The Analysis of Cluster-Randomized Test-Negative Designs:
Eliminating Dengue

by

Suzanne M. Dufault

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nicholas P. Jewell, Chair
Professor John Marshall
Professor Ellen A. Eisen

Spring 2020

The Analysis of Cluster-Randomized Test-Negative Designs:
Eliminating Dengue

Abstract

The Analysis of Cluster-Randomized Test-Negative Designs:
Eliminating Dengue

by

Suzanne M. Dufault

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Nicholas P. Jewell, Chair

According to the World Health Organization, dengue is the most critical and most rapidly spreading mosquito-borne viral disease in the world and is responsible for the infection of an estimated 380 million people across the globe annually. There is no cure for dengue, making prevention key to disrupting the rapid progression of this disease into the world's population.

Recent scientific advances target the mosquito's ability to carry and transmit viral diseases. The method motivating this research injects a safe, naturally occurring bacterium called *Wolbachia* into the mosquito population responsible for the spread of dengue and other arboviruses including Zika, chikungunya, and yellow fever. When successfully introduced into the mosquito population, *Wolbachia* prevents these viruses from replicating, which reduces the potential of transmission to humans.

This dissertation addresses the statistical evaluation of the impact of studies of such mosquito-based interventions. Collecting reliable evidence for mosquito-borne interventions is often expensive and logistically prohibitive. The Cluster Randomized Test-Negative Design discussed in this thesis addresses many of the barriers to such vital research. In this trial setting and several variations, I propose and evaluate estimators of intervention impact. These results can be used to better inform policies and protect vulnerable populations.

To my parents – thanks for teaching me that the first step when facing any challenge is to say “I’ll try.”

Contents

List of Figures

List of Tables

Acknowledgments

I must first thank my advisor, Nicholas Jewell, Ph.D. for his tremendous, unwavering mentorship. It has been an absolute joy and honor to learn from one of the field's greatest. You are an inspiration.

Second, I am indebted to the many funding sources that made this work and my doctoral studies a reality. In particular, thank you to the National Institute of Allergy and Infectious Diseases grant R56AI134724. Thank you also to the University of California Dissertation-Year Fellowship program and the faculty and staff who nominated me and helped me prepare my application. I would also like to thank all of the professors who offered me employment as a researcher or instructor, as well as all of the staff who worked alongside them to make sure everything went smoothly. I would especially like to thank Sharon Norris and Janene Martinez for advocating for me every step of the way.

This dissertation would not have been possible without the groundbreaking research being carried out by the World Mosquito Program. Specifically, I am grateful for the opportunity I have had to learn from my Australian collaborators: Cameron Simmons, Ph.D., Katie Anders, Ph.D., and Stephanie Tanamas, Ph.D.

I am further grateful for guidance on this dissertation and the additional research opportunities I have received from the members of my dissertation committee. Ellen Eisen, ScD, thank you for helping me expand on my statistical skill set. Working with your research group has challenged me to grow as a researcher and communicator and has exposed me to some of the most wonderful, brilliant people in academia. John Marshall, Ph.D., thank you for giving me one of my first research opportunities. The dynamics I engaged with in your mathematical modeling courses and lab helped form the basis of my understanding and interest in mosquito-borne infectious diseases.

Over my five years on Berkeley's campus, I have had the pleasure of learning from and working with so many of the School of Public Health's experts. Thank you to Amani Allen, Ph.D., Patrick Bradshaw Ph.D., Sandra McCoy, Ph.D., and Lia Fernald Ph.D.

Thank you to my classmates, colleagues, and friends who have so enriched my Berkeley experience. I am in awe of your accomplishments and intelligence. Thank you for reviewing my work, challenging me, and being all-around great people to spend time with.

Due to the COVID-19 pandemic and subsequent shelter-in-place orders, I am filing this dissertation during one of the most strange and isolating times in collective memory. This forced disconnection from daily routine and expectation has brought into sharp focus the underlying support systems that have pushed me further than I ever imagined was possible. Thank you to my Macalester College professors for the encouragement and guidance that have persisted long past my undergrad years. Thank you, Angela Clem, MScR for being my soul sister and proofreading not just this dissertation but my masters thesis and dozens of application essays. Your skillful eye improves my writing and your wit and friendship improve my life. Thank you to my extended family – Grandma Margaret, Grandpa Loren, Grandma Lucille, and Grandpa Lloyd, in particular. I hope I can continue to make you proud.

Finally, thank you to my parents and siblings. You mean everything to me.

Chapter 1

Introduction

This introduction aims to provide an overview of the global impact of mosquito-borne diseases and the current state of evidence regarding preventative efforts for readers not familiar with the field. Chapters 2 and 3 of this dissertation are comprised of papers that have been published in peer-reviewed journals. Chapter 4 will be submitted for peer review and publication in the near future. Each chapter has been written carefully and is consistent within itself, but notation may differ from chapter to chapter.

1.1 The global impact of mosquito-borne infectious diseases

Mosquitoes transmit diseases that are responsible for millions of human deaths each year, making them one of the most dangerous members of Kingdom Animalia with respect to their impact on human health.[40] *Aedes aegypti* mosquitoes, in particular, are the primary vectors for dengue, Zika, chikungunya, yellow fever and other arboviruses. Though Zika has captured global attention following the 2015 outbreak in Brazil, dengue remains one of the most critical and rapidly spreading mosquito-borne diseases.[13] The estimated number of dengue infections is around 390 million per year. Dengue is endemic in more than 100 countries and is circulating in the Americas, Europe, Africa, and Asia with the majority of the case burden occurring in Asia. It's estimated that roughly 3 billion people are at risk of infection. The true burden of dengue is still unknown given differences in detection and reporting practices around the world, but current measures are likely to be underestimates.[13, 6]

Dengue is caused by four distinct RNA viruses, also referred to as serotypes. Exposure and recovery from one serotype of dengue virus does not confer cross-protection or immunity to the others, allowing for multiple sequential infections. Further, subsequent infection with a second dengue serotype can increase the risk of severe dengue and dengue shock.[52] Additional risk factors for severe dengue infection include high body-mass index, genotype, female sex, and young age with severity of disease differing by dengue virus serotype as well.

The pathogenesis of severe dengue is an active area of research and is complicated by the lack of an adequate animal model.[52]

Upon infection with dengue virus, most people are asymptomatic. For the minority of patients with symptoms, the infection progresses through three clinical phases. The first “febrile phase” spans three to seven days and is characterized by high fever, vomiting, joint and muscle pain, and occasionally the presence of a rash. Most recover after this phase without additional complications. Those who do not enter a one- to two-day “critical phase” in which patients may experience a buildup of fluid in the lungs and abdomen, persistent vomiting, major skin or mucosal bleeding, and low pulse which can lead to shock and death. Those who survive the critical phase enter the “recovery phase” which can consist of a second rash and fatigue spanning several weeks.[52] There are no targeted treatments, such as antivirals, for those infected with dengue, so hospitalization and the timely management of complications is essential to the survival of those with severe dengue.[52, 6]

The World Health Organization has two distinct dengue classifications: dengue and severe dengue. Those who enter and exit the febrile phase without additional complications are classified as dengue cases. The minority of patients who enter the critical phase with complications are classified as severe dengue cases.[52, 13] Severe dengue mortality rates can be greater than 20% if adequate medical care is not available.[13]

In addition to the burden on human health, the disruption caused by outbreaks of mosquito-borne diseases takes an alarming toll on global and national economies. For the 2000-2007 period in the Americas alone, the annual total cost of dengue was conservatively estimated to be US$2.1 billion with a majority of this estimate the result of indirect costs due to productivity losses. Brazil shouldered the majority of this burden, with an estimated US$878 million in annual total costs.[51] In South East Asia, the region with the highest dengue burden, the overall annual total cost was estimated to be US$905 million, with Indonesia experiencing the majority of the economic burden (US$323 million).[50] Note that the reason for the apparent “lower cost” in South East Asia compared to the Americas is primarily due to differences in GDP which result in a higher cost per case for the Americas. In total, billions of dollars are spent annually on direct and indirect costs due to dengue illness. Shephard et al. anticipate that these costs will only increase if current trends continue.[51]

1.2 The current state of research for prevention of mosquito-borne infectious diseases

As there are no effective targeted antivirals for the treatment of dengue, prevention is crucial. Until recently prevention efforts primarily focused on reducing or eliminating the potential of being bitten by a mosquito through direct and indirect interventions. Direct interventions target the mosquito through space-spraying, fogging, the use of repellents or insecticides, and other such methods. Indirect interventions target the environment, primarily through community-based efforts to eliminate potential breeding sites, sanitation improvements, or changes to housing to prevent mosquito-entry.[7]

More targeted, and hopefully more effective, forms of prevention have been under development over the last decade. Recent scientific advancements have made it possible to modify the mosquito itself to either decrease its ability to transmit the disease or disrupt its ability to reproduce, techniques referred to as population replacement and population suppression, respectively. One such approach is through the use of genetic modification via gene drive strategies which aim to introduce and propagate desirable genes into the mosquito population. For example, the introduction of disease-refractory genes could make mosquitoes poor hosts to human diseases, thereby rendering them unable to transmit such diseases.[35] Another approach is through the introduction of the intracellular bacterium *Wolbachia*, which has been shown to render *Aedes aegypti* mosquitoes resistant to the replication and subsequent spreading to humans of arboviral diseases including dengue, Zika, chikungunya, yellow fever, and Japanese encephalitis. While gene drives for dengue prevention have yet to be studied “in the wild”, studies relating *Wolbachia* and dengue incidence are now underway, with positive findings from preliminary studies[26] painting an encouraging picture.

Between the multitude of approaches and the high cost associated with ineffectively preventing dengue illness from circulating, identifying the “best” direct and indirect interventions is critical. A recent systematic review attempted to compile the existing, contemporary evidence to answer, among other critical questions, “[w]hat are the best currently available dengue vector control tools, as measured by their impact on dengue infections...?” In the 41 articles that met the requirements for inclusion in the review, the majority failed to examine the impact of the intervention techniques on dengue incidence itself and instead examined the impact on vector indices which have shown to be imperfect proxies for dengue incidence in the human population. In addition, the prevalent use of weak experimental designs lacking the robustness of randomized trials casts uncertainty on observed associations due to internal and external threats to the validity of such observations. Because of the severe lack of evidence-generating research linking vector control methods to clinical manifestations of disease, the authors were unable to answer their primary question. The need for statistical methods that tie intervention efforts to dengue incidence through thoughtful, feasible study designs is essential for filling the existing gap in evidence generation and analysis — a need this dissertation aims to directly address.

1.3 Dissertation overview

In the following chapters of this dissertation, I develop and compare methodologies for analyzing data to estimate intervention effects from study designs that generate valid evidence and are feasible with respect to common barriers for implementation in mosquito-borne infectious disease research. In particular, Chapter 2 describes the novel cluster-randomized test-negative design (CR-TND), a variant of the gold standard approach for evaluating community-level interventions that minimizes logistical hurdles to enrolling a sufficiently large sample size for valid inference. This study design is the basis of the work presented in

Chapters 2 and 3. Statistical analyses of randomized trial data typically have three primary goals: 1) the accurate estimation of the size of the intervention effect, 2) an estimate of the reliability (i.e. precision) of the estimated size of the intervention effect, and 3) a reliable test to determine whether the size of the intervention is significantly different from what would occur in the absence of an intervention. Meeting each of these goals is essential for understanding the effectiveness of an intervention. Given the novelty of the design, these statistical methods were not previously available. In Chapter 4, we consider the interrupted time series (ITS) quasi-experiment. When a cluster randomized trial is infeasible, the ITS design can be a reliable alternative. In this chapter, we propose a flexible parametric method for simulating highly variable, short time series of dengue fever incidence from historic data. Such simulations can provide insight to statistical, logistical, and practical decision making during trial planning.

Chapter 2

Cluster-Level Estimators of CR-TND Data

This chapter has been published in the journal *Biostatistics*. I specifically focused on simulations and analysis of simulated datasets.[1]

2.1 Introduction: The test-negative design

The test-negative design (TND) is a modification of a case-control study that allows for the use of surveillance systems in assessing the impact of an intervention in reducing disease. The design has been widely used to assess the effectiveness of seasonal influenza vaccines since 2005.[53] The original intent of the TND design was, in large part, to deal with confounding associated with health-seeking behavior that can occur in case-control or cohort designs. However, it can also be seen as providing a viable approach to disease ascertainment in cases where longitudinal studies are likely to be ethically difficult or cost prohibitive.

In the TND, individuals seeking care for symptoms consistent with the disease of interest (but not unique to this disease) are recruited and formally tested for the presence of the specific disease. Those testing positive are then compared with those testing negative with regard to a pre-specified exposure or intervention. As noted, the design is popular for evaluating the effectiveness of seasonal influenza vaccination.[57] Here, patients seeking health care for an acute respiratory illness are recruited into the study and tested for influenza—those confirmed as incident cases of influenza are referred to as test-positives, whereas the others are the test-negatives. In addition, the patient's recent vaccination status is ascertained. In general, assuming that influenza ascertainment was complete, one could estimate influenza incidence proportions for both the vaccinated and unvaccinated population subgroups by dividing the number of influenza cases by the sizes of the vaccinated, and unvaccinated, susceptible and care-seeking populations, respectively, to yield estimates of incidence pro-

[1] I am grateful for permission from my collaborators to include this paper as a part of my dissertation: Nicholas P. Jewell, Zoe Cutcher, Cameron P. Simmons, and Katherine L. Anders.

portions. However, these population sizes can be extremely difficult to determine accurately. The TND uses the relative frequency of the test-negatives in the two exposure groups as a proxy for the ratio of these population sizes. This yields an estimated Relative Risk (RR) based on the ratio of the odds of vaccination in patients testing positive for influenza to the equivalent odds in patients testing negative—see [27], and below, for further details. No rare disease assumption is required.

Because cases and controls are recruited from the same patient population, and restricted to those seeking care at participating clinics, the design was assumed to eliminate bias caused by health-care seeking behavior[27, 19]; however it has recently been demonstrated that this bias is more likely reduced than entirely removed.[58] Several authors have explored the statistical rationale and underlying assumptions of the TND, showing that the Odds Ratio (OR)—i.e. the odds of vaccination in influenza cases versus that for test-negative controls—is directly equivalent to the RR of influenza comparing vaccinated with unvaccinated individuals, providing underlying assumptions are met ([27, 19] etc.). In addition, the causal structure underlying this approach has recently been studied in detail using directed acyclic graphs and causal inference ideas sullivan2016theoretical. For the traditional TND examining influenza, vaccination is applied at the individual level, not always possible for some interventions. This paper discusses TNDs for interventions randomly applied at a group level.

2.2 Cluster-randomized trials and application to test-negative designs

Randomized controlled trials are considered the gold standard for evaluating the efficacy of health interventions, providing the basis of non-model dependent inference. When an intervention is delivered to groups of individuals, e.g. in neighborhoods, or is expected to have a community-wide impact, randomization of the intervention necessarily occurs at the group, rather than individual, level. Such a trial is termed a cluster randomized trial (CRT)—see [22].

Investigators implement a CRT by recruiting a cohort of participants, randomly assigning the intervention to groups of individuals, and following the cohort over time to measure the endpoint in groups assigned to each study arm. The Relative Risk comparing intervention and non-intervention arms is used to quantify the intervention's efficacy, equal to one minus RR. The non-independence of individuals within each group in CRTs causes statistical inefficiency, termed the 'design effect' and inferential methods must appropriately account for the clustering induced by the design ([22], Part C).

While these statistical challenges have been effectively addressed, prospective CRTs frequently require very large cohorts of individuals to generate a sufficient number of events for hypothesis testing, particularly when the outcome is relatively rare (see, for example, [39]). This has significant cost, time, ethical and logistical implications. Trials of preventive interventions against infectious diseases with acute and transient presentations or narrow diagnostic windows face further challenges due to difficulties in obtaining complete and unbiased ascertainment of outcomes within a cohort. Where blinding to intervention status is not possible due to the nature of the intervention or other reasons, this introduces the potential for detection or performance bias[66] if care-seeking or testing behavior (and therefore case ascertainment) is differential by study arm due to participant's and/or health-care providers' perceptions of the intervention's efficacy. Estimation of disease incidence in treated and untreated populations through simple clinic-based surveillance of the disease of interest, without employing the test-negative design, also requires knowledge of the size of the source population from which cases arise. In most settings this cannot be estimated with any accuracy, given that case surveillance is likely to occur in only a subset of the total available healthcare providers and most (potentially care-seeking) populations will have the choice of a number of providers, including in the private sector, both within and outside their local residential area.

Here we propose a method to assess the endpoints in CRTs using the test-negative design that offers the advantage of being more efficient, cost effective, and logistically simpler to achieve than a large prospective cohort, and does not require knowledge of the sizes of populations at risk. We refer to our proposal as a Cluster Randomized Test-Negative Design (CR-TND)—see [1]. The CR-TND fundamentally alters the standard TND in two key ways: (i) randomization of the exposure, and (ii) clustering of participants' exposure status due to randomized delivery of the intervention at a group-level.

The next section describes the motivating application, a preventive intervention against dengue. Section 4 introduces estimation and inferential methods to assess an intervention's efficacy using data arising from the new design, based on summary measures at the cluster level (as compared to individual level data). Section 5 provides exploratory simulation results to assess the power of these approaches along with additional properties of the estimation and inferential methods. Inference is examined in terms of the permutation distribution induced by randomization; some comparisons are made with model-based methods—including Generalized Estimating Equations (GEE) and mixed effects logistic regression—techniques intended to account for within group correlation. Section 6 provides a brief discussion and points towards further research topics.

2.3 Application

The World Mosquito Program is an international research collaboration that is delivering a paradigm shift in the control of arboviral diseases transmitted by *Aedes aegypti* mosquitoes. The method utilizes *Wolbachia*, obligate intracellular endosymbionts that are common in insect species (see, for example, [23]) but were not present in *Aedes aegypti* mosquitoes until they were stably transinfected in the laboratory. In insects, *Wolbachia* is maternally transmitted via the egg and manipulates insect reproduction to favor its own population dissemination via cytoplasmic incompatibility. The result is that *Wolbachia* rapidly enters into naive mosquito populations in a self-sustaining, durable manner. Strikingly, the presence of *Wolbachia* in *Aedes aegypti* mosquitoes renders them more resistant to disseminated arbovirus infection, including dengue, Zika, CHIK and Yellow fever.[14, 30, 45] Thus the critical and signature effect of *Wolbachia* as an intervention is to severely reduce the vectorial capacity of mosquito populations to transmit arboviral infections between humans. For field implementation, the approach seeds wild mosquito populations with *Wolbachia* through controlled releases of small numbers of *Wolbachia* infected mosquitoes.

The motivation for the proposed CR-TND arises from a planned two-year trial to evaluate the efficacy of *Wolbachia*-infected mosquitoes in reducing dengue transmission in Yogyakarta City, Indonesia. The administrative area of the city, with a population in 2015 of 408,000, has a generally higher dengue incidence than surrounding districts. A parallel two-arm non-blinded CR-TND will be conducted. The study site was subdivided into 24 contiguous clusters, each approximately one km^2 in size, to allow for effective deployment while minimizing cluster interference. Clusters were randomly allocated in a 1 to 1 ratio to receive either *Wolbachia*-infected mosquito deployments or no intervention. Eligible febrile participants will be recruited from across the study area through primary healthcare clinics, and subsequently classified as virologically confirmed dengue cases (test-positives) or arbovirus-negative controls on the basis of laboratory testing (i.e. those suffering from other febrile illnesses or OFIs). The *Wolbachia* exposure distribution in test-positive cases (i.e. whether or not the individual lives in an intervention cluster) will be compared with that for test-negative controls, in order to estimate the efficacy of the intervention.

