2.2753 | 2.9431 | 2.2654 | ... |
| Critical Value for r= .5 | 1.2718 | 1.6517 | 1.8669 | 1.9829 | 2.0164 | ... |
| Critical Value for r =1 | 1.4763 | 2.0165 | 1.9829 | 2.0531 | 2.1782 | ... |
| Critical Value for r= 2 | 1.9523 | 2.1126 | 2.1126 | 2.2651 | 2.4497 | ... |
| Reject the Null | No | No | Yes | Yes | Yes | ... |

Table 4.3: SLR-TT rejects the null hypothesis of no adverse events of AMI at the third calendar period for all three $\alpha$-spending functions

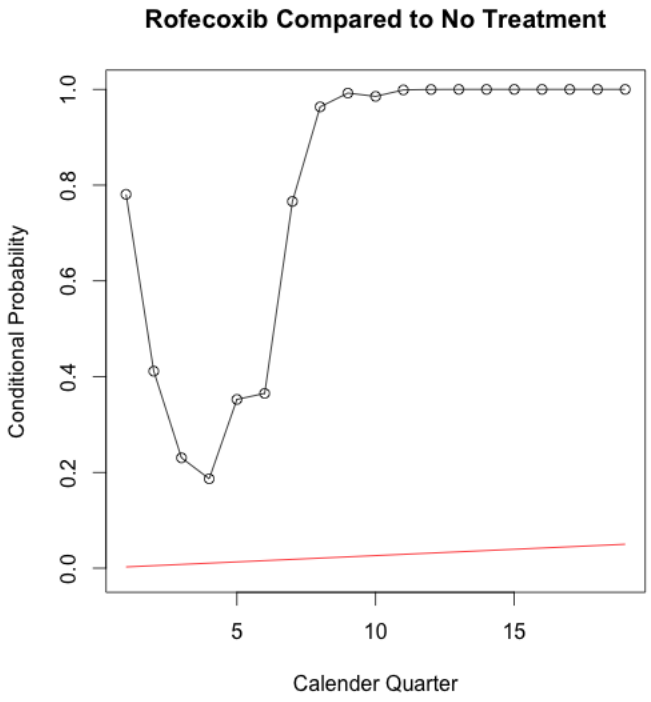



Figure 4.2: CSSP fails to reject the null hypothesis of no adverse events of AMI for all three $\alpha$-spending functions as the lowest conditional probability is 0.1867 at the forth time interval

# 5

# What Do IV Estimates Mean when A Long Gap Exists between IV Creation and Treatment: Implication from Mendelian Randomization

## 5.1 Introduction

Mendelian randomization is a method of using genetic variants as instruments to examine the causal effect of a modifiable environmental exposure on the risk of disease in epidemiological studies. Since Mendelian genes are inherited randomly and independently at conception, they are independent of confounding factors that are hard to be accurately accessed. Genetic variants usually remain constant throughout an individual's lifetime, thus are not affected by occurrences of diseases and changes in exposure levels. As a results, reserve causation, which generally distorts causal estimates, can be avoided using the Mendelian randomization Smith and Ebrahim (2003). Validity of the method relies on one crucial yet untestable assumption, that is there is no direct effect of Mendelian genes on disease nor any other mediated effect

*Joint work with Dylan Small

other than through the exposure of interest. This assumption, known as the exclusion restriction assumption, needs to be justified using background knowledge of the underlying biology and may incur a sensitivity analysis if necessary.

Benefiting from the development of a sophisticated range of instrumental variables methods in econometrics, the Mendelian randomization has been widely applied to cross-sectional and pooled panel data to estimate causal effects. One of the most well-known estimation techniques is the generalized method of moments (GMM), introduced by Hansen (1982), and the two-stage least squares (2SLS) regression is often adopted as a special case of GMM when both responses and exposures are continuous.

Causal conclusions drawn from the Mendelian randomization have interpretations as changes in a response resulted from hypothetical interventions imposed on an exposure. For instance, in a study that confirmed significant causal relationships between body mass index (BMI) and blood pressure, FTO and MC4R genotypes were used as instruments using data were collected from 37,027 unrelated individuals in Denmark Timpson et al. (2009). The study concluded that, if a person were able to reduce BMI by 10%, she would decrease systolic blood pressure by 3.85 mm Hg (95% CI:1.88 to 5.83 mm Hg, P=0.0002) and diastolic blood pressure by 1.79 mmHg (95% CI: 0.68 to 2.90 mmHg, P=0.002). Such interpretations correspond to difference between counterfacturals in the potential outcome framework developed by Rubin (1974b), which makes sense for static and discrete data. However, if we take a hard look at the underlying data generative process instead of the format of the observed data, we see that both BMI and blood pressure are more naturally to be considered as time-continuous variables. In the broader context of many epidemiological studies, biological variables of interest are often inherently dynamic yet observed discretely by design of experiments or data collection procedures. That being said, most cross-

sectional and pooled panel data that the Mendelian randomization is applied to are actually discrete snapshots of dynamic processes, in which (1) the outcome, the exposure, and covariates exhibit non-negligible serial correlations; (2) the outcome is an important determinant of future exposure levels either directly or indirectly through other unmeasured factors; (3) treatment levels are modifiable but can only change gradually in a continuous fashion.

When observations are inherently dynamic, conclusions from the Mendelian randomization are hard to be interpreted. Even worse is that the conventional exclusion restriction assumption is ungrounded in dynamic settings so that estimators derived based on static analysis could be severely biased from what they were intended to estimate. To illustrate this point, we consider a simple scenario in which there exist serial correlations, feedbacks, and unmeasured confinding, and approximate the dynamics using measurements taken at equally-spaced infinitesimal intervals as shown in Figure 5.1. We use $Y_t, D_t, C_t, t \in \{1, 2, \ldots, T\}$ to denote repeatedly observed outcomes, exposure levels, and confounders respectively. The exclusion restriction assumption asserts that the effect of the instrument $Z$ on the outcome only goes through the endogenous exposure whereas Figure 5.1 depicts multiple pathways from the instrument $Z$ to the last observed outcome $Y_T$ going through either $Y_{T-1}$, $D_T$, or $D_{T-1}$. When data are inherently dynamic, the exclusion restriction assumption is sensible only in the conditional sense, i.e., conditioned on previous exposure levels and outcome history, the instrument $Z$ affects the current outcome $Y_T$ only through the current exposure level $D_T$. Such conditional assumption is not implementable on discrete observations using adjustment techniques like stratification or matching as units in a study are almost impossible to have exactly the same historical trajectories.

As a results, the Mendelian randomization may distort the true causal relationships without a careful modeling of the entire dynamic system.