The CR-TND approach has been developed for the Yogyakarta trial in preference to a traditional cohort design, in which absolute and relative dengue incidence in treated and untreated populations is measured directly, because of challenges in reliably ascertaining dengue illness episodes prospectively in a large population and in quantifying true exposure to *Wolbachia* prior to any presenting infection. Passive ascertainment of dengue cases through existing routine disease surveillance systems would provide incomplete, and potentially biased, estimates of the population dengue incidence. The surveillance system in Indonesia, and many endemic settings, captures only hospital admissions, and completeness of reporting may be spatially and temporally variable. Specificity of the surveillance data is also imperfect and variable, as case notification is commonly based on a clinical diagnosis of dengue, with only a subset of cases confirmed by laboratory diagnostics. The alternative approach of actively ascertaining dengue cases within a cohort of individuals recruited from treated and untreated areas is challenging both operationally and ethically, due to the large size of the cohorts required, the intensity of contact required to detect and diagnose acute dengue infections, and the potential for biases arising from under-ascertainment of illness events or loss to follow-up. Further, it would be hard to measure individual mobility and *Wolbachia* levels at the time of any presenting infection without almost continuous monitoring.

2.4 Cluster-level analyses of CR-TND data

Test-positive fraction of all tests by cluster

Suppose there are $2m$ clusters with m randomly assigned to the intervention and the remainder untreated. All test-positive (cases) and test-negative (controls) individuals are selected, numbering n_D, and $n_{\bar{D}} = rn_D$, respectively. The Relative Risk, associated with the intervention, is denoted by λ. The simplest data are cluster counts of confirmed test-positives and test-negatives. As our primary approach is to develop inference based on a permutation approach, we take these numbers as fixed, at this point allowing randomness only to enter through the allocation of intervention status to each cluster. Further, p_{Dj} and $p_{\bar{D}j}$ represent the fraction of cases and controls, respectively, in the j^{th} cluster. Let a_j denote the ratio of the number of test-positives (n_{Dj}) to the total number of tests in the j^{th} cluster ($n_{Dj} + n_{\bar{D}j}$). Note that $a_j = \frac{n_D p_{Dj}}{n_D p_{Dj} + n_{\bar{D}} p_{\bar{D}j}} \equiv \frac{p_{Dj}}{p_{Dj} + r p_{\bar{D}j}}$. A proposed test statistic to assess the effect of the intervention is then $T \equiv \alpha_I - \alpha_C \equiv$ average(a_j|cluster j is intervention) − average(a_j|cluster j is untreated).

Under the null hypothesis of no intervention effect, for a randomly selected cluster (from either arm) $\mathbb{E}_0(p_{Dj}) = 1/2m$ since the p_{Dj}s sum to 1; here the expectation is over the permutation distribution of all possible random allocations and the subscript 0 reinforces that this expectation is under the null. Similarly, $\mathbb{E}_0(p_{\bar{D}j}) = 1/2m$. Thus, $\mathbb{E}_0(T)$ is approximately zero using the delta method. In fact, by the symmetry of its definition, $\mathbb{E}_0(T) \equiv 0$.

We now approximate $\mathbb{E}(T)$ when the intervention affects case counts, changing the relative distribution of cluster test-positives to test-negatives, and how this depends on λ. For a large number of test-positives in the intervention arm, this total is reduced by λ (for $\lambda < 1$). Thus, the fraction of the total number of test-positives that occur in the intervention clusters is approximately $\lambda/(1+\lambda)$. Hence, for a random cluster, j, in the intervention arm, $\mathbb{E}(p_{Dj}) \approx \frac{\lambda}{m(1+\lambda)}$. Similarly, for a random cluster, j, in the untreated arm, $\mathbb{E}(p_{Dj}) \approx \frac{1}{m(1+\lambda)}$. The $p_{\bar{D}j}$ are unaffected by the intervention so that by the delta method,

$$\alpha_I \approx \frac{2\lambda}{(2+r)\lambda + r}, \alpha_C \approx \frac{2}{r\lambda + (2+r)}. \tag{2.1}$$

Thus,

$$\mathbb{E}(T) \approx \frac{2\lambda}{(2+r)\lambda + r} - \frac{2}{r\lambda + (2+r)} = \frac{2r(\lambda^2 - 1)}{[(2+r)\lambda + r][r\lambda + (2+r)]}. \tag{2.2}$$

The RHS of (2.2) is zero if and only $\lambda = 1$, as noted above. When $\lambda < 1$, $\mathbb{E}(T) < 0$, as expected. Further, (2.2) is quadratic in λ for any given $\mathbb{E}(T)$. Thus, we can substitute an estimate of $\mathbb{E}(T)$, obtained from differencing the average observed ratios a_j in each arm, into (2.2) and solve for λ, yielding an approximate estimate of RR. This approach assumes a cluster specific interpretation of the Relative Risk as compared to a marginal version–we return to this point later.

We illustrate with a simple example, with $r = 1$ and $\lambda = 0.5$. Here (2.2) yields $\alpha_I \approx 2/5$ and $\alpha_C \approx 4/7$, so that $\mathbb{E}(T) \approx -6/35$. In reverse, substituting $-6/35$ into the LHS of (2.1) yields $22\lambda^2 + 15\lambda - 13 = 0$ with only one positive solution, namely $\lambda = 0.5$.

Turning to inferential methods based on T, the null hypothesis can be tested via the permutation distribution of T, calculated by examining the estimated Ts for all possible randomized intervention allocations, holding the observed data fixed. This permutation distribution can then be used to assess the significance of the observed value of T.

For simplicity in carrying out simulations, and for additional insight, we can analytically evaluate the null permutation variance of T. As noted, we consider the observed values of a_j for $j = 1, \dots, 2m$ to be fixed. Under the null hypothesis the a_js for the m intervention clusters are simply a random sample of m of these values, also true for the untreated a_js. The variance of the $2m$ fixed values of a_j across all clusters depends on: (i) the variability of both the p_{Dj}s and $p_{\bar{D}j}$s, i.e. the distribution of cluster test-positives and test-negatives, respectively, (ii) the covariance of the p_{Dj}s and the $p_{\bar{D}j}$s, i.e. how test-positives and test-negatives covary across the clusters, and (iii) the value of r. However, we do not need to analytically derive this variance as we will empirically estimate it in due course. We refer to the variance of the $2m$ a_js by $\sigma^2 = \sum_{j=1}^{2m}(a_j - \bar{a})^2/(2m-1)$ where $\bar{a} = \sum_{j=1}^{2m} a_j/2m$.

In the m intervention clusters, the permutation variance of α_I is $(\sigma^2/m) \times \left(\frac{2m-m}{2m}\right)$ where the second term is the finite population correction factor. The variance of T must accommodate that α_I and α_C are correlated due to the finite number of clusters, and the fact that we are conditioning on the observed data and computing expectations and variances according to the permutation distribution. In fact, under the null hypothesis, $m\alpha_I + m\alpha_C = 2m\mu$ (where $\mu = \frac{1}{2m}\sum_{j=1}^{2m} a_j$), so that $\alpha_C = 2\mu - \alpha_I$. Hence, under the null hypothesis, $T = 2(\alpha_I - \mu)$. Finally, therefore,

$$\text{variance}_0(T) = 2 \times (\sigma^2/m). \tag{2.3}$$

Note that, when m is sufficiently large, the RHS of (2.3) is just what you would obtain by naively treating α_I and α_C as independent averages with a common variance σ^2/m.

The variance, σ^2, can be estimated by the variance of the a_js in either the intervention or untreated clusters. Since these arms both contain m clusters, an average of these two estimates suffices–the pooled variance estimator of the two-sample t-test. With this estimate, $\hat{\sigma}^2$ used in (2.3), the standardized test statistic is thus $T/\sqrt{2(\hat{\sigma}^2/m)}$, equivalent to the two-sample t-test statistic comparing the observed a_js across the two arms. Of course, if the null hypothesis is true we would know the variability of the a_js exactly since all would be observed, but this variance would overestimate the variance of T away from the null. We can then compare the standardized test statistic to a t distribution with $2(m-1)$ degrees of freedom to assess significance, and this provides a close approximation to the permutation distribution result so long as m is large.

For any r, the difference in the average cluster means of the a_js across intervention arms estimates $\mathbb{E}(T)$ and thus directly relates to an estimate of λ, through (2.2). As λ moves away from the null value, the absolute size of $\mathbb{E}(T)$ monotonically increases. Hence, in addition to point estimation, a confidence interval for $\mathbb{E}(T)$ directly yields a corresponding confidence interval for λ. Note that, in the scenario with an intervention effect, it is not straightforward to approximate the permutation variance of T. Nevertheless, we recommend treating the average a_j's as if they come from independent samples since this is correct at the null. Away from the null, the variances of α_I and α_C are not equal so that it might be better to base confidence intervals on the Welch version of the t-test, using separate estimates of the variances in the two arms and modifyng the number of degrees of freedom appropriately using the Welch-Satterthwaite formula.[63, 49, 62] We examine the empirical coverage properties of this approach in Section 5. Note that the differentiation between the intervention arms—given by $\mathbb{E}(T)$—gets smaller as r increases from 1. Perversely, this approach is then optimal when $r = 1$ and loses power as r increases. However, this effect is small as examined quantitatively in simulations.

Odds ratios from collated cluster data

Consider now an alternative method to both assess the intervention's efficacy and provide an estimate of λ. Following [27], Table 2.1 is the 2 x 3 table that classifies care-seeking individuals by their intervention status and outcomes, with A_+ the total number of individuals who both experience the intervention, are detected by the surveillance system, and test positive for the outcome of interest with similar definitions for other entries. There is an analogous classification of (unobserved) individuals who do, or would, not seek care when experiencing such infections; the generalizability of any intervention efficacy estimate from observed data in Table 1 to the entire population depends on an untestable assumption that efficacy is not modified by care-seeking behavior—see [27] and [58].

	Seek Care			
	Infected with dengue	Infected with OFI	Not Infected	*Total*
Intervention	A_+	B_+	C_+	N_I
Control	G_+	H_+	I_+	N_C

Table 2.1: Stratification of population based on intervention status, infection, and health care-seeking behavior. OFI refers to other febrile illnesses with similar presenting conditions to the infection of interest that can be discriminated on the basis of a specific laboratory test. Adapted from Figure 1 of Jackson & Nelson (2013).

N_I is the total number of exposed *susceptible* individuals in the population who would seek care if they experience symptoms; an analogous definition describes N_C for controls. In principle, the incidence of the disease outcome in the exposed care-seeking population can then be estimated by A_+/N_I, and RR by $\frac{A_+}{N_I}/\frac{G_+}{N_C}$. Unfortunately, it is often difficult to obtain accurate details for the susceptible care-seeking population sizes N_I and N_C. The TND therefore exploits the incidence of OFIs in the population as a basis for assessing the relative population sizes N_I and N_C, under the assumption that incidence of OFIs is independent of the intervention. The latter assumption means that $\frac{B_+}{N_I} \approx \frac{H_+}{N_C}$ so that $\frac{H_+}{B_+} \approx N_C/N_I$. Thus $(A_+ \times H_+)/(G_+ \times B_+) \approx RR$. In other words, the empirical Odds Ratio from the data of Table 2.1 provides a direct estimate of RR. Note that this is a marginal Odds Ratio with no reference to within cluster characteristics.

In the CR-TND, the two rows of Table 2.1 correspond to data from different clusters since every individual in a given cluster is either exposed to the intervention or not. The observed log Odds Ratio from Table 2.1 is then $\log(A_+H_+/B_+G_+)$. We first assess the properties of this random variable (induced by randomization, keeping the data fixed) under the null hypothesis, i.e. under the permutation distribution. Under random sampling, $\mathbb{E}_0(A_+) = E_0(G_+) = n_D/2$ where $n_D = A_+ + G_+$. Analogously, $\mathbb{E}(B+) = E(H_+) = n_{\bar{D}}/2$. Then, by the delta method, $\mathbb{E}_0(\log(A_+H_+/B_+G_+)) \approx 0$; of course, this expectation is *exactly* 0 because of the symmetry between the counts caused by randomization. Thus, this log(Odds Ratio) is centered at the correct value assuming the null. For testing we can again revert to the full permutation distribution. Again, we can analytically approximate the variance of the cumulative log(Odds Ratio) estimate at the null hypothesis for use in simulations and to provide approximate permutation inference.

As before, under the null, the permutation variance of A_+ is simply $mV_D \times \left(\frac{2m-m}{2m}\right)$ where V_D is the variance of the test-positive counts $A_1, \ldots, A_m, G_1, \ldots, G_m$ across both intervention and untreated clusters (again using $(2m-1)$ in the denominator of the definition of V_D). Here we use A_j, G_j etc to refer to the entries of Table 2.1 specific to the j^{th} cluster. Conditional on the observed data, A_+ and G_+ are not independent under the randomization distribution since $A_+ + G_+ = n_D$. Thus the null variance of $\log(A_+/G_+) \approx (16/{n_D}^2)\text{var}(A_+)$ using the delta method and the fact that $\mathbb{E}_0(A_+) = n_D/2$. Finally, putting these two observations together yields

$$\text{var}(\log(A_+/G_+)) \approx (16/{n_D}^2)(m/2)V_D. \tag{2.4}$$

Similarly,

$$\text{var}(\log(B_+/H_+)) \approx (16/{n_{\bar{D}}}^2)(m/2)V_{\bar{D}}, \tag{2.5}$$

where $V_{\bar{D}}$ is the variance of all $2m$ cluster test-negative counts.

For approximate estimation of the variance of the log(Odds Ratio), note that A_+ and B_+ may be correlated because of characteristics of the clusters that may induce test-positives and test-negatives to tend to be high together (e.g. the population size and density within a given cluster) or possibly negatively associated. We can approximate the covariance of the two terms $\log(A_+/G_+)$ and $\log(B_+/H_+)$ by exploiting the delta method, so that

$$\text{cov}(\log(A_+/G_+), \log(B_+/H_+)) \approx \frac{n_D n_{\bar{D}}}{A_+(n_D - A_+)B_+(n_{\bar{D}} - B_+)}\text{cov}(A_+, B_+). \tag{2.6}$$

Putting (2.4), (2.5) and (2.6) together then yields

$$\begin{aligned}\text{var}(\log(\text{Odds Ratio})) \approx (16/{n_D}^2)(m/2)V_D + (16/{n_{\bar{D}}}^2)(m/2)V_{\bar{D}} \\ -2 \times \frac{n_D n_{\bar{D}}}{A_+(n_D - A_+)B_+(n_{\bar{D}} - B_+)}\text{cov}(A_+, B_+).\end{aligned} \tag{2.7}$$

It remains to estimate $V_D, V_{\bar{D}}$, and $\text{cov}(A_+, B_+)$. As before, under the null, V_D could be estimated by using a variance estimator of the test-positive counts $A_1, \ldots, A_m, G_1, \ldots, G_m$ from both intervention and untreated clusters. However, this yields a poor estimate with an intervention effect. We thus use a pooled estimate, the average of the variances of the $A_1, \ldots, A_m$ and $G_1, \ldots, G_m$, separately estimated (using $m-1$ in the denominator for the estimated variances). $V_{\bar{D}}$ can be estimated analogously. Finally, $\text{cov}(A_+, B_+) = m \times \text{cov}(A_j, B_j) \times \left(\frac{2m-m}{2m}\right)$ using finite population sampling methods—see [59] for the less familiar use for covariances. The term $\text{cov}(A_j, B_j)$ can be estimated from the covariance of the observed $A_1, \ldots, A_m$ and $B_1, \ldots, B_m$ in the intervention clusters, again using $m - 1$ in the denominator of the covariance estimate.

A simple example illustrates the effectiveness of a Gaussian approximation to the permutation distribution (at the null) based on the proposed sample estimate of (2.7). Table 2.6 in Section 2.7 shows an assumed distribution of dengue and OFI cases for 10 clusters (mimicking more complete data used in later simulations). From these distributions, a random set of 100 cases and 100 controls were selected once. There are $252 = \binom{10}{5}$ possible intervention cluster intervention allocations. The exact standard deviation of the permutation distribution of $\log(OR)$, with no intervention effect, is 0.1566. Over the 252 possible intervention allocations the average estimated standard deviation of $\log(OR)$ based on (2.7) is 0.1529, a close approximation. For comparison, for a simple random intercept logistic regression, the average model-based standard deviation estimate is 0.2843. The average estimated robust standard deviation based on a standard GEE (assuming an exchangeable correlation structure - see below) is unreliable at this small of a sample size. However, when 16 of the 252 permutation distributions with unreliably large estimates are removed, the average estimated standard deviation is 0.1407. Further discussion of these two approaches is provided below. For $\log(OR)$, using the 252 possible intervention allocations yields lower and upper 2.5% percentile thresholds of ±0.2870 for the exact permutation distribution; the lower and upper thresholds, induced by ±1.96 × the average standard deviation estimate based on (2.7), are ±0.2997. The performance of GEE improved as sample size increased, resulting in a stable underestimate of the permutation variance; the performance of the mixed effects estimator was more variable and not always as poor as this specific case and also improves as the sample sizes of cases and controls increase. Further simulation evidence of type 1 error rates and power associated with using this approximate Odds Ratio inference approach is given in the next section.

For confidence interval calculations (away from the null) we need to evaluate the randomization distribution of the log(Odds Ratio) estimate assuming an intervention effect. Following [27], note that the intervention only affects the counts $A_1, \ldots, A_m$ by assumption. These are each replaced in turn by $A_1^*, \ldots, A_m^*$ which reflect altered test-positive counts in the intervention clusters. For large populations, $A_j^* \approx \lambda A_j$ for the intervention clusters. Note that this specifically uses the assumption that the intervention effect is the same for all clusters. An alternative approach might model the intervention effect that allows variation with cluster characteristics, or, for example, with the size of A_j itself. We ignore this "second order" phenomenon and leave this for future analysis. The common reduction of the $A_1, \ldots, A_m$ has two immediate implications: first, under the randomization distribution, $\mathbb{E}\left(\log\left(A_+H_+/B_+G_+\right)\right) \approx \log(\lambda)$; second, there is no change to the variance formula (2.7) since all the Odds Ratios for different permutations are simply shifted by approximately $\log(\lambda)$. Despite the remaining usefulness of (2.7), we do now have to modify the estimates of n_D, V_D, and $\text{cov}(A_+, B_+)$ due to the replacement of each Aj with A_j^*. The necessary adjustment is achieved by simply increasing the observed A_j^*s by the common factor $1/\hat{\lambda}$ to obtain an estimate of A_j (in the j intervention clusters), en route to an estimate of n_D, V_D, and $\text{cov}(A_+, B_+)$ as before.

The above approach ignores any random variation of A_j^* around $\lambda \times A_j$. Two potential sampling models might be employed here to account for this additional variation when the counts $A_1, \ldots, A_m$ are smaller due to an intervention effect. The first is to assume that for the intervention clusters, and given A_j, A_j^* is Binomial with probability $\lambda \times A_j$. The alternative is to assume that A_j is Poisson with rate parameter $\lambda \times A_j$. Essentially we are assuming here that given all the test-positive and test-negative counts in all the clusters, the A_js in the intervention clusters are 'filtered' by an additional layer of randomness to generate the observed A_j^*s. The Binomial approach essentially assumes that RR < 1 but this can achieved without loss of generality by switching the intervention label. By first conditioning on all counts (including the unobserved $A_1, \ldots, A_m$ in the intervention clusters), we can see that the variance in (2.7) is subsequently increased approximately by either $(1-\lambda)/(\lambda A_+)$ for the Binomial case, or $1/(\lambda A_+)$ for the Poisson. (The latter is more conservative and therefore recommended.)

Odds ratio estimates via GEE and random effects logistic regression models

The CR-TND design yields clustered binary outcome data where interest may focus on estimation of the Odds Ratio. It is natural, therefore, to consider applying Generalized Estimating Equations (GEE), or random intercept logistic regression methods. For GEE we focus on use of a working exchangeable correlation structure within groups. Both of these methods are well known and easily implementable using standard software, and account for clustering through use of a robust variance method (GEE) or via an appropriate assumed random effects distribution. However, both methods are also known to suffer in performance in situations with small number of clusters, each containing a large number of observations. Also, at the outset, it is important to note that these two approaches are designed to estimate different parameters: GEE focuses on the marginal, or population averaged, Odds Ratio whereas the random effects model targets the cluster-specific Odds Ratio (see, [22], Chapter

9.3). As defined in Section 2.4, the Odds Ratio used to estimate intervention efficacy is a marginal Odds Ratio.

In the simple example above with 10 clusters, when the sample size is small GEE is unreliable in estimating the variability of the log OR; the random effects model also overestimates the relevant standard deviation although significantly less so. As the sample size increases, both methods tend to underestimate the standard deviation. In general, with small numbers of clusters, the GEE technique suffers from inflated Type 1 error rates [4]. Pan and Wall [41] describe approaches to correcting for this though use of a t distribution as reference as opposed to the standard Normal. Morel *and others* [38] suggest an alternative modification for making inferential statements. As pointed out by [22], however, this tactic requires addition calculations to determine the relevant degrees of freedom.

Recent research has examined the impact of small numbers of clusters on point and interval estimation of a fixed effect that covers both GEE and random effects logistic regression. See [36] for an overview. While comparisons differ depending on the context, the general consensus is that GEE performs poorly with small numbers of clusters whereas random effects models provide reasonable inference. We consider both methods in the simulations.

2.5 Simulations

We used simulations that exploited historical information on the incidence of dengue and OFIs in Yogyakarta City. As discussed above, the city was divided into 24 non-overlapping contiguous clusters. The design randomly allocates 12 of these clusters to the intervention, with the remainder left untreated. Dengue incident cases, along with relevant comparative OFIs will be collected over a two-year period. Table 2.7 contains the frequency of recorded (hospitalized) dengue cases in the 24 clusters for each of nine distinct two-year periods covering 2003-2014. During this period, there was no available data for 2004 and 2009, so that the first two-year interval was for 2003 and 2005; similarly the 2008-10 interval included data for 2008 and 2010. Otherwise each two-year period covered consecutive years. Data for the distribution for OFIs (in Table 2.8) is only available for one two-year period from 2014-15.

We carried out simulations assuming that the dengue distribution was exactly as identified for a given time period, but with a consistent OFI distribution across all time periods (the 2014-15 distribution). For each simulation, 1000 dengue cases were assigned to the 24 clusters according to the dengue distribution, and $r \times 1000$ OFIs were assigned according to the OFI distribution. These allocations provided the base data from which we subsequently applied intervention assignment labels to 12 of the 24 clusters at random (according to a permutation distribution) for each distinct simulation. The cluster intervention assignment was permuted 10,000 times at random.

At the null, no further data modifications were required in computing various estimates and test statistics. Away from the null, we applied various values of λ to deterministically reduce the dengue cases in the 12 intervention clusters before selecting cases (while maintaining the total number of cases at 1000). While this reduction could be applied stochastically, this was not considered necessary given that the permutation approach conditions on the true fixed number of cases in each cluster and bases inference on simply permuting the intervention assignment. Note that this applies a cluster-specific reduction in cases corresponding to the chosen λ.

We first examine the power of several approaches to testing for an intervention effect: (i) comparison of the average test-positive fraction across intervention arms as outlined in Section 2.4, using the t-test statistic (assuming variance homogeneity across the two arms), (ii) using a test statistic based on estimated Odds Ratio from aggregated cluster data (by arm) as described in Section 2.4, using the variance estimate (2.7) (on the log scale), (iii) GEE (with a working exchangeable correlation structure) Odds Ratio from aggregated Odds Ratio data that employs a robust variance calculation, and (iv) a mixed effects logistic regression model with random intercept terms by cluster, the latter two approaches mentioned in Section 2.4. For each method, the same 10,000 random permutations of intervention assignment were used to generate results. Subsequently simulation results were averaged across the nine distinct two-year time periods.

Given the relatively small number of clusters in the Yogyakarta trial, *constrained* randomization was used to ensure balance between study arms for some key cluster covariates including historical dengue incidence, OFI incidence, demographics, population and area. Constrained randomization restricts the number of permuted intervention assignments that are allowed in the random selection. After filtering 100,000 random allocations by these balancing criteria, 247 balanced allocations were retained, i.e. 494 potential allocations of the intervention to either arm. Computational considerations make it difficult to examine the exact permutation distribution of the average test-positive fraction difference or the aggregate Odds Ratio test statistic over all simulations and so we focus here on the approximations derived above.

Table 2.2 shows the power of the various methods for testing the null hypothesis of no intervention effect for values of λ ranging from 1.0 down to 0.3 with $r = 4$ (with results averaged across the nine historical dengue distributions discussed above). The average type 1 error control is extremely close to 5% for the test-positive fraction approach, and very slightly anti-conservative for the Odds Ratio method. Decomposing the results for the nine specific dengue cluster distributions used (not shown here) exhibits very little variation where the range of type 1 errors is from 4.6% to 5.2% for the test-positive fraction methods, and from 6.7% to 8.2% for the Odds Ratio test. The extremes do not necessarily occur for the same dengue distribution across clusters for the two techniques. GEE and the random effects model perform similarly to the direct Odds Ratio technique. This suggests that while 10 clusters were insufficient to reliably use such models, 24 may be enough.

With regards to power, both the test-positive fraction and direct Odds Ratio methods exhibit excellent power, on average, for values of the true Relative Risk equal or lower than 0.5, as exhibited in Table 2.2. The Odds Ratio method exhibits slightly improved power, as compared to the test-positive method, somewhat more than can be explained by its slight anti-conservativeness at the null. The random effects model is, in turn, very slightly more

Relative Risk (λ)	Test-Positive Fraction	Odds Ratio	GEE	Random Effects
1	0.0497	0.0749	0.0779	0.0743
0.6	0.4916	0.5795	0.5936	0.6143
0.5	0.7498	0.8238	0.8266	0.8445
0.4	0.9298	0.9620	0.9603	0.9670
0.3	0.9951	0.9985	0.9983	0.9988

Table 2.2: The proportion of simulations that returned significant results for each intervention effect of interest (λ). The GEE assumed an exchangeable correlation matrix. Each approach was applied to the results of the 10,000 random intervention allocations with 1,000 cases and 4,000 controls ($r = 4$).

powerful than the direct Odds Ratio technique. There is considerably more variation in the power of both methods over the nine varying dengue distribution scenarios. For example, with the Relative Risk, λ, at 0.5, the power of the test-positive fraction approach ranges from 52% to 94% (with average of 75%); at the same λ, the power ranges from 61% to 98% for the Odds Ratio strategy. Again, the extremes occur at differing assumed dengue distributions for the two methods. Table 2.9 provides similar results for $r = 1$. Given the way the simulations were performed, it is immediately apparent why the results are identical for the direct Odds Ratio approach. Results for the other three techniques are very similar with an incremental increase in power for the test-positive fraction method as previously noted (the results for the GEE method only appear the same–differences occur beyond the fourth decimal place).

Tables 2.10 and 2.11 present analogous results for the situation where assignment of the intervention labels is constrained as described above. Now, on average, the test-positive fraction and Odds Ratio methods are both quite conservative in this situation (as is the random effects model), although the Odds Ratio test remains considerably less so. For this case, at least, it appears that the test thresholds should be relaxed (thereby gaining additional power) to produce a 5% type I error; this is, of course, most easily achieved by using the exact permutation distribution. The constrained randomization power is only modestly greater than for unconstrained randomization although this would likely be improved by using the exact permutation distribution in each case.

We also examined point and interval estimation of the Relative Risk based on the test-positive fraction (using both the homogeneous variance assumption and the Welch adjusted method), Odds Ratio, and random effects logistic regression techniques. We examined the identical 10,000 random permutations used for our power calculations above for both $r = 1$ and $r = 4$. Table 2.3 shows the bias of point estimates based on the test-positive fraction quadratic estimation procedure, with a comparison of the average estimated standard deviation of the differences of the test-positive fractions (i.e., T), based on the estimator given in (2.3) with pooled variance, to the true standard deviation of T across the 10,000

	Ratio = 1			**Ratio = 4**		
Relative Risk (λ)	Bias	Average Estimated Standard Deviation of T	True Standard Deviation of T	Bias	Average Estimated Standard Deviation of T	True Standard Deviation of T
1	0.0264	0.0559	0.0564	0.0340	0.0397	0.0401
0.6	0.0386	0.0552	0.0557	0.0180	0.0392	0.0389
0.5	0.0390	0.0546	0.0551	0.0140	0.0389	0.0379
0.4	0.0380	0.0536	0.0540	0.0099	0.0385	0.0365
0.3	0.0351	0.0520	0.0521	0.0053	0.0379	0.0344

Table 2.3: The bias for the test-positive fraction quadratic estimates of the Relative Risk and the standard deviation of the difference in arm-specific averages (T) from 10,000 unconstrained intervention allocations.

simulations. We show the bias on the scale of λ and the standard deviation comparison on the (symmetric) scale for T on which confidence intervals are first calculated. In practice, these confidence intervals are subsequently transformed back to λ. The table shows very little bias in the estimation strategy (with very slight overestimation, i.e. closer to the null here) and that (2.3) provides a very good approximation to the variance of T, even away from the null where the common variance assumption is not exactly satisfied. This suggest that there will be little value in turning to the more complex Welch version of the t-test. This is examined further below when we consider confidence interval coverage.

A similar analysis was implemented for the direct Odds Ratio estimator where bias is assessed on the Odds Ratio scale but standard deviations are compared on the log scale (where confidence intervals are calculated). Note that here, the term 'bias' is a misnomer as the Odds Ratio estimator targets a marginal effect whereas, in the simulated data, λ denotes a cluster-specific effect. As noted above, the Odds Ratio estimate for a specific sample has zero bias as a single random draw from the permutation distribution of the Odds Ratio estimators. Thus, here the bias term refers to the difference between the population-averaged, or marginal Odds Ratio and the true cluster-specific Odds Ratio. This difference moves the Odds Ratio estimate slightly towards the null as would be expected. Specifically, the bias is 0.0287, 0.0172, 0.0143, 0.0115, and 0.0086 for $\lambda = 1, 0.6, 0.5, 0.4$, and 0.3, respectively. The average estimated standard deviation of $\log(\hat{\lambda})$ is 0.2348 whereas the true standard deviation is 0.2435, these values not depending on λ; there is no variation of the true standard deviation of $\log(\hat{\lambda})$ as λ changes since the simulations at differing λ do not allow for stochastic variation around the reduced $A+$ counts as previously noted. Thus, for any given permutation labeling, the estimator $\log(\hat{\lambda})$ is simply shifted by the fixed amount $\log(\lambda)$. When $\lambda \neq 1$, the simulation average of the estimated standard deviation of $\log(\hat{\lambda})$ over permuted intervention assignments also does not depend on the true value of λ when the variant of (2.7) is used in estimation after changing the observed A_j^* counts. This is because the A_j^*s are deterministically obtained by multiplying the (fixed but unobserved) A_js by λ. For variance estimation, Section 4.2 then suggests inflating A_j^* by $\hat{\lambda}$ to estimate the original A_j. But, as noted, $\hat{\lambda} = \lambda\frac{A+H+}{B+G+}$, so that $\lambda/\hat{\lambda}$ does not depend on the assumed λ. By the same token, the estimated $V_D, V_{\bar{D}}$, and $\text{cov}(A+, B+)$ in (2.7) using the modified A_j^*s also do not depend on λ. Note that, at the null, no modification of the observed A_j^*s is necessary if one assumes $\lambda \equiv 1$: in this case, without modification, the average estimated standard deviation of $\log(\hat{\lambda})$ is 0.2363, very similar to that recorded when adjustments are made even at the null. These simulations indicate that (2.7) provides a good approximation to the true permutation variation of the estimator.

Table 2.4 provides analogous average results based on a simple random effects logistic regression model. The bias is acceptably small (and here the random effects model indeed targets the appropriate cluster-specific effect), and the model-based standard deviation estimator (on the log scale) adequately estimates the true standard deviation with this number of clusters.

	Ratio = 1			**Ratio = 4**		
Relative Risk (λ)	Bias	Average Estimated Standard Deviation of $\log(\hat{\lambda})$	True Standard Deviation of $\log(\hat{\lambda})$	Bias	Average Estimated Standard Deviation of $\log(\hat{\lambda})$	True Standard Deviation of $\log(\hat{\lambda})$
1	0.0284	0.2239	0.2366	0.0293	0.2262	0.2390
0.6	0.0169	0.2240	0.2364	0.0171	0.2263	0.2388
0.5	0.0141	0.2240	0.2365	0.0142	0.2263	0.2390
0.4	0.0113	0.2242	0.2364	0.0113	0.2264	0.2388
0.3	0.0083	0.2248	0.2438	0.0081	0.2270	0.2367

Table 2.4: The bias and standard deviation for the random effects Odds Ratio estimates of the relative risk from 10,000 unconstrained intervention allocations.

Finally, Table 2.5 provides coverage properties of approximate confidence interval methods associated with the three methods. The Odds Ratio methods works on the log scale and uses (2.7) to provide the relevant estimated standard deviation, before subsequently transforming back by exponentiating; the test-positive method is based on the scale of T, using (2.3) for standard deviation estimates and then transforms back to the scale of the Relative Risk (noting that this assumes two independent samples of a_js in the two arms as discussed in Section 2.4); and the simple random effects model also works on the log scale using model-based variance estimation before transforming back to the original scale. The coverage estimates are again averaged across the simulated permuted intervention assignment labels for a specific dengue and OFI distribution and then averaged across the nine possible scenarios. All of the methods provide reasonable coverage for each case considered with little to choose between them. The Welch-Satterthwaite modification to the test-positive fraction method makes very little difference to coverage, very slightly improving the performance when $r = 1$ (for example, in the best case considered, increasing coverage to 0.9434 when $\lambda = 0.3$) but working in the opposite direction when $r = 4$ (for example, in the worst case considered, increasing coverage to 0.9651 when $\lambda = 0.4$).

	Any value of r	$r = \mathbf{1}$		$r = \mathbf{4}$	
Relative Risk (λ)	Direct Odds Ratio Method	Random Effects Method	Test-Positive Method	Random Effects Method	Test-Positive Method
1	0.9251	0.9258	0.9494	0.9257	0.9507
0.6	0.9628	0.9635	0.9478	0.9635	0.9544
0.5	0.9629	0.9636	0.9463	0.9636	0.9575
0.4	0.9629	0.9638	0.9445	0.9638	0.9626
0.3	0.9629	0.9642	0.9426	0.9640	0.9713

Table 2.5: The coverage averaged across the 10,000 intervention allocations and 9 time periods for the proposed Odds Ratio method, the random effects Odds Ratio estimates, and the test-positive method using pooled variance estimation. The proposed Odds Ratio Method is invariant to r.

In the simulation scenarios examined, the random effects logistic regression model is reasonable. However, it is premature to speculate that this will remain true in other simulation scenarios or with a different number of clusters. In Section 2.4 we showed that, with 10 clusters, the performance of a random effects logistic regression model is unsatisfactory. Further research will be needed to demonstrate conditions where the latter approach is reliable.

2.6 Discussion

The CR-TND provides an excellent approach to evaluating the efficacy of an intervention randomly applied to clusters that allows for clinic-based surveillance of disease outcomes. Our simulations are necessarily limited although motivated by the specific application. Even here, the simulations only consider one OFI distribution where, in reality, the observed distribution may differ. Further analysis of the various methods in a wider variety of other settings would be valuable. In particular, evaluation of the methods for data generated by designs other than parallel arm interventions is of immediate interest. Consideration of the CR-TND with a stepped wedge design [25] may provide an appealing alternative design in many contexts.

The methods considered here use cluster summaries of the observed frequencies of dengue and OFI outcomes. The methods also assume no interference across cluster boundaries in terms of the intervention and outcome. In the Yogyakarta trial, data will be collected on mobility of participants in the immediate time window proceeding the relevant clinic visit, information that will account for the percentage of time spent in intervention and untreated clusters. In addition, contemporaneous assessments of *Wolbachia* prevalence in trapped mosquitoes by cluster will be measured throughout the trial. Both of these will allow construction of an index of "*Wolbachia* exposure" prior to disease onset precipitating a clinic visit. The ability to capture an exposure assessment immediately prior to the onset of clinical symptoms presents a significant advantage to a cohort design where such measurements would be challenging absent constant surveillance. This kind of exposure measure, and other factors of interest, introduce individual level covariates that may explain some of the variation in dengue, as compared to other OFI, incidence. Such data then demands the extension of the permutation-based inference methods considered here to allow for individual-level explanatory covariates. This requires extension of the methods of [54] (and earlier authors) to the CR-TND context. This paper summarizes methods to extend exact permutation inference to account for covariance adjustment in cluster-randomized trials with continuous outcomes and sharp null hypotheses. More recent work extends these methods to handle composite null hypotheses with binary outcomes [16, 31]. Extension of these ideas to the CR-TND design represents an important area of current research.

2.7 Supplement

Cluster ID	Distribution of Dengue Cases	Distribution of OFI Controls
1	52	138
2	74	212
3	54	125
4	72	145
5	46	165
6	42	194
7	70	250
8	50	131
9	73	229
10	69	156
Total	602	1745

Table 2.6: Hypothetical dengue and other febrile illness (OFI) count data for an example of 10 clusters used for permutation distribution estimates.