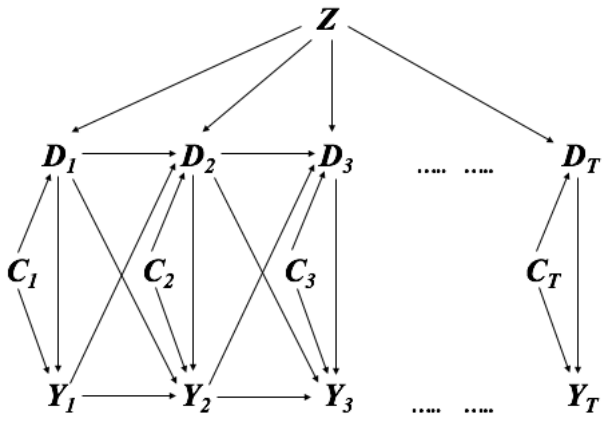



Figure 5.1: Illustration of multiple pathways from the instrument to the outcome

In the literature, many papers have addressed limitations of the Mendelian randomization, most of them are centered around genetic heterogeneity, the lack of suitable polymorphisms for studying modifiable exposures of interest, and confounding of genotype such as intermediate phenotype-disease associations Smith and Ebrahim (2003); Lawlor et al. (2008); Nitsch et al. (2006). But its applicability to dynamic or even longitudinal data has not received a lot of attention, at least to our knowledge. We think it's important to delve into this problem as many biological variables are indeed dynamic processes. We shall take a mechanistic view by modeling reactions among variables using stochastic differential equations and give an explicit formula of the 2SLS estimator derived from discrete observations of the system. We conclude that the 2SLS estimator is a biased estimator of the immediate causal effect, which corresponds to the usual constant treatment effect in static settings.

The paper is organized as follows. In Section 5.2, we review the main idea and limitations of the existing methods to provide practical insights into the potential fallacy of the Mendelian randomization applied to dynamic data. In section 5.3, we take a mechanistic view starting from generalizing notation and assumptions to dynamic data in the presence of a static instrument. In particular, we extend the

exclusion restriction assumption and relate it to local independence, a concept long existed in the literature of graphical models. We use a linear dynamic system, whose time-local presentation is described by a system of stochastic differential equations, to depict the causal mechanism in the Mendelian randomization. In Section 5.4, we derive the distribution of discrete measurements and contrast it with the true dynamic mechanism. In Section 5.5, we investigate whether the Mendelian randomization is applicable to discrete observations generated by the dynamic system. We show that discrete observations of a s time-continuous process generally obscure the underlying local independence between the outcome and the instrument. Hence, applying the Mendelian Randomization to discrete observational data without explicit time justification, could give insensible conclusions. We provide main results and simulated examples. We find that the 2SLS estimator can be used to test for whether the immediate causation is significantly different from zero but does not gauge its magnitude. Based on our derivation, the 2SLS estimator has a form that involves the truth immediate causation as well as many other properties of the dynamic model. In the end, we provide conclusion and discussion.

## 5.2 Review and Limitations of Static Models

The idea of using Mendelian genes as instrumental variables can be viewed as carrying out a randomized encouragement experiment to estimate the causal effect of an exposure as an alternative to measure all confounders Holland (1988). In the simplest case, each unit $i$ is associated with a quadruple $(Y_i, D_i, Z_i, U_i)$, where

- $Y_i$ is the observed response or outcome variable.
- $D_i$ is the observed exposure or treatment assignment.

- $Z_i$ is the randomly assigned encouragement intervention that encourages unit $i$ to experience a higher exposure level or enroll in the treatment group.

- $U_i$ is the joined effect of all excluded variables. $U_i$ is independent of $Z_i$ since $Z_i$ is randomized by the study design.

To quantify the causal effect of $D_i$ on $Y_i$, Holland (1988) proposed the following additive linear constant effects model and $\beta$ represents the average causal effect of interest.

$$Y_i = \alpha + \beta D_i + \rho Z_i + U_i \tag{5.1}$$

If $U_i$ is correlated with $D_i$, i.e. there are unmeasured confounders that affect both the response and the exposure, coefficients on $D_i$ and $Z_i$ from least squares regression are generally biased estimates of $\beta$ and $\rho$. In order to consistently estimate the additive effect of $D_i$, the direct effect of the encouragement intervention is assumed to be zero,i.e., $\rho = 0$. Under the exclusion restriction, $Z_i$ is independent of $Y_i - \beta D_i$. One can test $H_0 : \beta = \beta_0$ using a Wilcoxon rank sum test comparing $Y_i - \beta_0 S_i$ for $Z_i = 1$ and $Z_i = 0$ and estimate $\beta$ using the generalized method of moments (GMM) introduced by Hansen (1982). In particular, the two-stage least squares (2SLS) regression is commonly adopted as a special form of GMM when distributions of both $Y_i$ and $D_i$ are reasonably normal and homoscedastic.

However, in many epidemiological contexts, $Y_i$ and $D_i$ are usually time-continuous. For example, both BMI and blood pressure associated with an individual are dynamic processes. One can apply the model 5.1 to a single-time measure or pooled repeated measurements of BMI and blood pressure by justifying that the additive effect $\beta$ is unchanged over the course of study period. However, we argue that coefficients obtained from such static analysis is not interpretable and the model together with

its assumption is unrealistic given the underlying dynamics.

First of all, it's unlikely for a biological exposure like BMI to change by a certain amount in an instant and, if causal, it takes time for changes in blood pressure to be observed. This means that $Y_i$ has to be measured with some time lag $\Delta t$ after $D_i$ is measured and the interpretation of the additive effect $\beta$ corresponds to the change of $Y_i$ that takes $\Delta t$ for one unit change in $D_i$ to take effect. The size of $\Delta t$ is an important property of the additive effect but is not incorporated explicitly in the model 5.1.

Second, time-continuous variables usually exhibit non-negligible serial correlations and the response can be an important determinant of future exposure levels either directly or indirectly through other excluded variables. In the BMI and blood pressure example, both variables are naturally auto-correlated. The level of blood pressure could either promote or suppress the weight trajectory of an individual. On the one hand, people with severe hypertension tend to limit physical activity to avoid a short time increase in blood pressure. On the other hand, people with moderate level of hypertension are likely to eat a healthy diet, limit the amount of alcohol intake, and exercise regularly in order to lower blood pressure, which either directly or indirectly result in weight loss. The model 5.1 is limited to static settings and does not capture possible complicated reactions between current and previous values, and between current outcome and future exposure levels.

Third, the exclusion restriction assumption is ungrounded for time-continuous variables. Since $Z_i$ is associated with the exposure, it is correlated with the entire history of exposure levels before $D_i$ is measured. $Z_i$ could therefore affect $Y_i$ through the entire history of exposure levels, which is on top of the additive effect of $D_i$.

Therefore, we are motivated to take a mechanistic perspective by proposing time-continuous models that incorporate serial correlations and possible feedbacks for

Mendelian randomization analyses. We will show in the next section that parameters in our time-continuous models have straightforward causal interpretations and the exclusion restriction assumption is modified to "local" exclusion restriction assumption that is sensible for variables that are inherently dynamic.