	Period								
Cluster ID	'03 – '05	'05 – '06	'06 – '07	'07 – '08	'08 – '10	'10 – '11	'11 – '12	'12 – '13	'13 – '14
1	13	19	37	29	42	48	18	26	34
2	14	14	30	27	34	37	15	25	34
3	35	32	39	43	62	52	25	40	38
4	9	13	13	8	18	18	6	7	9
5	17	25	69	60	36	53	34	47	71
6	37	38	77	72	75	89	84	120	104
7	23	28	48	51	85	76	28	40	36
8	20	32	51	57	66	41	13	36	37
9	25	29	46	41	57	48	15	27	25
10	14	25	53	49	41	31	9	35	42
11	40	61	78	64	84	98	57	62	71
12	33	54	74	59	80	80	44	63	69
13	35	52	79	86	119	112	49	56	76
14	28	39	57	48	59	56	29	49	62
15	30	39	56	46	52	40	20	25	27
16	22	51	68	47	56	43	19	36	38
17	12	18	25	22	20	14	8	17	16
18	41	55	112	93	130	151	81	139	128
19	16	27	69	71	53	44	24	47	69
20	19	37	43	28	45	41	30	79	77
21	24	45	63	49	59	62	42	73	68
22	33	57	72	59	84	73	35	66	62
23	12	19	29	29	36	29	14	34	32
24	21	40	67	90	151	106	27	72	76

Table 2.7: *Dengue Case Counts.* The frequency of recorded (hospitalized) dengue fever cases in each of these 24 clusters for each of nine distinct two-year periods covering the time interval from 2003-2014. During this period, there was no available data for 2004 and 2009, so that the first two-year interval was for 2003 and 2005; similarly the 2008-10 interval included data for 2008 and 2010. Otherwise each two-year period covered consecutive years.

Cluster ID	'14 – '15
1	486
2	155
3	1197
4	255
5	249
6	710
7	658
8	714
9	478
10	376
11	388
12	426
13	842
14	547
15	285
16	586
17	344
18	484
19	151
20	223
21	522
22	804
23	286
24	792

Table 2.8: *OFI Counts.* Data for the distribution for OFIs is only available for one two-year period from 2014-15.

Relative Risk (λ)	Test-Positive Fraction	Odds Ratio	GEE	Random Effects
1	0.0506	0.0749	0.0744	0.0742
0.6	0.5258	0.5795	0.6161	0.6222
0.5	0.7798	0.8238	0.8446	0.8508
0.4	0.9418	0.9620	0.9657	0.9693
0.3	0.9965	0.9985	0.9988	0.9990

Table 2.9: The proportion of simulations that returned significant results for each intervention effect of interest (λ) as in Table 2 of the paper, but now with 10,000 random intervention allocations of 1,000 cases and 1,000 controls ($r = 1$).

Relative Risk (λ)	Test-Positive Fraction	Odds Ratio	GEE	Random Effects
1	0.0013	0.0117	0.0779	0.0743
0.6	0.4786	0.6136	0.5936	0.6144
0.5	0.8075	0.8866	0.8266	0.8445
0.4	0.9732	0.9831	0.9603	0.9670
0.3	1.0000	1.0000	0.9983	0.9988

Table 2.10: The proportion of simulations that returned significant results for each intervention effect of interest (λ) as in Table 2 0f the paper, but now with each approach applied to the results of the 247 constrained intervention allocations with 1,000 cases and 4,000 controls ($r = 4$).

Relative Risk (λ)	Test-Positive Fraction	Odds Ratio	GEE	Random Effects
1	0.0022	0.0117	0.0058	0.0054
0.6	0.5283	0.6136	0.6534	0.6601
0.5	0.8488	0.8866	0.9141	0.9213
0.4	0.9800	0.9831	0.9906	0.9926
0.3	1.0000	1.000	1.0000	1.0000

Table 2.11: The proportion of simulations that returned significant results for each intervention effect of interest (λ) as in Supplementary Table 4, but now constrained intervention allocations with 1,000 cases and 1,000 controls ($r = 1$)

Chapter 3

Analysis of Counts for Cluster Randomized Trials: Negative Controls and Test-Negative Designs

This chapter has been published in the journal *Statistics in Medicine.*[1]

3.1 Introduction

Randomized controlled trials are the gold standard for evaluating the efficacy of health interventions. Randomization makes comparison groups as similar as possible in all factors except for the intervention under study, and provides a basis of non-model based inference. When an intervention is delivered to groups of individuals, e.g. in neighborhoods, or may have a community-wide health impact, randomization of the intervention necessarily occurs at the group, rather than individual, level. Such a trial is termed a cluster randomized trial (CRT).[22] The non-independence of individuals within each cluster in CRTs causes statistical inefficiency–the design effect– necessitating inflation of the sample size to achieve power equivalent to an individually randomized trial.[22, 11, 21]

In many CRTs, outcome measurements are made at the cluster–rather than individual–level for a variety of reasons. For example, counts of events across a cluster may be collected by existing or designed surveillance systems. For CRT count outcome data, common estimators of the intervention effect include estimation of absolute and relative rate differences, usually based on demographic information on relevant population years of observation, or population size, per cluster.[22] When adjustment for cluster level covariates is desirable, model-based regression modeling approaches are often used, including marginal generalized estimating equation (GEE) approaches and mixed effects models with random effects at

[1] I am grateful for permission from my collaborator, Nicholas P. Jewell, to include this paper in my dissertation. Additionally, I am grateful to my collaborators at the World Mosquito Program for the motivating example.

the level of the cluster. The latter models can be extended to allow for individual level information.

In many situations, these standard approaches require modification. For example, in settings where few clusters are available for randomization, model-based estimation and inference may be less accurate and require small sample size adjustments.[22] In such cases, a randomization-based strategy (e.g. permutation tests) presents an attractive alternative. Further, population-based denominators may sometimes be unavailable or not appropriate. The latter can occur when the count ascertainment system does not cover the entire cluster populations, perhaps due to access to care issues. The statistical analysis then depends solely on randomization balancing the unobserved population denominators across intervention arms. This risks unobserved bias–particularly in unblinded studies–due to differential ascertainment coverage across arms that will confound any intervention effect.

As noted, it is possible to estimate and test an intervention effect using only cluster-level case counts given intervention randomization. Here, we discuss such inference, focusing on the Relative Risk and its permutation distribution (under permuted intervention assignments). We subsequently consider the impact of differential ascertainment bias, and introduce a method to remove, or reduce, such bias through the use of *negative controls.*[33] We discuss briefly the required properties for a valid negative-control count. We use simulations to address bias and precision comparisons between the various methods.

An example of a design that explicitly uses negative controls is the *test-negative design* that was recently extended to allow for cluster randomization of an intervention.[1] Test-negative designs are explicitly used to address ascertainment bias caused by differential health-care seeking behavior.[58, 27] We thus interpret our findings in the context of cluster randomized test-negative designs with analytic methods that either only use case count data or use negative control (in addition to case count) information. Test-negative designs also directly accommodate the absence of population level denominator information underlying the observed counts of interest.

These issues are motivated by the World Mosquito Program's ongoing balanced parallel-arm Cluster Randomized Test-Negative Design (CR-TND) trial to evaluate the efficacy of *Wolbachia*-infected mosquitoes in reducing the burden of dengue transmission in Yogyakarta City, Indonesia. In this study, Yogyakarta City, and its population of approximately 400,000, was divided into 24 contiguous clusters each measuring approximately 1 km^2 in size but with varying population density and socioeconomic status. Twelve of the clusters were randomly assigned to an intervention arm that received releases of *Wolbachia*-infected mosquitoes. *Wolbachia* successfully transinfected in non-native hosts such as *Aedes aegypti* mosquitoes, the primary vectors of dengue, have been shown to disrupt the transmission of dengue and other flaviviruses by minimizing virus replication within the vector.[30] The remaining twelve clusters were assigned as control clusters. Count ascertainment depends on individuals seeking care at *puskesmas* (community health clinics) who present with general symptoms consistent with the clinical case definition of dengue. Such individuals who consent to enroll in the trial are subjected to laboratory testing for dengue, which determines their test-positive (case) or test-negative (control) status. The trial has been described in greater detail elsewhere.[28, 2]

3.2 Direct comparison of counts in the absence of population denominators

We consider here a CRT for which the outcome is measured at the cluster level and comprises of a count of a number of "events" in each cluster. For example, the counts could represent the number of incident dengue infections over the study period as obtained through some well-defined ascertainment system. We let A_j denote the observed count in the j^{th} cluster assigned to the intervention and, analogously, G_j is the count in the j^{th} control cluster. Then, A_T and G_T are the *total* sum of the j cluster-level case counts (A_j, G_j) in the treatment and control arms, respectively. That is, $A_T = \sum_{j=1}^{m} A_j$ and $G_T = \sum_{j=1}^{m} G_j$ where we assume, for convenience, that m clusters are randomly assigned to both the intervention and control arm.

Given randomization, differences in the cluster counts between the intervention and control arms should only arise through the intervention so long as case ascertainment is not differentially applied across arms. In particular, the underlying population denominators for a rate should be balanced across arms. Thus, to test the null hypothesis that there is no difference in the rate of case counts between the intervention and control arms, we can use the test statistic in Equation 3.1.

$$\begin{aligned} T &= \sum A_j - \sum G_j \\ &= A_T - G_T \end{aligned} \tag{3.1}$$

With small numbers of clusters, we focus on the permutation distribution of T (across all permutations of clusters' intervention assignments). In the simple case considered here, there are $\binom{2m}{m}$ possible intervention assignments and computation of an estimate for each of these (while holding each cluster count fixed) yields the permutation distribution that can form the basis of randomization inference. It is immediate that $\mathbb{E}_P[A_T] = \mathbb{E}_P[G_T] = n_D/2$, where $\mathbb{E}_P$ refers to the expectation under the permutation distribution, and n_D is the total of all counts across all clusters, i.e. $A_T + G_T$, held fixed over all permutations. Further, from finite sampling methods, $\text{Var}_P(A_T) = mV_D/2$ where V_D is the variance of the combined counts $A_1, \ldots, A_m, G_1, \ldots, G_m$ in the intervention and control clusters combined, with this variance calculated using $(2m-1)$ in the denominator. This follows since, for a random permutation, the A_j counts are simply randomly selected from the combined counts across all clusters.

Thus, $\mathbb{E}_P[T] = \mathbb{E}_P[A_T] - \mathbb{E}_p[G_T] = 0$, and the permutation variance of T is just $\text{Var}_P(T) = 2mV_D$. Thus, to evaluate the null hypothesis of no intervention effect we can either use the full permutation distribution or approximate such an approach by comparing a standardized statistic, $T/\sqrt{2m\hat{V}_D}$–using an appropriate estimate of V_D–to a t distribution with the appropriate number of degrees of freedom.

V_D can be simply estimated by the empirical variance of the A_js in the intervention clusters or the G_js in the control clusters (or the variance of the counts combined across both arms). Since the arms contain the same number of clusters, a simple average of these two arm-specific variance estimates could be used, leading to the so-called pooled variance estimator for the two-sample t-test with $2(m-1)$ as the appropriate number of degrees of freedom. The combined variance, and, to a lesser extent the pooled estimator, are likely to be biased in estimating V_D in the presence of an intervention effect. This suggests an alternative approach when using the permutation distribution, or its approximation, as the basis for confidence intervals, which we discuss below.

	Seek Care				**Do Not Seek Care**			
	Test-Positive Cases	Test-Negative Controls	Not Infected	*Total*	Test-Positive Cases	Test-Negative Controls	Not Infected	*Total*
Intervention (E)	A_T	B_T	C_T	N_{IO}	D_T	E_T	F_T	N_{IU}
Control ($\bar{E}$)	G_T	H_T	I_T	N_{CO}	J_T	K_T	L_T	N_{CU}

Table 3.1: Stratification of population based on intervention status, infection, and health-care–seeking behavior. Adapted from Figure 1 of Jackson & Nelson.[27]

We now turn to estimation of λ, the Relative Risk comparing intervention and control arms. One can think of λ as the ratio of the underlying rates that generates the cluster counts in each arm. Alternatively, λ is simply the ratio of the mean of the cluster counts across the two arms. Here, we focus on the estimator $\lambda_R = A_T/G_T = A_T/(n_D - A_T)$, where R simply stands for ratio (of the counts). For confidence intervals, we move to the symmetrically distributed version, $\log(\lambda_R)$. By definition, $\mathbb{E}_P \log(\lambda_R) = 0$ at the null. Away from the null, we need to evaluate the permutation distribution of the $\log(\lambda_R)$ assuming an intervention effect. Note that the delta method can be used to approximate the permutation variance of $\log(\lambda_R) \approx (16/n_D^2)(m/2)V_D$.

Note that the intervention only affects the counts $A_1, \ldots, A_m$ by assumption. These are each replaced in turn by $A_1^*, \ldots, A_m^*$ which reflect altered counts in the intervention clusters. For large populations, $A_j^* \approx \lambda A_j$ for the intervention clusters, assuming that the intervention effect is the same for all clusters. The common modification of the $A_1, \ldots, A_m$ has two immediate implications: first, under the permutation distribution, $\mathbb{E}_P \log(\lambda_R) \approx \log(\lambda)$; second, there is no change to the variance formula $\log(\lambda_R) \approx (16/n_D^2)(m/2)V_D$ since all count ratios for different permutations are shifted by approximately $\log(\lambda)$.

However, away from the null, we have to modify the estimates of n_D, V_D due to the replacement of each A_j with A_j^*. The necessary adjustment is achieved by simply increasing the observed A_j^*s by the common factor $1/\lambda_R$ to obtain an estimate of A_j (in the j intervention clusters), en route to an estimate of n_D, V_D as at the null.

3.3 Differential case ascertainment

A fundamental threat to the validity of the approach of Section 3.2 – even with randomization – arises when there is differential "counting" methods across the two arms. In such cases, when passive surveillance approaches are used to generate the necessary counts, differential case ascertainment may occur across treatment arms. For example, individuals' health-care–seeking behavior may be differential based on knowledge of their intervention assignment and this will affect any ascertainment system that is based on attendance in some health-care setting. This behavior is particularly relevant in trials where blinding of the participants and/or investigators to the intervention is infeasible for logistical, ethical or other reasons. We refer to this phenomenon as differential count ascertainment. We stress that this threat to validity persists even if the relevant denominator information is known for the cluster counts.

We quantify this effect through the relative propensity π of treated and untreated populations to "be counted", e.g. seek health care. We allow this propensity to differ across treatment arms denoted by E here, for convenience. That is E refers to individuals in the intervention arm and $\bar{E}$ to those in the control arm. Then we let

$$\alpha_{RA} = \frac{Pr(A = 1|E, D)}{Pr(A = 1|\bar{E}, D)},$$

where A stands for ascertainment, RA for relative ascertainment, and the binary indicator D denotes a 'case' that would be counted if ascertainment was guaranteed.

It is obvious that with the comparison of counts across arms as described in Section 3.2, the effects of risk reduction and relative ascertainment are completely confounded and could not be disentangled without direct knowledge of α_{RA}. One approach to address this fundamental bias, is through use of negative controls. Negative controls, commonly used to calibrate measurements in laboratory experiments, have recently been re-examined for epidemiological applications.[33] The key requirements for a useful negative control outcome is that (i) no intervention effect is expected on the negative control outcome, and (ii) negative control outcomes must be affected by identical relative ascertainment effects as our outcome of interest. Note that the latter assumption allows differential ascertainment across intervention arms but this must occur in identical fashion as to what occurs for the outcome of interest as quantified by α_{RA}. It is exactly this assumption that allows estimation of α_{RA} and subsequent removal of ascertainment bias in estimation of λ.

The second of these conditions may appear difficult to achieve in any practical intervention study. We nevertheless introduce exactly such an example in the context of what are referred to as test-negative designs.

The Test-Negative Design

In infectious disease research, issues relating to differential case ascertainment, typically under the influence of differential health-care–seeking behavior, have been mitigated by the implementation of the test-negative design (TND). TNDs represent a variant on a traditional case-cohort design: studies enroll subjects who seek care for a clinical syndrome, defining those who test positive and negative for a pathogen of interest as cases and controls, respectively. Specifically, the popularity of the TND arose from its ability to use existing surveillance systems (e.g. clinic data) to estimate seasonal influenza vaccine effectiveness while minimizing bias due to health-care–seeking behavior. A nuanced discussion of this design can be found in the recent literature [58, 64, 15] that includes a formal analysis of causal diagrams associated with the design. The design and analytical methods were recently extended to cluster randomized interventions, [28, 2] yielding the so-called cluster randomized test-negative design (CR-TND). A recent review of test-negative designs to mosquito vaccine effectiveness discusses 348 such studies.[10]

In a TND, test-positives play the role of our case counts in Section 3.2, and are ascertained through attendance, diagnosis and testing at a clinic or other health-care setting. Subsequently, a critical component of the TND is the definition of test-negatives. As negative controls, the objective is to identify a disease that is unaffected by the intervention of interest, and symptomatically similar to the disease outcome of interest. Upon recruitment at a clinic, a highly sensitive and specific laboratory test is used to distinguish test-positive cases (those with the disease of interest) from the test-negative controls (those without). The full extent of these assumptions have been critically discussed in the literature.[1, 58] The key property of negative controls regarding differential ascertainment is explicitly achieved since participants do not know their disease status until they are ascertained and so it theoretically not possible for the test-positives and test-negatives to suffer from differential relative ascertainment; that is, the relative ascertainment α_{RA} is the same for Ds (test-positives) as for $\bar{D}$s (test-negatives).

Using the cumulative notation provided in Table 3.1, that describes totals across clusters, the negative control assumption that the intervention has no impact on test-negatives leads to the proportion of test-negative individuals among the intervention care-seeking population (B_T/N_{IO}) being approximately equivalent to the proportion of test-negative individuals among the negative control care-seeking population (H_T/N_{CO}). Note that, in this context, N_{IO} and N_{CO} represent the unobserved denominators discussed in Section 3.2. It is then possible to approximate the natural, but unobserved, estimate of the relative risk of disease across the intervention and control populations ($(A_T/N_{IO})/(G_T/N_{CO})$) by substituting the ratio of test-negative individuals from the intervention and control subpopulations (H_T/B_T) as a proxy for the unobserved relative sizes of the care-seeking intervention and control denominators (N_{IO}/N_{CO}). This results in the simple TND estimator, $\lambda_{TND} = A_T H_T / B_T G_T$.[58]

Estimation of differential case ascertainment

Note that the assumptions of an appropriate negative control allow for estimation of the common relative ascertainment parameter α_{RA}. The first assumption indicates that the relative counts of test-negatives in the intervention and control arms are not affected by the intervention which has a null effect on the negative control outcome. The second assumption then yields that the relative ascertainment of test-negatives is the same as for test-positives representing the outcome of interest.

Consider the scenario in which individuals within the intervention arm are ascertained differentially from individuals within the control arm. Focusing on the test-negatives, our assumptions show that $\alpha_{RA} = Pr(A = 1|E, \bar{D})/Pr(A = 1|\bar{E}, \bar{D})$. Provided the other CR-TND assumptions hold,[1] and with the assumption of no intervention effect on the negative controls, α_{RA} can be estimated by the identical approach previously outlined for case count only estimation of the RR; that is, $\hat{\alpha}_{RA} = B_T/H_T$. This provides an unbiased estimator of the relative ascertainment parameter.

The variance of $\hat{\alpha}_{RA}$ can be estimated exactly as we described for the intervention effect estimate in Section 3.2: $\text{Var}_P(\hat{\alpha}_{RA}) \approx (16/n_{\bar{D}}^2)(m/2)V_{\bar{D}}$, where $V_{\bar{D}}$ is the variance of the clusters' test-negative counts combined across intervention arms and $n_{\bar{D}} = B_T + H_T$. To assess whether ascertainment (of the negative controls) differs across arms, a suitable test statistic is, again, the difference in counts, $T = B_T - H_T$, scaled by the variance $\text{Var}_P(T) = 2mV_{\bar{D}}$, where $V_{\bar{D}}$ is the population variance of the $2m$ test-negative counts, and compared to a t distribution with $2(m-1)$ degrees of freedom (assuming we use a variance estimate that averages variability across the two arms as described in Section 3.2). This test is of interest in its own right when negative control information is available as it assesses differential ascertainment effects across arms independently of any intervention. Such information may be useful in planning and interpreting future trials.

Estimating the intervention effect, λ, in the presence of differential case ascertainment

When $\alpha_{RA} \neq 1$, the estimated intervention effect given by λ_R is necessarily biased, as noted above; that is, the estimate is shifted multiplicatively by α_{RA} (or, additively, by $\log \alpha_{RA}$ on the log scale). Without further information, this reflects the vulnerability to bias of the 'count-only' approach of Section 3.2. However, knowledge of the negative control counts allows estimation of α_{RA} as shown in Section 3.3. Thus, a 'de-biased' intervention Relative Risk can then be estimated by $\hat{\lambda} = \lambda_R \times \alpha_{RA}^{-1} = \frac{A_T H_T}{G_T B_T}$. This, of course, is precisely the simple TND estimator (λ_{TND}) proposed for all test-negative designs including the CR-TND. Randomization-based inference associated with this estimator is presented in Chapter 2.

3.4 Simulations

Data-based simulations evaluate the performance of the proposed estimation methods. As a practical basis for simulations, historical counts of dengue from 24 contiguous clusters within a city in Indonesia collected from 2003 to 2014 were divided into 9 consecutive[2] two-

[2] There are two exceptions to the consecutive two-year period counts. Data was missing in 2004 and 2009 which were ignored in making a two year time period in both cases.

year periods. Other febrile illnesses (OFIs) with similar presenting symptoms will be used as negative controls. Counts of OFIs for each of the 24 clusters from 2014 through 2015 provided the historical distribution of these negative controls. Exact distributions of these historical counts can be found in Tables 2.7 and 2.8. For each historical period, complete random assignment was performed such that $m = 12$ of the total 24 clusters were assigned to a putative intervention and the remainder to control.

Instead of building an exhaustive permutation distribution of the more than 2 million distinct intervention allocations for each time period, each simulation assigned intervention according to the same 10,000 distinct potential intervention allocations and examined the results of these intervention allocations across all 9 historical time periods.

For a specific period, the distribution of the case counts (n_D) and negative control counts ($n_{\bar{D}}$) amongst clusters are assumed to follow multinomial distributions parameterized by the observed historical cluster-level proportions of cases (or negative controls) that fell in cluster j, p_{Dj} or $p_{\bar{D}j}$, respectively. Given an intervention effect λ, $p^*_{Dj} = \lambda p_{Dj}$ for all clusters in the intervention arm with the other proportions in the control cluster left unchanged. These adjusted proportions are then standardized such that $\sum_{i=1}^{2m} (E = 1) \times \lambda p_{Di} + \{1 - (E = 1)\} \times p_{Di} = 1$. The negative control distribution is unaffected by the intervention by definition.

To allow for potential differential ascertainment by intervention arm, we assume that α_{RA} can be applied in a similar manner except that it also modifies the distribution of negative controls. Since α_{RA} is a relative measure of differential ascertainment, we modify all case counts and negative control counts within the intervention arm only. After this modification, the proportions are again standardized such that the proportions of case counts and negative control counts each sum to one across all clusters.

The marginal ratio of cases (Ds) to negative controls ($\bar{D}$s) was 1:4, with 1,000 cases and 4,000 controls selected for each simulation. Five[3] intervention relative risks ($\lambda = 1, 0.8, 0.6, 0.4, 0.2$) are examined and four different levels of differential ascertainment ($\alpha_{RA} = 1, 0.95, 0.85, 0.5$). The performance of the count ratio method of Section 3.2 (λ_R) was compared to the bias-adjusted method of Section 3.3 (λ_{TND}) using the variance estimates noted in 2.4.

For model-based comparisons we also consider mixed effects models and GEE. For the estimation of the Relative Risk using only case counts in the absence of a population-based denominator, the GEE and mixed effects models assume Poisson distributed counts and use a canonical log link. To estimate the Relative Risk with the inclusion of negative controls counts, the GEE and mixed effects models assume binomially distributed counts and use a canonical logit link. All mixed effects models include a random intercept for each cluster and all GEEs assume an exchangeable correlation structure.

All simulations and subsequent analyses were performed in R version 3.6.1 “Action of the Toes”.[43] GEE models were fit using “geeglm” from the “geepack” package.[24, 68, 67] Mixed effects models used “glmer” from the “lme4” package.[3] Plots were generated using

[3] The supporting material also shows results for two additional intervention relative risks $\lambda = 0.5, 0.3$.

the "ggplot2" package.[65] All additional simulation code is available as a GitHub repository managed by the first author.[4]

3.5 Results

Detecting an intervention effect

Figures 3.1, 3.2, and 3.3 compare the performance of the count ratio estimator (λ_R) to the simple de-biased estimator (λ_{TND}), as well as the mixed effects, and GEE approaches. The simulation results are averaged across the 10,000 unique intervention allocations applied to each of the 9 different observed historical time periods. Thus the simulations reflect overall performance over nine somewhat different scenarios. These results are summarized numerically in Tables 3.3 through 2.5 included in Section 3.7.

Power, shown in Figure 3.1, is estimated as the proportion of permuted allocations that return a significant test result at a significance level of 0.05. Significance for the count ratio method is determined on the basis of the test statistic proposed in Equation 3.1, standardized by its estimated variance, as compared to a t-distribution with $2(m-1)$ degrees of freedom. In the case of the simple ascertainment de-biased estimator (λ_{TND}), a significant result is determined by the absence of the null value in the 95% confidence interval around the estimated intervention RR, as performed on the log scale. Finally, significance is determined by the model-based coefficient p-value corresponding to intervention in the mixed effects and GEE models. The power for each intervention and differential ascertainment scenario is relatively stable for the approaches that make use of both count and negative control information (Figure 3.1B). The count only approach shows the most desirable estimated Type I error in the setting where there is no differential ascertainment (power = 0.058). However, it seriously deteriorates for a high level of differential ascertainment. This is explained by the introduced bias in estimation. This does not effect the approaches that use the negative control information (Figure 3.1B) although there is some anti-conservativeness in the simple TND estimator for a high level of differential ascertainment.The increasing power of the count only methods (Figure 3.1A) for any fixed value of λ is an artefact of the fact that, for the simulations considered here, the intervention effect and the differential ascertainment work in the same direction (of reducing counts in the intervention clusters); for simulations with $\alpha_{RA} > 1$ (not shown here), the power of the count only approaches substantially worsens as differential ascertainment widens.

Bias (Figure 3.2) is estimated as $\mathbb{E}_P[\hat{\lambda}] - \lambda$. The estimated bias is reported on the scale of the Relative Risk for interpretability. In the setting of no differential ascertainment ($\alpha_{RA} = 1$), the estimators perform similarly, as expected, as most of the estimators enjoy zero asymptotic bias (note that the mixed effects model estimates a cluster specific odds ratio that is not identical to the marginal odds ratio targeted by GEE and the other methods). The small gain when using the count ratio estimator is less than 1% which is negligible.

[4] `https://github.com/sdufault15/case-only-crtnd`

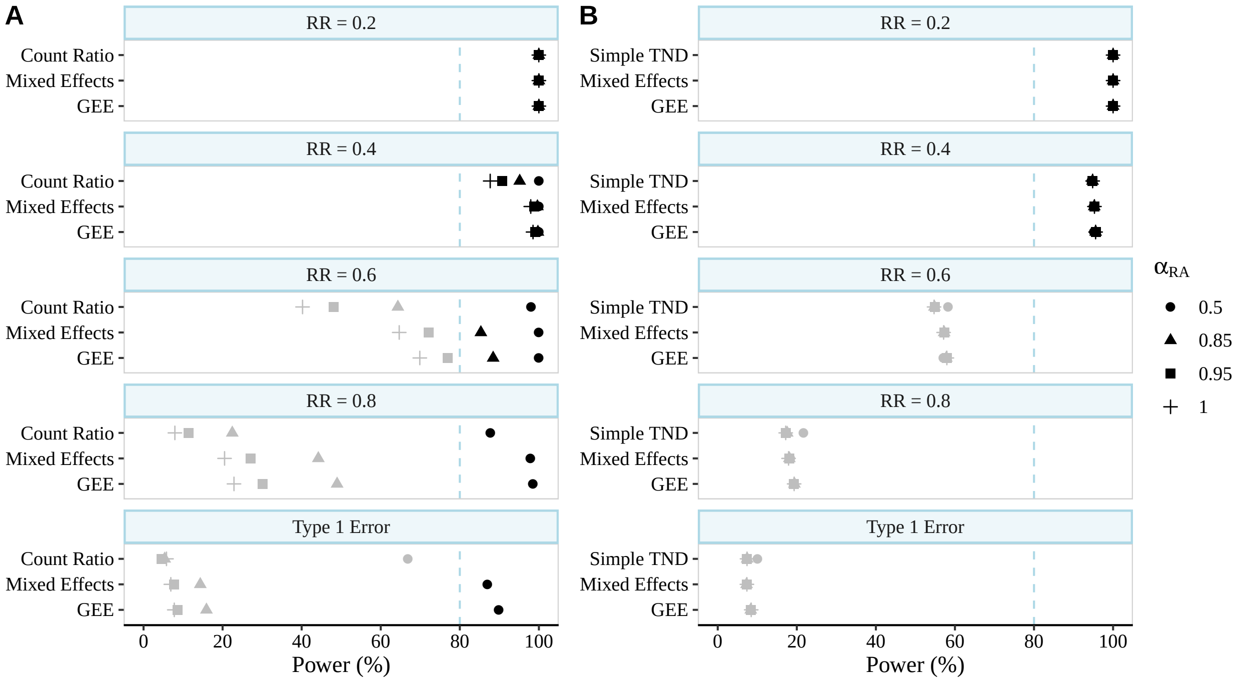

Figure 3.1: The power, and Type I error rates, in testing departure from the null of no intervention effect based on various estimation methods for a range of Relative Risks (RR), over 10,000 intervention allocations applied to each of 9 historical time periods with 1,000 cases and 4,000 negative controls (when applicable). Differential ascertainment (α_{RA}) is allowed to increase in severity. **A)** Results from count only methods in the absence of a population denominator. The mixed effects and GEE models assume the case counts are Poisson distributed and use the canonical log link. **B)** Negative control bias-adjusted results. The mixed effects and GEE models assume the case and negative control counts are binomially distributed and use the canonical logit link.

Further, as differential ascertainment increases, the count only estimators (Figure 3.2A) are unable to reliably estimate the intervention effect. The simple TND estimator, binomial GEE, and binomial mixed effects methods (Figure 3.2B) all maintain low bias (bias ≤ 0.05).

Finally, coverage (Figure 3.3) represents the proportion of estimated 95% confidence intervals which contain the true intervention Relative Risk. Again, in the absence of differential ascertainment, the count ratio estimator (λ_R) enjoys slightly improved coverage across each of the examined intervention RRs ($\approx$ 93.4% coverage). As expected, however, the coverage deteriorates as the bias from differential ascertainment increases (Figure 3.3A). Slight deterioration in coverage as differential ascertainment worsens was observed across each of the estimators, though for the approaches accounting for negative controls (Figure 3.3B) coverage fell only to 90%.

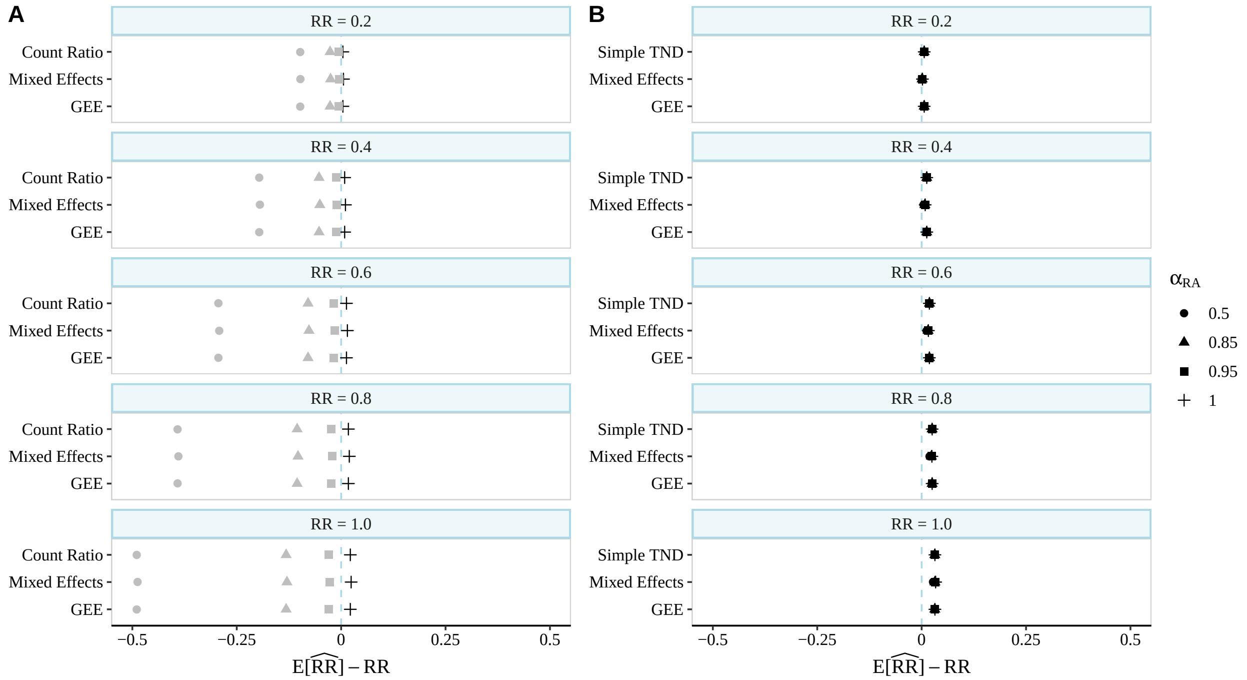

Figure 3.2: Bias in estimation of the intervention Relative Risk for various methods over 10,000 intervention allocations applied to each of 9 historical time periods with 1,000 cases and 4,000 negative controls as differential ascertainment increases in severity. **A)** Results from count only methods in the absence of a population denominator. The mixed effects and GEE models assume the case counts are Poisson distributed and use the canonical log link. **B)** Negative control bias-adjusted results. The mixed effects and GEE models assume the case and negative control counts are binomially distributed and use the canonical logit link.

Detecting differential ascertainment

As described in Section 3.3, the count ratio estimator can be used to estimate the Relative Risk of differential ascertainment (α_{RA}) using the negative control counts, when available. Table 3.2 displays the bias, power, and coverage statistics for estimation of α_{RA} when the true α_{RA} is null ($\alpha_{RA} = 1$), low ($\alpha_{RA} = 0.95$), medium ($\alpha_{RA} = 0.85$), and high ($\alpha_{RA} = 0.5$). As the distribution of the negative controls is assumed unaffected by the intervention, these results are true for any size of intervention effect λ. Despite the low bias in estimation, good coverage and type I error (i.e. power when $\alpha_{RA} = 1$), the power to detect differential ascertainment away from the null (i.e. $\alpha_{RA} \neq 1$) is necessarily low except with high differential ascertainment.

Note that, in Table 3.2, when $\alpha_{RA} = 1$ (i.e. at the null of no differential ascertainment), the power represents the Type I error and should be complementary to the coverage rate in that the two values should sum to 1. However, for the count ratio estimator, hypothesis

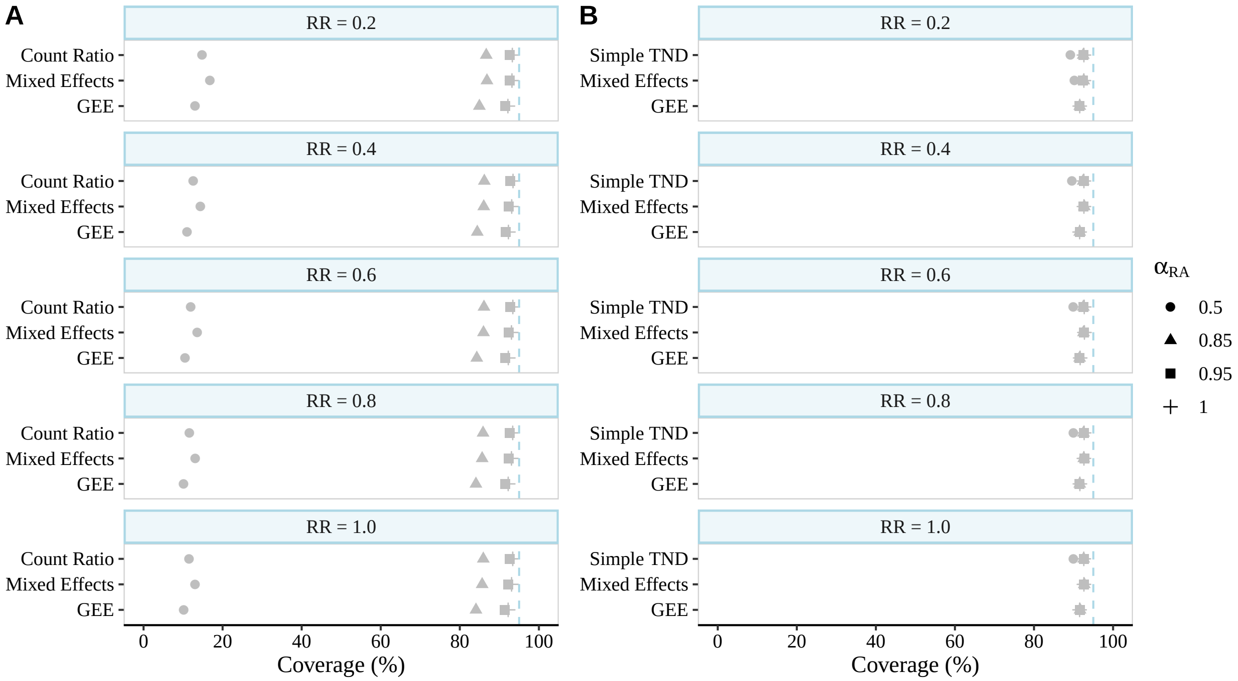

Figure 3.3: 95% confidence interval coverage based on estimation of the intervention Relative Risk for various methods over 10,000 intervention allocations applied to each of 9 historical time periods with 1,000 cases and 4,000 negative controls as differential ascertainment increases in severity. **A)** Results from the comparison of counts in the absence of population denominator. The mixed effects and GEE models assume the case counts are Poisson distributed and use the canonical log link. **B)** Bias-adjusted results. The mixed effects and GEE models assume the case and negative control counts are binomially distributed and use the canonical logit link.

testing is based on the normalized t-statistic of Section 3.3, whereas coverage is based on the confidence interval associated with the ratio estimator of the relative ascertainment α_{RA}, also introduced in Section 3.3. Thus, the corresponding entries only approximately add to one.