## 5.3 Time-continuous Models for Mendelian Randomization Analyses

### 5.3.1 Additive Linear AR(1) Models

We first extend the model 5.1 to the following additive linear AR(1) model by adding first-order lagged terms and specifying how the exposure is affected by the previous response. Parameters and error terms are functions of $\Delta t$ representing additive effects that take $\Delta t$ to take effect from the corresponding variables.

$$Y_{t+\Delta t,i} = \alpha_1(\Delta t) + \beta_{11}(\Delta t)Y_{t,i} + \beta_{12}(\Delta t)D_{t,i} + \rho(\Delta t)Z_i + \epsilon_i(\Delta t) \tag{5.2}$$

$$D_{t+\Delta t,i} = \alpha_2(\Delta t) + \beta_{12}(\Delta t)Y_{t,i} + \beta_{22}(\Delta t)D_{t,i} + \gamma(\Delta t)Z_i + \eta_i(\Delta t) \tag{5.3}$$

Certain functional restrictions are required to equalize both sides of the equations 5.2 and 5.3 when $\Delta t = 0$, which are $\alpha_1(0) = \alpha_2(0) = \beta_{12}(0) = \beta_{22}(0) = \rho(0) = \gamma(0) = \epsilon_i(0) = \eta_i(0) = 0$ and $\beta_{11}(0) = \beta_{21}(0) = 1$. This additive linear AR(1) mode reduces to the simple additive linear constant effects model 5.1 if $\beta_{11}$ is constantly zero and if $\Delta t$ is not explicitly expressed.

Assumptions are needed so that $Z_i$ is still a "valid" instrument in the extended AR(1) model. The exogenous condition of $Z_i$ requires that $Z_i$ is independent of $\epsilon_i(\Delta t)$ and $\eta_i(\Delta t)$ for any $\Delta t$, which is reasonable if $Z_i$ is randomized prior to the start of the dynamic processes. As the Mendelian gene is inherited independently and randomly

at conception for each individual, this assumption holds in Mendelian randomization analyses. The relevance condition is extended to the statement that $\gamma(\Delta t) \neq 0$ for any $\Delta t$ so that $Z_i$ is associated with the exposure throughout the entire processes.

A naive translation of the exclusion restriction assumption in Holland's model 5.1 is that $\rho(\Delta t) = 0, \forall \Delta t$. However, such condition is strong and contradicts our previous arguments that $Z_i$ could affect the outcome through the exposure history between $t$ and $t+\Delta t$ for a sizable $\Delta t$. As an alternative, we find that a weakened condition, which we call "local exclusion restriction", is better suited to restrict the direct effect from $Z_i$ on the outcome during $\Delta t$ and paves the way for eventual mechanistic modeling when $\Delta t$ shrinks to zero.

### 5.3.2 Local Exclusion Restriction

The local exclusion restriction states that

$$\frac{\rho(\Delta t)}{\Delta t} \to 0, as\ \Delta t \to 0 \tag{5.4}$$

In words, both the direct effect of the instrument on the outcome over the course of $\Delta t$ and the rate of the change in its magnitute converge to zero. For example, $\rho(\Delta t) = (\Delta t)^2$ conforms to the local exclusion restriction but is non-zero for a sizable $\Delta t$.

The local exclusion restriction is related to the concept of local independence formulated by Aalen (1987) for dynamic modeling of causality. The main idea of local independence is that the intensity of one type of event is independent of certain past events once we know about specific other past events and observed covariates) while the local exclusion restriction basically asserts that the dependence of $Y_{t+\Delta t,i}$ on $Z_i$ given $D_{t,i}$ converges to zero in a faster rate than the diminishment of time interval

$\Delta t$. Local independence has been applied to graphical models Didelez (2007, 2008) and time series models Eichler (2007); Eichler and Didelez (2010) and received much attention in the literature. But local exclusion restriction has never been considered in IV models yet is crucial for Mendelian randomization analyses to quantify the local behavior of how the Mendelian gene affects the outcome through the exposure.

### 5.3.3 Dynamic Models as Limits of AR(1) Models

Given the local exclusion restriction assumption, we assume further that all parameters (not including the error terms) are differentiable at $\Delta t = 0$ and decrease $\Delta t$ to zero. Specifically, we subtract off $Y_{t,i}$ and $D_{t,i}$ on both sides of equations 5.2 and 5.3 respectively and take the limit as $\Delta t$ goes to an infinitesimal time interval $dt$.

$$Y_{t+\Delta t,i} - Y_{t,i} = \alpha_1(\Delta t) + (\beta_{11}(\Delta t) - 1)Y_{t,i} + \beta_{12}(\Delta t)D_{t,i} + \rho(\Delta t)Z_i + \epsilon_i(\Delta t) \quad (5.5)$$

$$D_{t+\Delta t,i} - D_{t,i} = \alpha_2(\Delta t) + \beta_{12}(\Delta t)Y_{t,i} + (\beta_{22}(\Delta t) - 1)D_{t,i} + \gamma(\Delta t)Z_i + \eta_i(\Delta t) \quad (5.6)$$

Given the boundary conditions that $\alpha_1(0) = \alpha_2(0) = \beta_{12}(0) = \beta_{22}(0) = \rho(0) = \gamma(0)$ and $\beta_{11}(0) = \beta_{21}(0) = 1$, limits of parameters in equations 5.5 and 5.6 are the first order derivatives at $\Delta t = 0$ multiplied by $dt$. We also impose distributional assumptions on $\epsilon_i(dt), \eta_i(dt)$ so that they are normally distributed and correlated with each other. In particular, we write the error terms as linear transformations of changes in two independent Wiener processes $W_t^1$ and $W_t^2$. It follows that the limit of the additive linear AR(1) model is a system of stochastic differential equations (SDEs) with parameters corresponding to instantaneous additive effects.

$$dY_{t,i} = a_1 dt + b_{11}Y_{t,i}dt + b_{12}D_{t,i}dt + s_{11}dW_t^1 + s_{12}dW_t^2 \quad (5.7)$$

$$dD_{t,i} = a_2 dt + b_{21}Y_{t,i}dt + b_{22}D_{t,i}dt + rZ_i dt + s_{21}dW_t^1 + s_{22}dW_t^2 \quad (5.8)$$

in which $\alpha_1'(0) = a_1, \beta_{11}'(0) = b_{11}, \beta_{12}'(0) = b_{12}, \alpha_2'(0) = a_2, \beta_{21}'(0) = b_{21}, \beta_{22}'(0) = b_{22}, \gamma'(0) = r$. It is noteworthy $\rho'(0) = 0$ by the local exclusion restriction so that the instrument $Z_i$ does not affect $dY_{t,i}$.

From now on we shall use equations 5.7 and 5.8 to model the data generative process for time-continuous response and exposure in the presence of a static instrument variable. By our arguments, the newly proposed model is more suitable to describe the relationship between BMI and blood pressure and may overturn the causal conclusions in Timpson et al. (2009) drawn from discrete observations. We shall elaborate on this point in the next section.

Before moving on to inference using discrete observations, we want to clarify our contributions of proposing the above dynamic model in the form of SDEs. The use of SDEs for modeling biological variables has been long existed in the literature because SDEs resemble many natural laws in biology and medicine. A few examples are the relationship between the number of CD4 cells and virus concentrations in HIV infection Røysland et al. (2012) and biological pathways for cell signaling Perelson (2002); Bardwell et al. (2007). The causal interpretation of SDEs has also been investigated previously by Sokol and Hansen (2014) using post-intervention equations resulting from a perturbation in the functional form. Sokol and Hansen (2014) showed that, under regularity conditions, the solution to a post-intervention SDE is the limit of a sequence of interventions in Euler structural equation models, which are discrete approximations of the preintervention (observational) SDE. As a result, prescribing a hypothetical perturbation of a dynamic system has a counterfactual interpretation situation as in discrete settings. However, no researchers have ever related SDEs to the original encouragement design that defines a valid instrumental variable. Our derivation of equations 5.7 and 5.8, as limits of AR(1) models, which are extension of Holland's constant additive effect models, reveals a natural way to employ instrumen-

tal variables methods on dynamic data and, at the same time, exposes the potential fallacy of drawing causal conclusions from discrete observations.

## 5.4 Inference from Discrete Measurements

Parameter estimation for SDEs has been highly tackled in several areas of mathematics and statistics, often motivated by financial applications; for reviews, see Sørensen (2004); Bishwal (2008). Estimation methods are centered around constructing approximations of the continuous-time observation likelihood, which require high frequency data with a small time step between two successive observations Pedersen (1995). Unfortunately, data in epidemiological studies and medical fields are usually collected sparsely, in the time unit of month, calendar quarter, or even year. As a result, parameters in equations 5.7 and 5.8 are generally not identifiable from sparse discrete observations. From now on, we assume that discrete observations are snapshots of dynamic processes generated by equations 5.7 and 5.8 and investigate the possibility of drawing sensible causal conclusions without estimation of their parameters.

### 5.4.1 Distribution of Discrete Measurements

We first compute distributions of discrete observations by aggregating $dY_{t,i}$ and $dD_{t,i}$ using stochastic calculus. To ease computation, we write $X_{t,i} = (Y_{t,i}, D_{t,i}, Z_i)^T$ and $dX_{t,i} = (dY_{t,i}, dD_{t,i}, 0)^T$. The dynamics of $X_{t,i}$ is

$$dX_t = \mu dt + AX_t dt + \sigma dW_t \tag{5.9}$$

where

$$\mu = \begin{bmatrix} a_1 \\ a_2 \\ 0 \end{bmatrix}, \quad A = \begin{bmatrix} b_{11} & b_{12} & 0 \\ b_{21} & b_{22} & r \\ 0 & 0 & 0 \end{bmatrix}, \quad \sigma = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \\ 0 & 0 \end{bmatrix}, \quad dW_t = \begin{bmatrix} dW_t^1 \\ dW_t^2 \end{bmatrix} \tag{5.10}$$

We call $A$ the transition matrix as it describes how local characteristics of $X_{i,t}$ depend functionally on past values. It is noteworthy that the upper right corner of $A$ is zero, which restricts the immediate direct effect of $Z_i$ on $Y_{i,t}$ and represents the local exclusion restriction assumption of the instrument. By Ito's isometry, the distribution of $X_{t+\Delta t}$ given $X_t$ can be explicitly computed as a sum of deterministic terms and an integral of a deterministic function with respect to a Wiener process with normally distributed increments.

$$X_{t+\Delta t} = (e^{A\Delta t} - I)A^{-1}\mu + e^{A\Delta t}X_t + \sigma \int_0^{\Delta t} e^{A(\Delta t - s)} dW_s \tag{5.11}$$

The distribution of $X_{t+\Delta t}$ given $X_t$ is thus normal and $E(X_{t+\Delta t}|X_0) = (e^{A\Delta t} - I)A^{-1}\mu + e^{A\Delta t}X_t$, where $e^{A\Delta t} = \sum_{k=0}^{\infty} \frac{A^k(\Delta t)^k}{k!}$. Therefore, associations between the change in $X_{t,i}$ over an infinitesimal interval and the change over an sizable interval $\Delta t$ is as the relationship between the transition matrix $A$ and its exponential form matrix $e^{A\Delta t}$. These two matrices, though functionally related, have drastically different entries and structures. It particular, the upper right corner of $e^{A\Delta t}$ is likely to be non-zero, which means that the local exclusion restriction would be smeared out due to courser observations of the system. Moreover, the error terms as a stochastic integral demonstrates the very complex covariance matrix as a function of $\Delta t$.

The equation 5.11 shows that discrete observations with time step $\Delta t$ follow the

additive linear AR(1) model bellow.

$$Y_{t+\Delta t,i} = \alpha_1(\Delta t) + \beta_{11}(\Delta t)Y_{t,i} + \beta_{12}(\Delta t)D_{t,i} + \rho(\Delta t)Z_i + \epsilon_i(\Delta t) \tag{5.12}$$

$$D_{t+\Delta t,i} = \alpha_2(\Delta t) + \beta_{12}(\Delta t)Y_{t,i} + \beta_{22}(\Delta t)D_{t,i} + \gamma(\Delta t)Z_i + \eta_i(\Delta t) \tag{5.13}$$

All parameters in equations 5.2 and 5.3 are functions of $\Delta t$, representing the additive effects of the corresponding variables that take $\Delta t$ to take effect. The aggregated error terms $\epsilon_i(\Delta t)$ and $\eta_i(\Delta t)$ are still normally distributed but their variances are proportional to $\Delta t$. We pay special attention to parameters of $Y_{t,i}$, $D_{t,i}$ and $Z_i$, which are related to parameters of the dynamic model via the matrix exponential function and reduces to them when $\Delta t$ approaches zero.

$$A = \begin{bmatrix} b_{11} & b_{12} & 0 \\ b_{21} & b_{22} & r \\ 0 & 0 & 0 \end{bmatrix}, \quad e^{A\Delta t} = \begin{bmatrix} \beta_{11}(\Delta t) & \beta_{12}(\Delta t) & \rho(\Delta t) \\ \beta_{21}(\Delta t) & \beta_{22}(\Delta t) & \gamma(\Delta t) \\ 0 & 0 & 1 \end{bmatrix} \tag{5.14}$$

Hence, one may get a different impression of the relationship between the response and the exposure depending on whether one knows the true dynamic structure, represented by A, or one just has the empirical results for a few measurements of the process. We shall use simulations to illustrate this point in the next section.