3.6 Conclusions

The count only approaches for CRTs perform comparably in estimation of an intervention Relative Risk as compared to alternatives that use additional negative control information (albeit at reduced power), but only in the absence of differential ascertainment. The count only methods have reasonable bias and coverage properties (near 94%), and comparable power while maintaining a desirable Type I error rate. These properties depend entirely on

	Bias	Power	Coverage
$\alpha_{RA} = 1$	0.0215	0.0583	0.935
$\alpha_{RA} = 0.95$	0.0207	0.0479	0.934
$\alpha_{RA} = 0.85$	0.0181	0.0443	0.935
$\alpha_{RA} = 0.5$	0.0108	0.6870	0.934

Table 3.2: Bias, power (Type I error when $\alpha_{RA} = 1$), and 95% confidence interval coverage based on estimation of differential ascertainment by intervention arm from 10,000 permuted intervention allocations across 9 time periods of historical data for a ratio of 1,000 cases to 4,000 negative controls.

randomization and so cannot be used directly when the clusters are *not* randomized.

Further, the performance of the count only approaches falter in the presence of even relatively low differential ascertainment ($\alpha_{hcsb} = 0.95$) as demonstrated by increases in bias and decreases in coverage. In contrast, methods that adjust for differential ascertainment by incorporating proposed negative control counts maintain desirable performance even under major differential ascertainment. Thus, the count only estimators should only be used when there is no other alternative (despite this being currently standard), and should be treated with considerable caution if there is any possibility of differential ascertainment. The use of negative controls in CRTs provides an attractive option to remove, or reduce, the effect of differential ascertainment and should be used more widely.

Potentially, the results have more significance when considering stepped wedge designs rather than the parallel arm scenario considered here. Currently, almost all stepped wedge studies only consider an outcome of interest and do not employ negative controls to remove bias. Analytical results for the stepped wedge design in this context will be provided elsewhere.

Finally, determining whether differential ascertainment exists by the estimation approach proposed here is informative but lacks sufficient power to detect moderate differences by intervention arm. As such, determining whether a setting is appropriate for future estimation by the count only approach will likely return uninformative results unless ascertainment is exceptionally differential ($\alpha_{RA} \leq 0.5$ or $\alpha_{RA} \geq 2.0$).

Recommendations

The findings suggest two key recommendations. First, in CRTs where only counts are available for analysis, the proposed estimator is a viable option with desirable statistical properties. However, even with randomized interventions, it is only appropriately employed in settings where there is little to no differential ascertainment by intervention arm. This is likely most plausible under blinded intervention assignment. Second, in CRTs where differential ascertainment is likely or inevitable, negative control data is important for validity.

3.7 Supplement

Note that when $\lambda = 1$ (i.e. at the null hypothesis of no intervention effect), coverage and Type I error are complementary so that the corresponding entries of Table 3.3 and Table 3.5 are expected to add exactly to one. This is exactly true for the simple TND, GEE and mixed effects approaches since both tests and confidence intervals are based on estimation of the same parameter. However, for the count ratio only estimator (last column of both Tables 3.3 and 3.5), hypothesis testing is based on the normalized t-statistic of Section 3.2, whereas coverage is based on the confidence interval associated with the ratio estimator of the Relative Risk, λ_R, also introduced in Section 3.2. Thus here, the corresponding entries of Tables 3.3 and 3.5 only approximately add to one.

		Count Only Methods[*]			Negative Control De-biased Methods[†]		
λ	α_{RA}	GEE	Mixed Effects	Count Ratio	GEE	Mixed Effects	Simple TND
Type 1 Error: 1.0	1.00	0.078	0.069	0.058	0.085	0.074	0.074
	0.95	0.086	0.078	0.045	0.084	0.073	0.073
	0.85	0.159	0.144	0.054	0.084	0.073	0.074
	0.50	0.898	0.870	0.669	0.084	0.074	0.100
0.8	1.00	0.229	0.205	0.079	0.193	0.179	0.172
	0.95	0.302	0.271	0.113	0.193	0.181	0.172
	0.85	0.490	0.442	0.224	0.194	0.181	0.176
	0.50	0.985	0.979	0.877	0.193	0.184	0.217
0.6	1.00	0.699	0.647	0.402	0.580	0.571	0.547
	0.95	0.769	0.722	0.481	0.580	0.573	0.549
	0.85	0.884	0.853	0.643	0.578	0.573	0.548
	0.50	1.000	0.999	0.980	0.571	0.574	0.583
0.5	1.00	0.900	0.871	0.668	0.813	0.808	0.791
	0.95	0.931	0.909	0.731	0.813	0.806	0.790
	0.85	0.973	0.964	0.835	0.812	0.807	0.791
	0.50	1.000	1.000	0.997	0.804	0.808	0.797
0.4	1.00	0.985	0.979	0.877	0.956	0.953	0.948
	0.95	0.991	0.988	0.907	0.956	0.953	0.948
	0.85	0.998	0.996	0.951	0.956	0.952	0.948
	0.50	1.000	1.000	1.000	0.952	0.951	0.945
0.3	1.00	1.000	0.999	0.980	0.997	0.997	0.997
	0.95	1.000	1.000	0.987	0.997	0.997	0.996
	0.85	1.000	1.000	0.996	0.997	0.997	0.996
	0.50	1.000	1.000	1.000	0.996	0.996	0.996
0.2	1.00	1.000	1.000	1.000	1.000	1.000	1.000
	0.95	1.000	1.000	1.000	1.000	1.000	1.000
	0.85	1.000	1.000	1.000	1.000	1.000	1.000
	0.50	1.000	1.000	1.000	1.000	1.000	1.000

[*] GEE and mixed effects count only models assume Poisson distributed counts and use the canonical log link.

[†] GEE and mixed effects negative control de-biased models assume a binomial distribution and use the canonical logit link.

Table 3.3: The power, and Type I error rates, in testing departure from the null of no intervention effect based on estimation methods with and without de-biasing by negative control counts for a range of Relative Risks (λ), over 10,000 intervention allocations applied to each of 9 historical time periods with 1,000 cases and 4,000 negative controls (when applicable). Differential ascertainment (α_{RA}) is allowed to increase in severity.

		Count Only Methods*			Negative Control De-biased Methods†		
λ	α_{RA}	GEE	Mixed Effects	Count Ratio	GEE	Mixed Effects	Simple TND
1.0	1.00	0.022	0.024	0.022	0.032	0.033	0.032
	0.95	-0.030	-0.027	-0.030	0.031	0.032	0.031
	0.85	-0.132	-0.130	-0.132	0.031	0.032	0.031
	0.50	-0.489	-0.487	-0.489	0.032	0.028	0.032
0.8	1.00	0.017	0.020	0.017	0.025	0.025	0.025
	0.95	-0.024	-0.021	-0.024	0.025	0.024	0.025
	0.85	-0.105	-0.103	-0.105	0.025	0.023	0.025
	0.50	-0.392	-0.390	-0.392	0.025	0.019	0.025
0.6	1.00	0.013	0.015	0.013	0.018	0.016	0.018
	0.95	-0.018	-0.016	-0.018	0.019	0.016	0.019
	0.85	-0.079	-0.077	-0.079	0.019	0.015	0.019
	0.50	-0.294	-0.292	-0.294	0.019	0.012	0.019
0.5	1.00	0.010	0.013	0.010	0.015	0.012	0.015
	0.95	-0.015	-0.013	-0.015	0.015	0.012	0.015
	0.85	-0.066	-0.064	-0.066	0.015	0.011	0.015
	0.50	-0.245	-0.243	-0.245	0.015	0.008	0.015
0.4	1.00	0.008	0.010	0.008	0.012	0.009	0.012
	0.95	-0.012	-0.010	-0.012	0.012	0.008	0.012
	0.85	-0.053	-0.051	-0.053	0.012	0.008	0.012
	0.50	-0.196	-0.195	-0.196	0.012	0.005	0.012
0.3	1.00	0.006	0.008	0.006	0.009	0.005	0.009
	0.95	-0.009	-0.007	-0.009	0.009	0.005	0.009
	0.85	-0.040	-0.038	-0.040	0.009	0.005	0.009
	0.50	-0.147	-0.146	-0.147	0.009	0.002	0.009
0.2	1.00	0.004	0.006	0.004	0.006	0.002	0.006
	0.95	-0.006	-0.005	-0.006	0.006	0.002	0.006
	0.85	-0.026	-0.025	-0.026	0.006	0.001	0.006
	0.50	-0.098	-0.097	-0.098	0.006	-0.001	0.006

Note: GEE and Count Ratio, Simple TND results only appear identical due to rounding.

* GEE and mixed effects count only models assume Poisson distributed counts and use the canonical log link.

† GEE and mixed effects negative control de-biased models assume a binomial distribution and use the canonical logit link.

Table 3.4: Bias in estimation of the intervention Relative Risk for estimation methods with and without de-biasing by negative control counts for a range of Relative Risks (λ), over 10,000 intervention allocations applied to each of 9 historical time periods with 1,000 cases and 4,000 controls (when applicable). Differential ascertainment(α_{RA}) is allowed to increase in severity.

		Count Only Methods*			Negative Control De-Biased Methods†		
λ	α_{RA}	GEE	Mixed Effects	Count Ratio	GEE	Mixed Effects	Simple TND
1.0	1.00	0.922	0.931	0.934	0.915	0.926	0.926
	0.95	0.914	0.922	0.926	0.916	0.927	0.927
	0.85	0.841	0.856	0.859	0.916	0.927	0.926
	0.50	0.102	0.130	0.115	0.916	0.926	0.900
0.8	1.00	0.922	0.931	0.934	0.916	0.926	0.927
	0.95	0.914	0.923	0.926	0.915	0.927	0.926
	0.85	0.841	0.856	0.859	0.916	0.926	0.925
	0.50	0.101	0.131	0.116	0.916	0.926	0.900
0.6	1.00	0.923	0.931	0.934	0.917	0.928	0.927
	0.95	0.915	0.923	0.927	0.915	0.926	0.925
	0.85	0.843	0.860	0.861	0.915	0.925	0.925
	0.50	0.105	0.136	0.119	0.916	0.926	0.899
0.5	1.00	0.922	0.931	0.934	0.916	0.926	0.926
	0.95	0.913	0.922	0.925	0.915	0.926	0.925
	0.85	0.843	0.859	0.860	0.917	0.926	0.926
	0.50	0.106	0.137	0.121	0.916	0.925	0.898
0.4	1.00	0.923	0.931	0.935	0.915	0.926	0.926
	0.95	0.916	0.923	0.927	0.916	0.925	0.926
	0.85	0.844	0.861	0.862	0.916	0.927	0.924
	0.50	0.110	0.144	0.126	0.915	0.925	0.896
0.3	1.00	0.922	0.931	0.934	0.917	0.927	0.927
	0.95	0.914	0.925	0.926	0.916	0.926	0.927
	0.85	0.846	0.864	0.864	0.916	0.927	0.925
	0.50	0.117	0.152	0.134	0.916	0.924	0.895
0.2	1.00	0.922	0.932	0.933	0.916	0.926	0.926
	0.95	0.915	0.926	0.927	0.915	0.924	0.926
	0.85	0.849	0.869	0.867	0.916	0.926	0.924
	0.50	0.130	0.168	0.148	0.916	0.902	0.892

* GEE and mixed effects count only models assume Poisson distributed counts and use the canonical log link.

† GEE and mixed effects negative control de-biased models assume a binomial distribution and use the canonical logit link.

Table 3.5: 95% confidence interval coverage based on estimation of the intervention Relative Risk for estimation methods with and without de-biasing by negative control counts for a range of Relative Risks (λ), over 10,000 intervention allocations applied to each of 9 historical time periods with 1,000 cases and 4,000 negative controls (when applicable). Differential ascertainment (α_{RA}) is allowed to increase in severity.

Chapter 4

Model-based Simulations for Short, Highly Variable Interrupted Time Series Studies: Dengue Surveillance Data

4.1 Introduction

Many public health interventions cannot be implemented using full treatment randomization for ethical, logistical, or economical reasons. When exposure cannot be randomized, analyses face additional threats to their internal and external validity. In randomized trials, the objective is typically to minimize threats to internal validity in order to establish causality. This is accomplished by breaking causal relationships between factors related to selection, maturation, and other sources that may confound any true causal relationship between the exposure and the outcome of interest. Studies that cannot implement randomization are then subjected to a broader range of potential threats to internal validity than their randomized counterparts.[8, 20]

Quasi-experimental designs aim to address many of these validity concerns. An increasingly popular quasi-experiment is the interrupted time series (ITS) design. A time series is comprised of repeated measures often taken at regular points in time. A natural source of time series data includes the regular accumulation of counts for processes under regular surveillance. An ITS is a particular case in which the time series has been interrupted by something such as the implementation of an intervention. The ITS quasi-experiment then uses the pre-interruption data as the null case by observing the outcome trend in the absence of an intervention. The assumption is that this estimated trend provides the counterfactual of what would have continued had the interruption not occurred. The post-interruption data is similarly used to estimate the trend in the outcome in the presence of the intervention.

Due to its straightforward nature and applicability to public health natural experiments, ITS studies are rapidly becoming one of the most popular cases of quasi-experimental designs in the epidemiological literature.[5] Public health ITS studies have been used to investigate everything from campaigns to increase hand washing [56], changes in morbidity and mortality following natural disasters [37], reduced street lighting and automobile casualties [55], and the effect of smoking bans on acute myocardial infarction.[17] In examples such as these, researchers examine exposures that would be logistically impossible or entirely unethical to randomize, often through the use of surveillance data.

A clear advantage ITS has over classic prepost designs — designs that simply compare the average outcome levels from the pre- and post-intervention periods — is that ITS designs can account for secular trends such as seasonality through the use of ARIMA or alternative modeling techniques. ITS can also provide more nuance in the type of effect expected and subsequently estimated in these quasi-experiments. As outlined by Campbell and Cook, there are many ways in which an ITS intervention effect can manifest. Does the intervention have an immediate effect or is there a lag? Does the effect endure (a continuous effect) or does it fade (discontinuous) with time? Does the intervention affect the average outcome level (the intercept) and/or does it cause a change in trend (the slope)? A researcher interested in implementing ITS can consider each of these questions and propose a plausible, and appropriately complex impact model that clearly communicates the expected effect.

In order to estimate and evaluate the hypothesized effect, researchers must develop methods that account for the threats to internal validity. In classical randomized controlled trials, proper randomization of the intervention and extensive control over other experimental conditions mitigate investigator concern that the effect observed is spurious or due to another source. However, in ITS studies it is critical to identify, measure, and adjust for external sources of influence in order to isolate the intervention effect to the best of the researcher's ability. ARIMA models are often the most popular approach for estimating and subsequently removing the "noise" in time series data.[8] Segmented regression adjusting for some function of time is far more applicable when the time series is "short", e.g. less than the 50 recommended observations for adequate ARIMA modeling. As there is already considerable work on ARIMA and long series, we will focus on the setting where the time series is "short".

ITS studies are increasingly common in interventions targeting vector-borne infectious disease. Unlike ITS studies of the impact of particular public policies on relatively steady-state phenomena such as cigarette sales, vector-borne infectious diseases often have drastic, "explosive" outbreaks that are not easily explained by seasonality alone. In stark contrast to these massive outbreaks, there is also an abundance of low-transmission months where it is not unusual to have zero reported cases. The high monthly and seasonal variability, the presence of zero counts, and short follow-up periods complicate standard approaches for model selection and estimation in ITS studies of infectious disease. The ability to simulate data and assess various estimators on metrics such as power, bias, and coverage, is therefore essential in planning a rigorous evaluation of the evidence collected in an infectious disease ITS study.

Valuable work has been done to develop simulations for ITS studies of public health interventions,[69] but very little guidance exists for researchers working with the highly variable, short time series such as those found for infectious disease interventions. This can present a barrier to planning and conducting critical research on the effectiveness of vector-borne disease interventions. A recent sytematic review, described in Chapter 1, highlighted the serious lack of evidence on the effectiveness for a multitude of current vector control methods, including common methods such as spraying.[7] ITS studies may present an approach that is feasible and cost-effective if properly implemented. In planning such trials (prospective or based on natural experiment), simulations may provide significant insight.

This paper explores a model-based framework to generate simulations from complex historical data in order to inform statistical decision making for short ITS analyses. This approach is demonstrated in an application to a planned prospective ITS study evaluating the effectiveness of *Wolbachia* in decreasing the incidence of reported dengue using historic data from Yogyakarta, Indonesia. In the application, we find that splines and at least two years of follow-up are necessary for reliable estimation of the intervention effect in this short ITS analysis. In the discussion, we identify additional ways simulations can be used in designing ITS trials as well as potential simulation alternatives for future work.

4.2 Methods

Exposure

Let A denote a point treatment exposure. Specifically, for a time series with measurements at each time t, A is an intervention occurring at a particular point t' that results in the distinct identification of pre-intervention ($t < t'$) and post-intervention ($t \geq t'$) follow-up. We will focus strictly on ITS with clear pre- and post-intervention periods. In many cases, there is a period of transition from pre-intervention to post-intervention in which the intervention itself is establishing, resulting in a less abrupt demarcation of the pre- and post- intervention periods. Rather than discard the data from this in-between state, other work [12] has shown methods, both deterministic and adaptive, for accounting for this in order to de-bias intervention effect estimates.

It can be helpful, particularly in short, highly variable settings, to have a nearby control region that does not receive the intervention but can be used to adjust for secular trends in the modelling process. The proposed simulation approach can accommodate single and multi-arm ITS study designs.

Outcome

The outcome of interest in an ITS study is commonly a count Y collected across all time points t. The count distributions that often appear in the epidemiological literature include Poisson and negative binomial, in both standard and zero-inflated form. While overdispersion and zero-inflation may not be concerns for certain, well-behaved time series, these complications certainly arise in the realm of vector-borne infectious disease research. It is critical that outcome counts are recorded during both the pre- and post-intervention periods. The pre-intervention counts are used to establish the outcome trend in the absence of an intervention. The post-intervention data is used to establish the counterfactual trend in Y in the presence of the intervention.

Follow-Up

The need to collect information during both the pre- and post-intervention periods of an ITS study necessitates additional points of planning: 1) how much follow-up is necessary? and, 2) is balance between the pre- and post-intervention periods essential? Assuming that the counts Y_t are readily available through existing surveillance systems (e.g. reported monthly case counts), there may be decades of historical data that could be used to establish estimates of the outcome trend in the absence of an intervention. As such, the difficulty arises not necessarily in accessing data for the pre-intervention period, but in identifying the necessary duration of post-intervention follow-up. This is a relevant task for both natural and planned ITS studies. A prospective ITS study is often restricted by limited funding which in turn restricts the length of follow-up that can be expected in any particular study. For a natural experiment, the importance of including an appropriate amount of follow-up is highlighted in nearly all introductory material of this subject.[8]

When the duration of a prospective ITS is restricted to two or three years of post-intervention followup, as is common for prospective grants, this results in fewer than 36 post-intervention monthly counts. One hypothesis may be that the short follow-up could be offset by incorporating a longer pre-intervention series. However, previous work has demonstrated that the gain in power when increasing the number of data points is minimal when the balance between pre- and post- intervention periods is lost. [69] Further, for infectious disease outbreaks, it may be unreasonable to expect the infectious disease trends years prior to the study of interest to be relevant to modern trends, even barring changes in reporting standards or quality.

Simulations from Historical Data

The objective of this model-based simulation procedure is to use sampling rather than forecasting in an attempt to preserve the unique dependencies, fluctuations, and covariances in the observed historical data. For highly variable, short ITS it may be difficult to simulate plausible data for a future that hasn't yet occurred. However, if we can rely on history as a guide, it is far easier to simulate data that looks like what we have already seen. The proposed approach uses the observed counts as estimates of the average rate of the outcome at time t and generates new simulated data from a parametric model fit to the historical data, with modifiable parameters for the various pieces of the impact model. In this way, we can construct data that looks like the historical data *while knowing the true intervention effect*, allowing us to evaluate the performance of various methods to determine which approach will make the most efficient and unbiased use of the data, according to some desirable standard. Further, we can gain insight into the necessary duration of follow-up for reliable effect estimation.

To put this into concrete terms, assume we have access to historical surveillance data of outcome Y_t for $t = 1, \ldots, M$ total time points and plan on carrying out a quasi-experiment of length $m \leq M$, where m includes the duration of pre- and post-intervention followup. The pre- and post-intervention followup need not be balanced, but we will assume there is enough pre-intervention follow-up to establish an outcome trend in the absence of an intervention.

To begin the simulation procedure, draw a sequence $Y_{|m|}$ of m consecutive historical Y_t. Drawing a consecutive sequence $Y_{|m|}$ will preserve any complexities in the transmission of the disease. Next, fit a negative binomial count model to $Y_{|m|}$, flexibly accounting for seasonality or other time-related trends. The model fit of $Y_{|m|}$ will serve as the average rate of Y at each time t across the sequence of length m, denoted in Equation 4.1 as μ_t. Therefore, to simulate the data around these average rates, at each time t, a value for the simulated Y_t^* is then drawn from the negative binomial distribution characterized by Equation 4.2, with the dispersion parameter ν estimated from the data.

$$\mu_t = E[Y_t|t, N_t] = N_t\text{exp}\{\beta_0 + f(t)\} \tag{4.1}$$

$$Y_t^* \sim \text{Negative Binomial}(\mu_t, \mu_t + \frac{\mu_t^2}{\nu}) \tag{4.2}$$

The same procedure can be used when there are multiple arms, Z, by incorporating the proposed arm information into the negative binomial model. Specifically, the model should note which counts belong to which arm of the hypothetical study and should include offsets for differences in population sizes. For example, the average dengue count Y at time t under the absence of an intervention could be modeled as a function of the "arm" Z, some function of time to account for seasonality $f(t)$, and an offset for the population size in each "arm" N_t, which may change with time (Equation 4.3). Note, this model assumes that once seasonality, $f(t)$, has been accounted for, the average rate of the outcome Y in each arm is constant across time. In the event the average rate is not constant across time, an appropriate interaction term between Z and t would be required.

$$\mu_t = [Y_t|Z, t, N_t] = N_t\text{exp}\{\beta_0 + \beta_1 Z + f(t)\} \tag{4.3}$$

To introduce intervention effects into the simulated data, one must simply take advantage of the multiplicative nature of the model. If the proposed impact model shows a step change in the average rate, then the simulated counts in the intervention arm need only be modified by the intervention $RR = \lambda = \text{exp}\{\beta_2 A_t\}$, where A_t is an indicator denoting the intervention status at time t. This will always be equal to zero for the control arm and will only be equal to one for the treatment arm during the post-intervention period, when $t \geq t'$.

This basic approach can be modified to produce simulated data from increasingly complex proposed impact models. For example, if a step change and a trend change during the post-intervention period is anticipated, this can be simulated by multiplying the simulated counts in the intervention arm with $\lambda_1 = \exp\{\beta_2 A_t\}$ and $\lambda_2 = \exp\{\beta_3 A_t \times t\}$. Here, λ_1 captures the immediate discontinuity in rates from pre- to post-intervention period and λ_2 ensures that the post-intervention trend continues to change monotonically with time. A host of flexible functions and model parameterizations can be used to create increasingly complex intervention effects, hopefully reflecting the reality of the underlying data and hypothesized impact model.

4.3 Application

In this section we demonstrate the usefulness of this technique with a real-world application.

Data

Surveillance data of monthly reported dengue hemorrhagic fever case counts from two non-contiguous regions bordering Yogyakarta, Indonesia was available from January 2006 until December 2017. These two regions were identified to be the sites of a controlled prospective ITS study of the effect of *Wolbachia*-infected mosquitoes on the incidence of dengue. The northern region received large-scale *Wolbachia* deployment starting in August 2016 while the southern region was held as an untreated control area. The specifics and results of this trial have been described elsewhere. [26]

The establishment of *Wolbachia* among local *Aedes aegypti* mosquitoes was expected to decrease the level of dengue incidence. In terms of simulating data for the ITS study, we proposed an impact model with a change in step, not necessarily a change in slope. The assumptions are therefore: 1) there has been no systematic increasing or decreasing in the average incidence rate of surveilled dengue infection during the pre-intervention period, 2) the application of the intervention will only affect the intervention arm, 3) the effect of the intervention will not lead to a systematic increasing or decreasing over time in the average incidence rate of dengue infection during the post-intervention period, but rather a step change in the average rate.

Simulations

To simulate ITS data, we used only the historical data preceding any deployment of *Wolbachia*, restricting to only the counts from January 2006 until August 2016. Given the motivating proposed trial was a prospective ITS study with a limited post-intervention period, we examined post-intervention periods of 1, 2, and 3 years each. We considered pre-intervention periods of 3, 4, 5, 6, and 7 years. This allowed us to examine the tradeoffs in estimation between more balanced designs (those with shorter pre-intervention periods) and those with increased series length (those with longer pre-intervention periods). As described in Section 4.2, a negative binomial model was fit to a random consecutive sequence ranging from 6 years to 10 years of historical data. This model was used to simulate the null case, or the trend under no intervention effect. An intervention RR respecting an impact model with a change in step was applied. Simulated intervention effects included RR = 0.2, 0.4, 0.6, 0.8, 1.0 (no effect). Power, bias, and coverage were based on estimation of the intervention effect in 1,000 simulated datasets for each unique duration and intervention effect combination.

Analyses

In this application, we used the simulated data to determine which modeling techniques may be most appropriate in an analysis. We considered three popular methods of controlling for seasonality and secular trends and compared their effects across the simulated datasets. Specifically, we compared modeling approaches where the $f(t)$ term in Equation 4.3 is specified as a cubic spline, Fourier terms, and quarterly indicators. Equation 4.4 lists each of these functions explicitly. The numeric month (e.g. January = 1, December = 12) at time t is denoted by m_t, c_k denotes the left boundaries set by knot k, and ω is the inverse of the period which, in this case, is taken to be 1/12. The model coefficients are represented by γ in each of these models for simplicity and are not expected to be equivalent across models in estimation or interpretation.

$$f(t) = \begin{cases} \gamma_1 t + \gamma_2 t^2 + \sum_{k=1}^{K} \gamma_{1+k} \mathbb{I}\{t \geq c_k\}(t - c_k)^3 & \text{spline} \\ \gamma_1 \sin \omega m_t + \gamma_2 \cos \omega m_t & \text{Fourier series} \\ \gamma_1 \mathbb{I}\left\{m_t \in \{4, 5, 6\}\right\} + \gamma_2 \mathbb{I}\{m_t \in \{7, 8, 9\}\} + \gamma_3 \mathbb{I}\{m_t \in \{10, 11, 12\}\} & \text{quarterly} \end{cases} \tag{4.4}$$

Rather than use the naive estimates of the variance-covariance matrix that are returned by default by most software, we used the HC3 covariance estimator. This should produce heteroskedasticity-consistent estimation of the covariance matrix for the model coefficients.

Simulations and analyses were carried out using R version 3.6.2 “Dark and Stormy Night”.[43] Negative binomial modeling was performed using the “MASS” package.[60] Splines and harmonic terms were incorporated thanks to R packages “spline” [44] and “tsModels”,[42] respectively. All code necessary to recreate this analysis is available at a GitHub repository[1] managed by the corresponding author.

4.4 Results

Assessing Simulation Model Fit

The historical incidence rate data from the two regions in Yogyakarta, Indonesia is shown in Figure 4.1. The intervention status of the two regions has been assigned according to their designations in the prospective quasi-experiment. The *Wolbachia* intervention has not

[1] `https://github.com/sdufault15/short-variable-its`

yet been applied for the range of data shown. It is difficult to identify a clear pattern or even seasonality in the historical data when displayed this way. When a smoother is fit to the grouped monthly data (Figure 4.2) a distinct transmission season begins to appear with rates increasing after the month of October and decreasing after the month of April. However, the variability is still quite high and there does not seem to be a clear relationship in monthly or yearly lags.

A more formal approach to assessing seasonality or secular patterning is through the use of the autocorrelation function (ACF) and partial autocorrelation function (PACF). Figure 4.3A displays the autocorrelation in the original data. There does appear to be autocorrelation up to a lag of 3, but there is no clear evidence of annual seasonality, which would be indicated by spikes at a lag of 12 months. When carried out properly, the residuals between the observed data and the model estimates should be normally distributed and free of autocorrelation. The model-based approach described in Section 4.3 for capturing the underlying secular trends in the historical data do seem to accomplish this task. The ACF and PACF of the residuals are plotted in Figure 4.3B. Though there is a small amount of lingering autocorrelation at a lag of 1, the majority of the autocorrelation has dissipated. Further, Figures 4.3C and 4.3D demonstrate the distribution of the residuals by time and marginally, respectively.

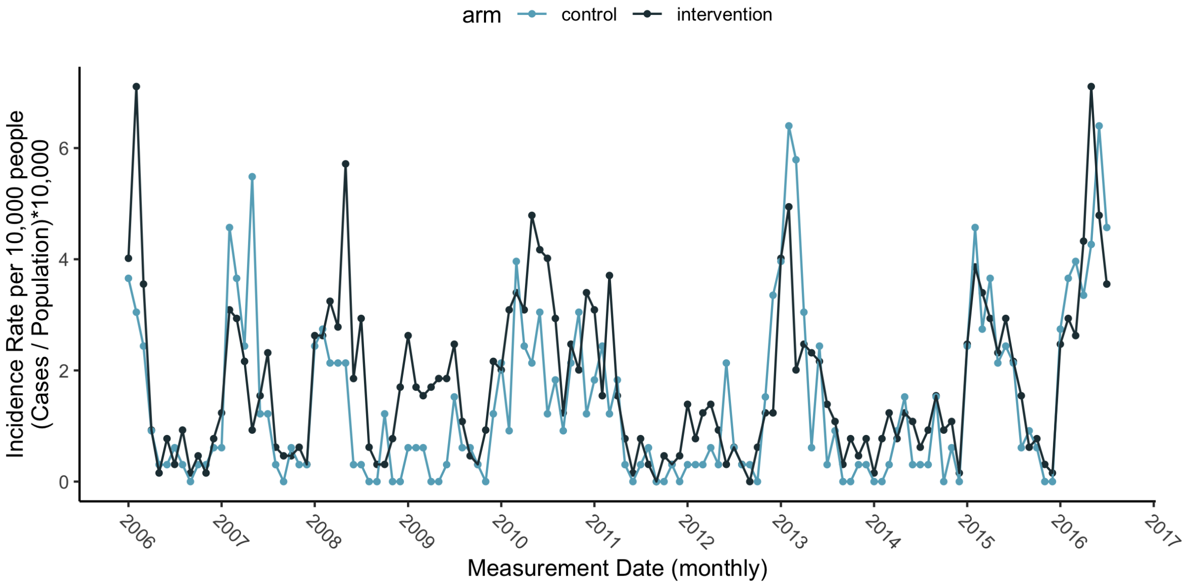

Figure 4.1: Observed monthly historical rates of dengue for two regions near Yogyakarta, Indonesia from January 2006 until July 2016.

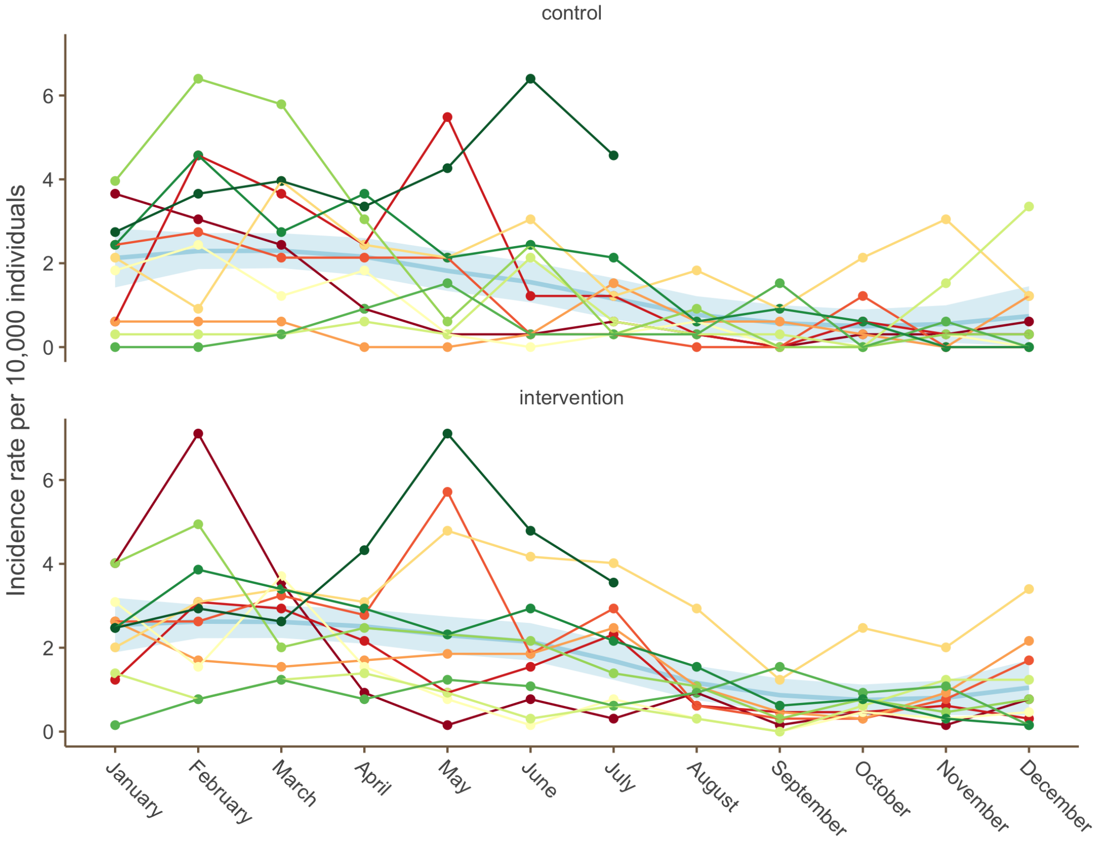

Figure 4.2: Monthly rates of dengue from January 2006 until July 2016 in two regions outside of Yogyakarta, Indonesia. The blue smooth line represents a linear smoother and its 95% confidence interval, as fit to the monthly rates.

Comparing Analysis Model Performance

This section contains the results from applying the model-based approach to simulating data to the example described in Section 4.3. Recall, the objectives in simulating data for this example included determining a reliable modeling approach for estimating the intervention effect, determining a minimum length of post-intervention follow-up, and identifying whether there are benefits to including additional years of pre-intervention data. This is the order in which we will examine the results.

Power, presented in subsequent tables, was estimated as the proportion of 1,000 simulations that returned a treatment coefficient p-value below the *a priori* specified cutoff of $\alpha = 0.05$. This p-value was estimated based on the adjusted covariance matrix. Table 4.1 displays the power and Type I error rates in testing for a departure from the null of no intervention effect based on a negative binomial model with harmonic, cubic spline, or quar-

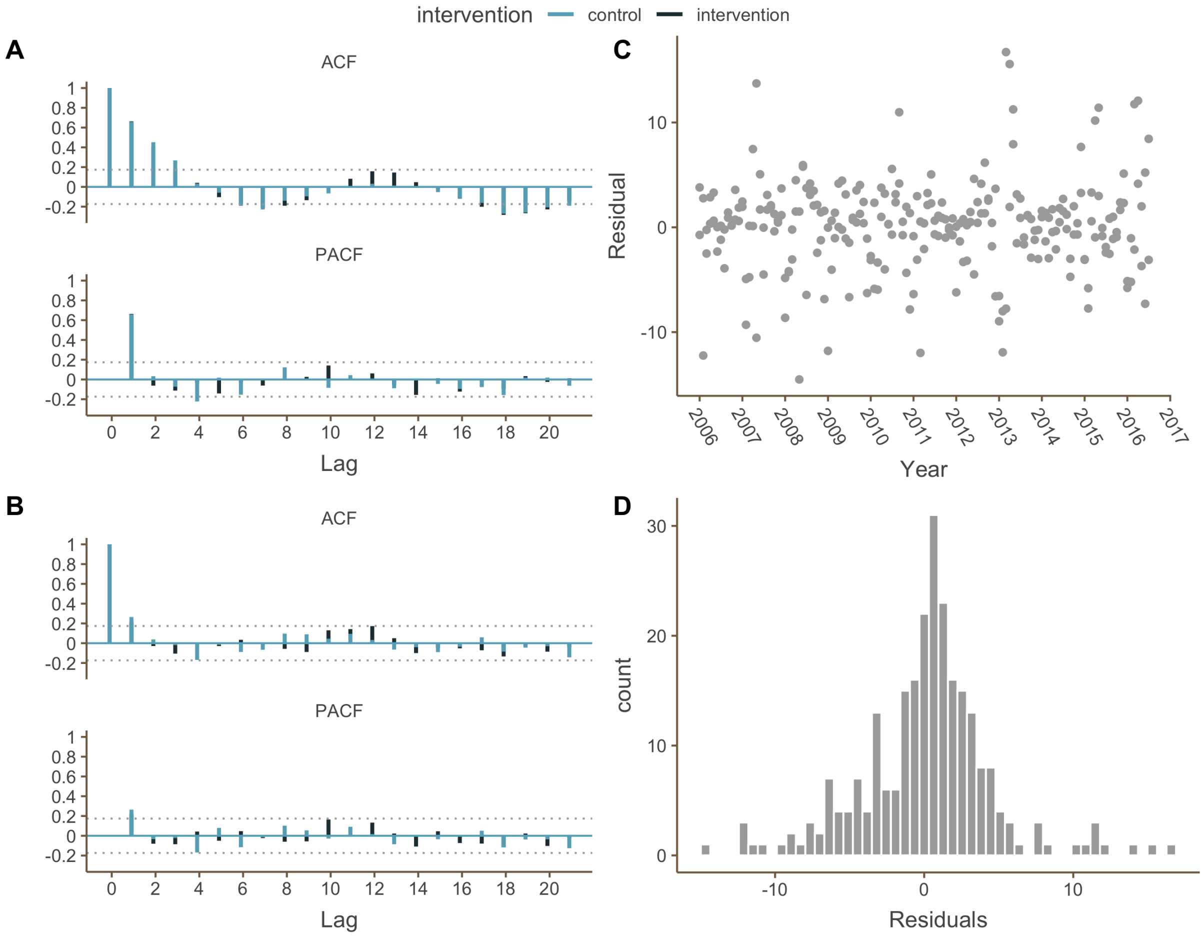

Figure 4.3: Assessing the quality of the model fit versus the observed historical data. The model will be used to simulate ITS data. **A)** The autocorrelation function (ACF) and partial autocorrelation (PACF) plots for the original data. **B**) The autocorrelation function (ACF) and partial autocorrelation (PACF) plots for the model residuals. **C**) The model residuals plotted against time. **D**) The marginal distribution of the model residuals.