### 5.4.2 Change of Observational Associations Over Time

The first observation is that $\rho(\Delta t) \neq 0, for\ \Delta t > 0$ and converges to zero in a faster rate than the diminish of $\Delta t$ to conform to the local exclusion restriction of $Z_i$. Such result indicates that the instrument would have a direct effect on the outcome due to courser observations of the dynamic process. The direct effect has nothing to do with the physical law of instrument, but is introduced artificially because of latent

interactions between the response and the exposure during the time interval $\Delta t$.

The second observation is that all parameters in $e^{A\Delta t}$ depend on the size of $\Delta t$, meaning that one may get a different impression of the relationship among $(Y_{t,i}, D_{t,i}, Z_i)$ . We use two numerical examples to investigate patterns of possible changes of $\rho(\Delta t)$ and $\beta_{12}(\Delta t)$, which is usually interpreted as the causal effect of the exposure.

We consider two different transition matrices $A_1$ and $A_2$. Since all nonzero entries of $A_1$ are positive, both $\rho(\Delta t)$ and $\beta_{12}(\Delta t)$ grow exponentially at $\Delta t$ increases, as shown in the left panel of the Figure 5.2. $A_2$ corresponds to a scenario in which the exposure has a positive effect on the response but the increased response reversely reduces the response in a larger magnitude. So the observational additive effect of the exposure on the response $\beta_{12}(\Delta t)$ increases first and then drops to negative values as $\Delta t$ becomes large, as shown in the right panel of the Figure 5.2

$$A_1 = \begin{bmatrix} .2 & .2 & 0 \\ .2 & .2 & .2 \\ 0 & 0 & 0 \end{bmatrix} \quad A_2 = \begin{bmatrix} .1 & .5 & 0 \\ -1 & .1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \tag{5.15}$$

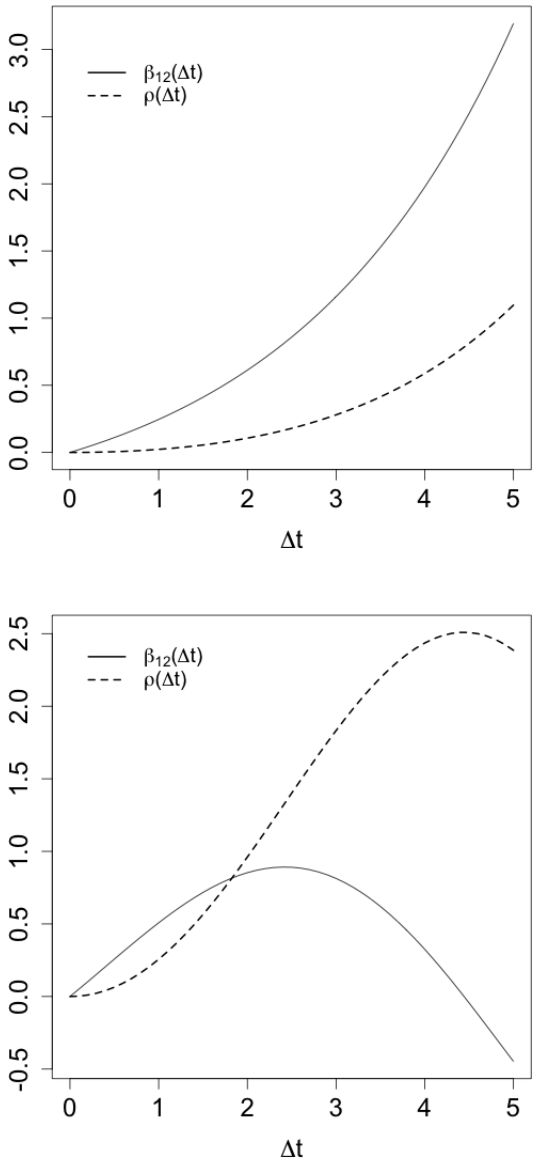



Figure 5.2: Illustration of variations in additive effect of the exposure and direct effect of the instrument over time when the underlying mechanism follows a system of stochastic differential equations. The left plot corresponds to $A_1$ and the right plot corresponds to $A_2$

### 5.4.3 Inference from Equally-spaced Repeated Measurements

Until now, we have made the point of the modeling the cause-effect relationship as a time-local character for time-continuous variables and have discussed how the mechanistic structure given by stochastic differential equations is often distorted when we only get to observe the process at a few time points. We have also mentioned that recovering the dynamic model which characterizes the true continuous cause-effect relationship over time requires quite frequently measured data of a large number of individuals. The questions is how "frequent" is sufficient for different settings.

If the data are discrete measurements at regular times, say anthropometric measurements every 6 months in a clinical trial for hypertension taking place over many years. One could estimate $e^{A\Delta t}$ based on the additive linear AR(1) model 5.2 by regressing $Y_{t+\Delta t,i}$ and $D_{t+\Delta t,i}$ onto $Y_{t,i}$, $D_{t,i}$ and $Z_i$. Theoretically, the obtained estimates are asymptotically unbiased as the error terms are normally distributed with mean zero and independent of all explanatory variables. But variances error terms increase as $\Delta t$ becomes large, implying that a large sample size is needed to obtain narrow confidence intervals of the estimates.

Suppose accurate estimation of $e^{A\Delta t}$ is available given appropriate sampling interval $\Delta t$ and large enough study population, recovering the transition matrix $A$ is still not guaranteed simply because of the mathematical fact that logarithm of a matrix is not an inevitable function. We have shown in the section 3.3 that $e^{A\Delta t}$ and $A$ are likely to have entries with opposite signs, hence we may even draw wrong qualitative conclusions of the relationship between the exposure and the outcome using discrete observations.

## 5.5 Applying Static Analysis to Single Time Measurement

In this section, we shall discuss the most extreme case in which we were only to see the process once and possible consequences of applying the usual static analysis without appropriately adjusting for the inherent dynamics. Suppose variables for each unit are measured at some (unknown) time $T$ after the start of the process. We investigate whether it is still possible to get useful information from ($Y_{T,i}$, $D_{T,i}$, $Z_i$) and what misleading results we would have by improperly applying static models. We shall use 2SLS regression as a demonstrating method.

Following the results in Section 5.4.1, we have

$$Y_{T,i} = \alpha_1(T) + \beta_{11}(T)Y_{0,i} + \beta_{12}(T)D_{0,i} + \rho(T)Z_i + \epsilon_i(T) \tag{5.16}$$

$$D_{T,i} = \alpha_2(T) + \beta_{12}(T)Y_{0,i} + \beta_{22}(T)D_{0,i} + \gamma(T)Z_i + \eta_i(T) \tag{5.17}$$

where $Cov(\epsilon_i(T), \eta_i(T)) \neq 0$ and $(Y_{0,i}, D_{0,i})$ are unknown initial states. Because $Z_i$ is randomized prior to the start of the dynamic process, $Cov(\epsilon_i(T), Z) = Cov(\eta_i(T), Z) = 0$.

### 5.5.1 Detecting Immediate Causation

Equations 5.16 and 5.17 reveal the true relationship between the measured response $Y_{T,i}$ and $D_{T,i}$. If one falsely assume that there is no direct effect of $Z_i$ on $Y_{T,i}$, the estimator obtained using the 2SLS regression is $\frac{\widehat{Cov}(Y_T,Z)}{\widehat{Cov}(D_T,Z)}$, which converges to the following quantity

$$\widehat{2SLS} = \frac{\widehat{Cov}(Y_{T,i}, Z)}{\widehat{Cov}(D_{T,i}, Z)} \rightarrow \frac{\beta_{12}(T)Cov(D_{0,i}, Z_i) + \rho(T)Var(Z_i)}{\beta_{22}(T)Cov(D_{0,i}, Z_i) + \gamma(T)Var(Z_i)} \tag{5.18}$$

If we further assume that the instrument $Z_i$ is independent of the initial state of exposure $D_{0,i}$,the 2SLS estimator estimates the ratio of the direct effect of the instrument on $Y_{T,i}$ to the direct effect of the instrument on $D_{T,i}$, which has no straightforward causal interpretation.

$$\widehat{2SLS} = \frac{\widehat{Cov}(Y_{T,i}, Z)}{\widehat{Cov}(D_{T,i}, Z)} \to \frac{\rho(T)Var(Z_i)}{\gamma(T)Var(Z_i)} \tag{5.19}$$

However, the 2SLS estimator can be used to detect the immediate causation and we have the following theorem.

*Theorem 1.* If the exposure has no immediate causation on the response and the instrument satisfies the local exclusion restriction, i.e., $dY_{i,t}$ does not depend on $D_{i,t}$ and $Z_i$, then neither the exposure nor the instrument has a direct effect on the response at any time.

Intuitively, all pathways from $Z_i$ to $Y_{T,i}$ are blocked although $Y_{T,i}$and $D_{T,i}$ are still correlated due to confounding and reverse causation. *Theorem 1* also implies that if dynamics between the exposure, the outcome, and the instrument follow equations 5.7 and 5.8 with constant parameters and the exposure has no immediate causation on the response, i.e., $b_{12} = 0$, then the 2SLS estimator derived from observations at any time is asymptotically zero.

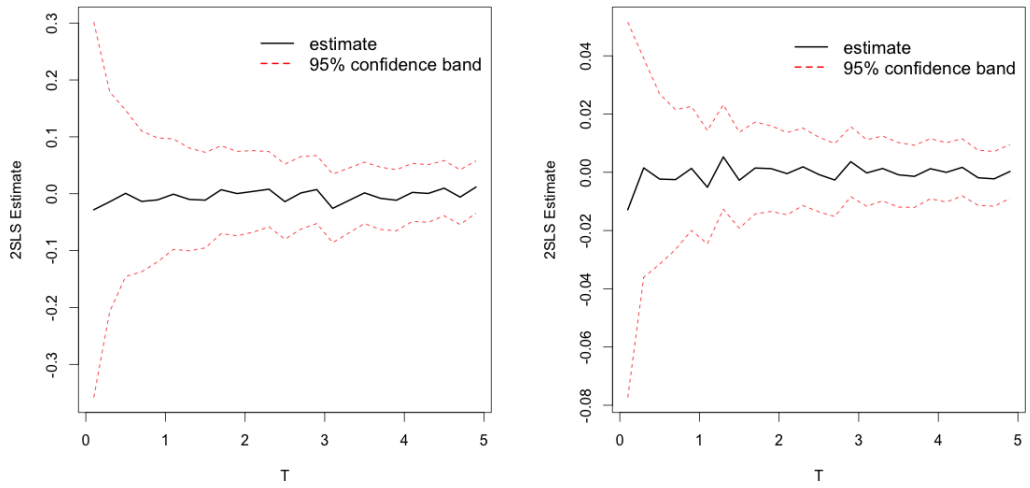



Figure 5.3: Illustration of the 2SLS estimator to detect immediate causal effect. The left and right plots correspond to $A_1$ and $A_2$ respectively with $b_{12} = 0$.

### 5.5.2 Possible Misleading Results

Even though the 2SLS estimator can be used to detect the immediate causation, it does not give a sensible estimate of its magnitude. If $b_{12} \neq 0$, the 2SLS estimator is generally a biased estimator of the immediate causal effect $b_{12}$ and we use simulations to investigate the bias of $\widehat{2SLS}$. We consider two different setups with the following parameters and the corresponding results are summarized in the table 5.1

$$A1 = \begin{bmatrix} .2 & b_{12} & 0 \\ .2 & .2 & .2 \\ 0 & 0 & 0 \end{bmatrix} \qquad A2 = \begin{bmatrix} .1 & b_{12} & 0 \\ -1 & .1 & .1 \\ 0 & 0 & 0 \end{bmatrix} \tag{5.20}$$

| $A$ | $b_{12}$ | T=1 | | | T=5 | | |
|---|---|---|---|---|---|---|---|
| | | $\rho(T)$ | $\gamma(T)$ | $\widehat{2SLS}$ (SD) | $\rho(T)$ | $\gamma(T)$ | $\widehat{2SLS}$ (SD) |
| | .5 | 0.058 | 0.225 | 0.254 (0.066) | 3.141 | 2.745 | 1.129 (0.062) |
| | .75 | 0.087 | 0.227 | 0.375 (0.073) | 5.257 | 3.364 | 1.550 (0.075) |
| $A_1$ | 1 | 0.116 | 0.229 | 0.510 (0.081) | 7.800 | 4.062 | 1.905 (0.089) |
| | 1.25 | 0.146 | 0.231 | 0.638 (0.086) | 10.822 | 4.847 | 2.214 (0.104) |
| | 1.5 | 0.176 | 0.233 | 0.756 (0.095) | 14.382 | 5.726 | 2.497(0.115) |
| | .5 | 0.256 | 0.964 | 0.257 (0.013) | 2.385 | -1.372 | -1.764 (0.019) |
| | .75 | 0.376 | 0.921 | 0.416 (0.014) | 1.419 | -1.956 | -0.676 (0.016) |
| $A_2$ | 1 | 0.491 | 0.881 | 0.546 (0.016) | 0.371 | -1.618 | -0.187(0.019) |
| | 1.25 | 0.601 | 0.841 | 0.705 (0.017) | -0.360 | -0.913 | 0.499 (0.032) |
| | 1.5 | 0.705 | 0.802 | 0.882 (0.020) | -0.645 | -0.171 | 3.709 (0.141) |

Table 5.1: 2SLS estimator as a biased estimator of the immediate causal effect with possibly negative signs

## 5.6 Summary and Discussions

In fields such as medicine, biology, and social science, variables of interest can be time-continuous. Given the same exposure trajectory and starting values, the sample path of the exposure varies between persons and time points. Assuming that the underlying dynamics following s diffusion process, we proved that local exclusion restriction would be smeared out and additional connections with no causal interpretations would be introduced due to courser observations of the system. In particular, the Mendelian randomization with the 2SLS regression cannot be applied to discrete measures to gauge the magnitude of causation. Significance level of the Mendelian randomization, however, can be used to detect the existence of causal relationship between the exposure and the outcome. These complete paths would of course never be available to the researcher who would merely observe the process at a few distinct locations. It is expected that one would benefit from finer observations, but

the time between consecutive measures relative to the length of study needs further investigation. Now we revisit the BMI and blood pressure example that concluded 10% increase in body mass index would increase systolic blood pressure and diastolic blood pressure by 3.85 mm Hg and1.79 mmHg, respectively. Based on *Theorem 1*, the study confirms that BMI affects blood pressure in a cause-effect fashion, but the numerical conclusions on cumulative changes in blood pressure caused by a given percentage change in BMI are not sensible.

Of course the actual mechanism for a biological process can be far more complex than our working model. For instance, parameters can change over time when the system is non-stationary. We opt for the simplest model to demonstrate that associations in discrete observations may be dramatically different from dependencies in the underlying system, therefore inferring causality from discrete observations could be problematic.

# Bibliography


Aalen, O. O. (1987). Dynamic modelling and causality. *Scandinavian Actuarial Journal*, 1987(3-4):177–190.

Asenso-Okyere, W. K., Osei-Akoto, I., Anum, A., and Appiah, E. N. (1997). Willingness to pay for health insurance in a developing economy. a pilot study of the informal sector of ghana using contingent valuation. *Health policy*, 42(3):223–237.

Bardwell, L., Zou, X., Nie, Q., and Komarova, N. L. (2007). Mathematical models of specificity in cell signaling. *Biophysical journal*, 92(10):3425–3441.

Bellan, S. E., Pulliam, J. R., Pearson, C. A., Champredon, D., Fox, S. J., Skrip, L., Galvani, A. P., Gambhir, M., Lopman, B. A., Porco, T. C., et al. (2015). Statistical power and validity of ebola vaccine trials in sierra leone: a simulation study of trial design and analysis. *The Lancet Infectious Diseases*.

Benjamin, E. J., Levy, D., Vaziri, S. M., D'agostino, R. B., Belanger, A. J., and Wolf, P. A. (1994). Independent risk factors for atrial fibrillation in a population-based cohort: the framingham heart study. *Jama*, 271(11):840–844.

Bishwal, J. P. (2008). *Parameter estimation in stochastic differential equations*. Springer.

Braun, T. M. and Feng, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association*, 96:1424–1432.

Brown, C. and Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology*, 6:54.

Brown, J. S., Kulldorff, M., Chan, K. A., Davis, R. L., Graham, D., Pettus, P. T., Andrade, S. E., Raebel, M. A., Herrinton, L., Roblin, D., et al. (2007). Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiology and drug safety*, 16(12):1275–1284.

Cain, L. et al. (2009). Effect of highly active antiretroviral therapy on incident aids using calendar period as an instrumental variable. *American journal of epidemiology*, 169(9):1124 –1132.

Cook, T. D., Campbell, D. T., and Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings*, volume 351. Houghton Mifflin Boston.

Davis, R. L., Kolczak, M., Lewis, E., Nordin, J., Goodman, M., Shay, D. K., Platt, R., Black, S., Shinefield, H., and Chen, R. T. (2005). Active surveillance of vaccine safety: a system to detect early signs of adverse events. *Epidemiology*, 16(3):336–341.

De Allegri, M., Kouyaté, B., Becher, H., Gbangou, A., Pokhrel, S., Sanon, M., and Sauerborn, R. (2006). Understanding enrolment in community health insurance in sub-saharan africa: a population-based case-control study in rural burkina faso. *Bulletin of the World Health Organization*, 84(11):852–858.

Devadasan, N., Ranson, K., Van Damme, W., Acharya, A., and Criel, B. (2006). The landscape of community health insurance in india: an overview based on 10 case studies. *Health Policy*, 78(2):224–234.

Didelez, V. (2007). Graphical models for composable finite markov processes. *Scandinavian Journal of Statistics*, 34(1):169–185.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264.

Dimairo, M., Bradburn, M., and Walters, S. (2011). Sample size determination through power simulation; practical lessons from a stepped wedge cluster randomised trial (sw crt). 12.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. *The American Statistician*, 53(3):177–190.

Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353.

Eichler, M. and Didelez, V. (2010). On granger causality and the effect of interventions in time series. *Lifetime data analysis*, 16(1):3–32.

Ekman, B. (2004). Community-based health insurance in low-income countries: a systematic review of the evidence. *Health Policy and Planning*, 19:249–270.

Fink, G., Robyn, P. J., Síe, and Sauerborn, R. (2013). Does health insurance improve health? evidence from a randomized community-based insurance rollout in rural burkina faso. *Journal of Health Economics*, 32:1043–1056.

Fisher, R. A. (1935). Design of experiments.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15:1069–1092.

Ghosh, M., Mukhopadhyay, N., and Sen, P. K. (2011). *Sequential estimation*, volume 904. John Wiley & Sons.

Govindarajulu, Z. (2004). *Sequential statistics*. World Scientific.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Greevy, R., Silber, J. H., Cnaan, A., and Rosenbaum, P. R. (2004). Randomization inference with imperfect compliance in the aaa randomized trial. *Journal of the American Statistical Association*, 99:7–15.

Hall, A. J., Inskip, H. M., Loik, F., Day, N. E., O'Conor, G., Bosch, X., and Muir, C. S. (1987). The gambia hepatitis intervention study. *Cancer Research*, 47:5782–5787.

Hansen, B. B. and Bowers, J. (2009). Attributing effects to a cluster randomized get-out-the-vote campaign. *Journal of the American Statistical Association*, 104:873–885.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hernández-Díaz, S., Varas-Lorenzo, C., and García Rodríguez, L. A. (2006). Non-steroidal antiinflammatory drugs and the risk of acute myocardial infarction. *Basic & clinical pharmacology & toxicology*, 98(3):266–274.

Ho, D. E. and Imai, K. (2006). Randomization inference with natural experiments: an analysis of ballot effects in the 2003 california recall election. *Journal of the American Statistical Association*, 101:888–900.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*, volume 3. Wiley New York.

Hoel, D., Weiss, G., and Simon, R. (1976). Sequential tests for composite hypotheses with two binomial populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 302–308.

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1).

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

Imbens, G. and Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: an introduction*.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J. M., and Thibaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*, 51(10):5142–5154.

Jennison, C. and Turnbull, B. W. (1997). Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92(440):1330–1341.

Ji, X., Small, D., Leonard, C., and Hennessy, S. (2016). The trend-in-trend research design for causal inference. *Epidemiology*.

Johnston, K., Gustafson, P., Levy, A., and Grootendorst, P. (2008). Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, 27(9):1539 –1556.

Jüni, P., Nartey, L., Reichenbach, S., Sterchi, R., Dieppe, P. A., and Egger, M. (2004). Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *The lancet*, 364(9450):2021–2029.

Kulldorff, M., Davis, R. L., Kolczak, M., Lewis, E., Lieu, T., and Platt, R. (2011). A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential Analysis*, 30(1):58–78.

Lachin, J. M. (1981). Sequential clinical trials for normal variates using interval composite hypotheses. *Biometrics*, pages 87–101.

Lan, K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163.

Lechner, J. (1962). Optimum decision procedures for a poisson process parameter. *The Annals of Mathematical Statistics*, pages 1384–1402.

Lechner, M. et al. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3):165–224.

Li, L. (2009). A conditional sequential sampling procedure for drug safety surveillance. *Statistics in medicine*, 28(25):3124–3138.

Mdege, N. D., Man, M.-S., Taylor, C. A., and Torgerson, D. J. (2011). Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology*, 64:936–948.

Meeker Jr, W. Q. (1981). A conditional sequential test for the equality of two binomial proportions. *Applied Statistics*, pages 109–115.

Meirik, O. (2008). Cohort and case-control studies. *Geneva: World Health Organization*.

Moulton, L. H., Golub, J. E., Durovni, B., Cavalcante, S. C., Pacheco, A. G., Saraceni, V., King, B., and Chaisson, R. E. (2007). Statistical design of thrio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clinical Trials*, 4:190–199.

Mukhopadhyay, N. and De Silva, B. M. (2008). *Sequential methods and their applications*. CRC Press.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. *Statistical Science*, 5:463–480.

Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B. L., Whittaker, J. C., and Leon, D. A. (2006). Limits to causal inference based on mendelian randomization: a comparison with randomized controlled trials. *American journal of epidemiology*, 163(5):397–403.

Pedersen, A. R. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian journal of statistics*, pages 55–71.

Perelson, A. S. (2002). Modelling viral and immune system dynamics. *Nature Reviews Immunology*, 2(1):28–36.

Peskir, G. and Shiryaev, A. N. (2000). Sequential testing problems for poisson processes. *Annals of Statistics*, pages 837–859.

Raz, J. (1990). Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach. *Journal of the American Statistical Association*, 85:132–138.

Rhoda, D. A., Murray, D. M., Andridge, R. R., Pennell, M. L., and Hade, E. M. (2011). Studies with staggered starts: multiple baseline designs and group-randomized trials. *American Journal of Public Health*, 101:2164–2169.

Robyn, P. J., Fink, G., Síe, A., and Sauerborn, R. (2012). Health insurance and health-seeking behavior: Evidence from a randomized community-based insurance rollout in burkina faso. 75:595–603.

Rosenbaum, P. R. Covariance adjustment in randomized experiments and observational studies (with discussion). *Statistical Science*, 17:286–327.

Rosenbaum, P. R. Observational studies.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, pages 41–55.

Røysland, K. et al. (2012). Counterfactual analyses with graphical models based on local independence. *The Annals of Statistics*, 40(4):2162–2194.

Rubin, D. B. (1974a). Estimating causal effects of treatments in randomized and non-randomized studies. 66:688–701.

Rubin, D. B. (1974b). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Sen, P. K. (1968). On a class of aligned rank order tests in two-way layouts. *The Annals of Mathematical Statistics*, pages 1115–1124.

Shih, M.-C., Lai, T. L., Heyse, J. F., and Chen, J. (2010). Sequential generalized likelihood ratio tests for vaccine safety evaluation. *Statistics in medicine*, 29(26):2698–2708.

Small, D. S., Ten Have, T. R., and Rosenbaum, P. R. (2008). Randomization inference in a group randomized trial of treatments for depression: covariate adjustment, noncompliance and quantile effects. *Journal of the American Statistical Association*, 103:271–279.

Smith, G. D. and Ebrahim, S. (2003). 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 29(2):1–22.

Sokol, A. and Hansen, N. R. (2014). Causal interpretation of stochastic differential equations. *Electron. J. Probab*, 19(100):1–24.

Solomon, D. H., Rassen, J. A., Glynn, R. J., Lee, J., Levin, R., and Schneeweiss, S. (2010). The comparative safety of analgesics in older adults with arthritis. *Archives of internal medicine*, 170(22):1968–1978.

Sørensen, H. (2004). Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, 72(3):337–354.

Stratton, K. G. (2012). *Sequential safety monitoring using observational data: A comparison of methods appropriate for newly-licensed vaccines in children.* PhD thesis, University of Washington.

Timpson, N. J., Harbord, R., Smith, G. D., Zacho, J., Tybjærg-Hansen, A., and Nordestgaard, B. G. (2009). Does greater adiposity increase blood pressure and hypertension risk? *Hypertension*, 54(1):84–90.

Tukey, J. (1993). Tightening the clinical trial. *Controlled Clinical Trials*, 14:266–285.

Van der Tweel, I., Kaaks, R., and Van Noord, P. A. (1996). Comparison of one-sample two-sided sequential t-tests for application in epidemiological studies. *Statistics in medicine*, 15(24):2781–2795.

van der Tweel, I. and van der Graaf, R. (2013). Issues in the use of stepped wedge cluster and alternative designs in the case of pandemics. *American Journal of Bioethics*, 13:23–24.

Varas-Lorenzo, C., Riera-Guardia, N., Calingaert, B., Castellsague, J., Salvo, F., Nicotra, F., Sturkenboom, M., and Perez-Gutthann, S. (2013). Myocardial infarction and individual non-steroidal anti-inflammatory drugs meta-analysis of observational studies. *Pharmacoepidemiology and drug safety*, 22(6):559–570.

Vestergaard, P., Rejnmark, L., and Mosekilde, L. (2006). Fracture risk associated with use of nonsteroidal anti-inflammatory drugs, acetylsalicylic acid, and acetaminophen and the effects of rheumatoid arthritis and osteoarthritis. *Calcified tissue international*, 79(2):84–94.

Viera, A. J. (2008). Odds ratios and risk ratios: what's the difference and why does it matter? *Southern medical journal*, 101(7):730–734.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wald, A. (1947). *Sequential Analysis*. New York, Wiley.

Wang, H., Yip, W., Zhang, L., and Hsiao, W. C. (2009). The impact of rural mutual health care on health status: evaluation of a social experiment in rural china. *Health Economics*, 18(S2):S65–S82.

Welch, B. L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika*, 29:21–52.

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

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060.