		Method		
	Post-intervention follow-up (years)	Harmonic	Quarter	Spline
Type 1 Error	1	0.394	0.381	0.039
	2	0.074	0.063	0.036
	3	0.006	0.004	0.037
RR = 0.8	1	0.326	0.319	0.111
	2	0.435	0.371	0.180
	3	0.293	0.225	0.260
RR = 0.6	1	0.594	0.553	0.411
	2	0.923	0.886	0.693
	3	0.983	0.962	0.855
RR = 0.4	1	0.967	0.939	0.832
	2	0.999	0.998	0.982
	3	1.000	1.000	1.000
RR = 0.2	1	1.000	1.000	0.985
	2	1.000	1.000	1.000
	3	1.000	1.000	1.000

Table 4.1: Power and Type I error rates in testing for a departure from the null of no intervention effect based on a negative binomial model with various methods of adjusting for seasonality and secular trends. All models were GLMs assuming a negative binomial count distribution and using a log link. To adjust for secular trends, the harmonic model used a pair of harmonic terms, the quarter model used quarterly indicator variables, and the spline model used cubic splines with knots at the boundaries of the transmission season. These results are estimated from 1,000 simulations based on ten years of monthly historical dengue incidence data.

terly indicator terms to adjust for seasonality. For concision, the pre-intervention length was fixed at 7-years though the observed trends persisted across the entire range of considered pre-intervention lengths. Generally, power increases as the duration of follow-up increases for all models. The harmonic and quarterly indicator models appear to have higher power for detecting an effect than the spline model, but their Type 1 error is far more variable than that of the spline model. Particularly, when follow-up is short, the Type 1 error is quite high.

Coverage, shown in Table 4.2, was estimated based on the proportion of simulations that returned estimated 95% heteroskedasticity-adjusted confidence intervals that contained the true intervention effect. This estimation was carried out on the log scale for improvements in distribution. As the duration of follow-up increases, the coverage tends to improve for each method. In contrast to power, we now get a clear sense of the superiority of the flexible

		Method		
	Post-intervention follow-up (years)	Harmonic	Quarter	Spline
	1	0.606	0.619	0.961
RR = 1	2	0.926	0.937	0.964
	3	0.994	0.996	0.963
	1	0.661	0.681	0.959
RR = 0.8	2	0.926	0.939	0.965
	3	0.996	0.998	0.965
	1	0.646	0.662	0.961
RR = 0.6	2	0.912	0.929	0.968
	3	0.995	0.998	0.958
	1	0.680	0.688	0.955
RR = 0.4	2	0.929	0.937	0.956
	3	0.990	0.994	0.953
	1	0.730	0.727	0.967
RR = 0.2	2	0.937	0.945	0.964
	3	0.988	0.988	0.969

Table 4.2: 95% confidence interval coverage rates in estimating the intervention effect based on a negative binomial model with various methods of adjusting for seasonality and secular trends. All models were GLMs assuming a negative binomial count distribution and using a log link. To adjust for secular trends, the harmonic model used a pair of harmonic terms, the quarter model used quarterly indicator variables, and the spline model used cubic splines with knots at the boundaries of the transmission season. Confidence interval estimation relied on heteroskedasticity-adjusted standard errors. These results are estimated from 1,000 simulations based on ten years of monthly historical dengue incidence data.

spline model in estimation of the intervention effect. The coverage for the spline model is consistently around 95%. The harmonic and quarterly indicator models have far lower, less consistent coverage by comparison.

Bias, Figure 4.4, is measured as the difference in the estimated intervention effect and the true intervention effect, taken on the log scale. The spline method consistently hovers around a difference of zero for all follow-up lengths, with a relatively symmetric distribution around zero indicating occasional over and under estimation in approximately equal proportion. The models using harmonic terms and quarterly indicators show deviation from zero bias when the follow-up is short (i.e. at or under two years). In these settings, these methods tend to estimate an intervention effect that is smaller than the truth, resulting in an anti-conservative bias. When there are three years of follow-up, this negative behavior seems to

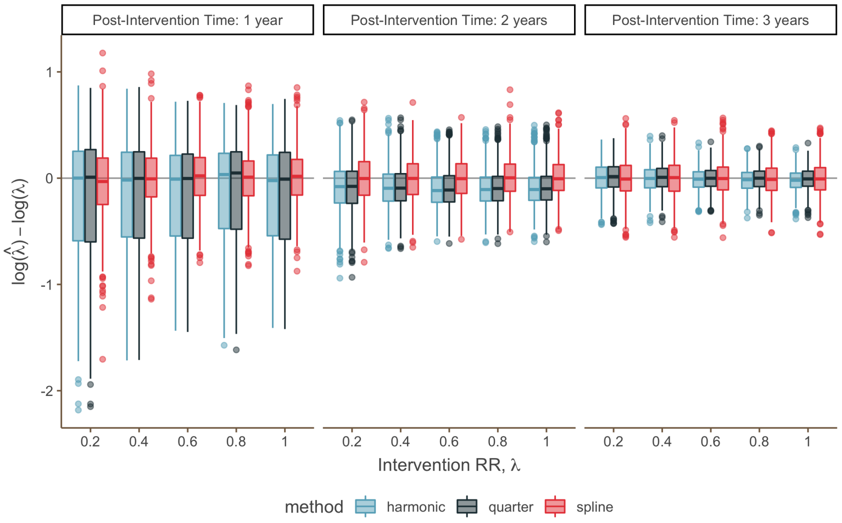

Figure 4.4: The bias in estimation of the intervention effect based on a negative binomial model with various methods of adjusting for seasonality and secular trends. All models were GLMs assuming a negative binomial count distribution and using a log link. To adjust for secular trends, the harmonic model used a pair of harmonic terms, the quarter model used quarterly indicator variables, and the spline model used cubic splines with knots at the boundaries of the transmission season.

adequately disappear and they seem to outperform the spline model in terms of the variance of the estimated bias.

Selecting Follow-Up Time

Next, we examine the importance of balance and duration of follow-up for estimation of the intervention effect. Given the reliable performance of the spline model, these results will be specific to its performance. Table 4.3 presents the power results for the spline model across all pre- and post-intervention follow-up times and intervention effects considered in the simulations. Power was estimated in the manner previously described. Moving row-wise across Table 4.3 demonstrates how critical the extension of follow-up can be in powering a study. In particular, for a relatively substantial, though not unreasonable effect size of RR = 0.6, it is not until there are at least 3 years of follow-up data that the analysis will reach the desirable threshold of 80% power.

To consider the impact of additional pre-intervention data, one can fix the post-intervention follow-up and trace the change in power down the columns. In this simulation study, though adding an additional year of pre-intervention occasionally provides a boost in power, its effect is not nearly as substantial as that of an additional year of follow-up.

Finally, the impact of balance on power can be examined by moving diagonally from right to left across Table 4.3. Power for a fixed trial period (the cumulative pre- and post-intervention time) is highest when there is balance. For example, consider a study that is fixed at 6 total years in length. According to these simulations, if the intervention effect is expected to be RR = 0.6, there is 71.6% power for a trial that divides pre- and post-intervention periods evenly at 3 years each. When balance shifts so that the pre-intervention period is 4 years in length and the post-intervention follow-up is an additional 2 years, power drops to 59.9%. Coverage can be seen in the supplemental materials (Table 4.4) but remains around 95% for all considered pre- and post-intervention combinations. Bias (Table 4.5) reflects the power trends with added years of pre-intervention data having a minor impact on de-biasing estimation, but additional years of post-intervention having a more considerable impact.

4.5 Discussion

This paper explores a flexible parametric model-based approach to simulating highly variable, short ITS data. This work is particularly relevant for evaluating the efficacy of popular interventions targeting vector-borne infectious diseases such as dengue. Simulations can be used to plan a prospective trial as well as identify an appropriate method of analysis. Hopefully, researchers will continue to develop tools such as these to continue to improve evidence generation in settings where random experiments may not be feasible or ethical, but natural experiments do tend to arise.

In order to produce useful simulated data from a parametric model, it is essential that the model appropriately reflects reality. In the time series literature there are a number of recommended metrics and visualizations to determine precisely this. We have shown the results from a few of the most common approaches, but researchers should be thorough in determining whether their model adequately captures the observed data trends and can simulate plausible data. This parametric model-based approach is only as useful as its ability to capture and replicate reality. As is the case with any attempts at power analysis, these approaches should not be seen as a way to extract proscriptive guidelines, but rather, when thoughtfully carried out, as one of many sources of insight that can be incorporated to improve trial design and analysis.

In the application, we find that when historical data is abundant the inclusion of an extra year of historical data in the pre-intervention period is not nearly as valuable as an extra year of follow-up data. This result agrees with common intuition around ITS studies. Further, the example indicates the importance of continuing follow-up until at least two cycles of data on seasonality and other secular trends have been measured. The inconsistency of the trend in power for one year of follow-up as pre-intervention data increases from three to seven years is evidence of the high variability in the data and an inability to observe the effect of interest long enough to untangle it from external transmission trends.

In the setting where researchers are planning an entirely prospective ITS in which the pre-intervention data has yet to be collected, balance in the planned pre- and post-intervention

		Post-intervention follow-up (years)		
Intervention effect	Pre-intervention (years)	1	2	3
Type 1 Error	3	0.052	0.040	0.042
	4	0.048	0.038	0.043
	5	0.043	0.038	0.026
	6	0.040	0.043	0.033
	7	0.062	0.031	0.035
RR = 0.8	3	0.128	0.150	0.189
	4	0.106	0.171	0.189
	5	0.120	0.153	0.220
	6	0.124	0.172	0.259
	7	0.121	0.180	0.237
RR = 0.6	3	0.369	0.605	0.716
	4	0.374	0.599	0.737
	5	0.409	0.674	0.785
	6	0.446	0.669	0.819
	7	0.424	0.680	0.844
RR = 0.4	3	0.810	0.972	0.991
	4	0.820	0.974	0.996
	5	0.820	0.984	0.996
	6	0.830	0.986	1.000
	7	0.840	0.988	0.999
RR = 0.2	3	0.978	1.000	1.000
	4	0.976	1.000	1.000
	5	0.974	1.000	1.000
	6	0.986	1.000	1.000
	7	0.988	1.000	1.000

Table 4.3: Power and Type I error rates in testing for a departure from the null of no intervention effect based on a negative binomial model with a log link. The model adjusted for seasonality via cubic splines with knots at the boundaries of the transmission season and used heteroskedasticity consistent standard error estimation. These results are estimated from 1,000 simulations based on ten years of monthly historical dengue incidence data.

periods produces the most desirable effects in terms of powering the study to detect the intervention effect. With highly variable outcomes of interest and short windows of trial length, it is essential to get sufficient data to estimate both the null effect and the intervention effect well.

Though including additional years of pre-intervention data demonstrated mild improvements in power, there may be settings in which this will not be a desirable approach. One critical consideration before using extensive historical data in simulating and estimating the pre-intervention trend is whether the standards and consistency of reporting have remained constant across the entire range. Additionally, expert knowledge may be useful in identifying a temporal threshold at which historical data is so temporally distant that it is no longer relevant to understanding current trends.

One unignorable limitation to this approach for simulating data is that it relies entirely on access to historical data, particularly a time series that is of an equal or greater length than the intended study. Sampling rather than forecasting to simulate data attempts to preserve the variability of the observed data, but bars the possibility of reliable extrapolation beyond the length of the observed historical time series. Alternatively, simulations based on self-excitatory spatio-temporal point processes and other individual-based models common in the mathematical modeling of infectious disease may be useful for generating highly variable, time series without the reliance on such extensive historical data.[46, 47] One potential direction is to build off of recent work combining stochastic Susceptible-Infected-Recovered (SIR) and Susceptible-Exposed-Infected-Recovered (SEIR) models with a finite population Hawkes model, accounting for absorbing states.[48, 9] In the meantime, the flexible modeling approach proposed here can be readily implemented by most standard statistical software and used to simulate data that reflects reality.

There are many additional directions for future work on this topic. First, if using a parametric model for simulating data, it may be beneficial to use a larger, more flexible model than is intended for analysis. A limitation of the example analysis is that spline-based model used for estimation was correctly specified to the model used for simulating the data. This makes the comparison between the spline, harmonic, and quadratic approaches slightly artificial in that it likely overestimates the quality of performance of the spline model. However, determining a larger model for simulating the data should be done with care and in careful consultation with the set of proposed and plausible impact models.

In the example provided here, we considered a multi-group ITS design where one group was treated and the other was not. As has been shown elsewhere,[34] controlled ITS designs strengthen the internal validity by allowing for additional control over external factors. A complementary approach to strengthening the comparability of the control group is through the use of synthetic controls. Though not explored here, this innovative reweighting approach is useful when there may be multiple relevant control groups from which to create a credible counterfactual for comparison against the treatment group[32] or for identifying an appropriate control group.[34] The simulation procedure described here could be useful in comparing the statistical properties of such methods to a standard multigroup ITS analysis.

In the analysis of a multi-group ITS design with many independent groups, one standard approach is through the use of mixed effects regression models with random effects for cluster and fixed effects for time, treatment effect, and other shared covariate effects. In this way, the models are able to account for within and across cluster variation while still estimating the common intervention effect. This approach can also account for multiple intervention groups receiving treatment at different times. The stepped wedge trial is a distinct case of this type of crossover study, in which all clusters sequentially cross over from untreated to treated over the study period, typically at regularly scheduled intervals. Rather than randomly assigning clusters to treatment or control, clusters are randomly assigned a time to crossover. This study design has an advantage over parallel arm studies when there is high variation at the cluster level as each cluster can serve as its own treatment counterfactual. However, this benefit must be balanced against the inevitable threat of temporal confounding. Building on the work of previous researchers,[25, 18] the simulation procedure described here could be extended to stepped wedge CRTs for insight into relevant analysis metrics and trial decisions.

4.6 Supplement

		Post-intervention follow-up (years)		
Intervention effect	Pre-intervention (years)	1	2	3
Type 1 Error	3	0.955	0.959	0.950
	4	0.974	0.953	0.971
	5	0.963	0.965	0.974
	6	0.955	0.953	0.960
	7	0.961	0.964	0.963
RR = 0.8	3	0.958	0.962	0.956
	4	0.952	0.969	0.965
	5	0.958	0.962	0.947
	6	0.959	0.967	0.974
	7	0.959	0.965	0.965
RR = 0.6	3	0.968	0.959	0.966
	4	0.965	0.970	0.958
	5	0.961	0.967	0.968
	6	0.954	0.947	0.966
	7	0.961	0.968	0.958
RR = 0.4	3	0.950	0.956	0.972
	4	0.968	0.961	0.957
	5	0.958	0.965	0.956
	6	0.965	0.967	0.956
	7	0.955	0.956	0.953
RR = 0.2	3	0.956	0.966	0.956
	4	0.966	0.968	0.959
	5	0.961	0.969	0.968
	6	0.954	0.964	0.966
	7	0.967	0.964	0.969

Table 4.4: Coverage based on 95% confidence intervals around the estimated intervention effect using a negative binomial model with cubic splines to adjust for seasonality and secular trends. These results are estimated from 1,000 simulations based on ten years of monthly historical dengue incidence data.

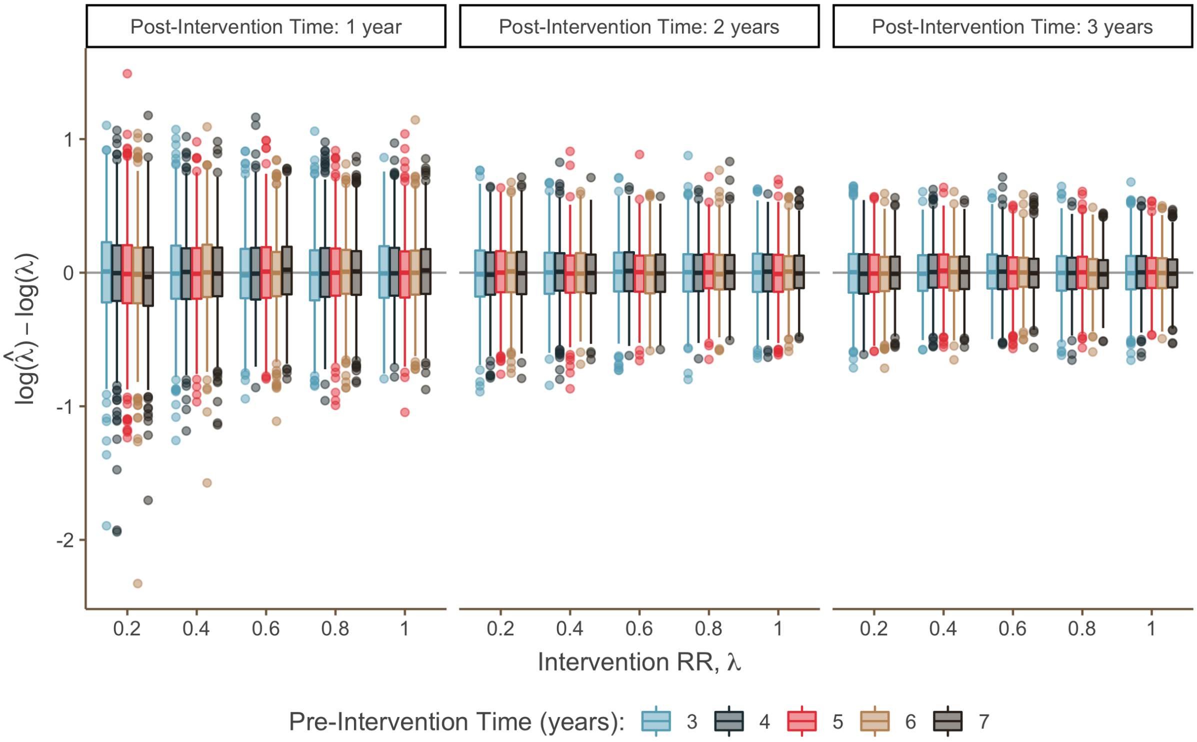

Figure 4.5: The bias in estimation of the intervention effect based on a negative binomial count model with log link. The model used cubic splines with knots at the boundaries of the transmission season to adjust for seasonality and secular trends.

Chapter 5

Conclusion

The objective of this dissertation is to develop statistical methods for the analysis of vector-borne disease preventative techniques wherein the treatment directly targets the vector and the outcome is epidemiological in nature. The motivating example transinfects the bacterium *Wolbachia* into *Aedes aegypti* mosquitoes, the primary vector of dengue, and as such disrupts the transmission of dengue. The focus in Chapters 2 and 3 is statistical analysis in the context of the Cluster-Randomized Test-Negative Design (CR-TND), which was proposed to combine the randomization benefits of the gold standard cluster randomized trials with the sampling benefits of the test-negative design. In Chapter 4, we examine simulations in the context of the interrupted time series (ITS), a strong quasi-experiment and reliable alternative when the CR-TND is infeasible.

In Chapter 2, we propose two distinct estimators of the intervention effect using cluster-level counts and an intention-to-treat approach, assuming no interference: 1) the test-positive fraction estimator, and 2) an aggregate Odds Ratio estimator that is equivalent to the simple TND estimator examined by Jackson et al.[27] Using permutation-based inferential methods, these estimators are compared to the standard mixed effects and generalized estimation equations approaches. In a simple 10 cluster example with small sample size, the GEE and mixed effects models exhibit well-established poor performance in properly estimating variance.[38, 41] In contrast, the proposed aggregate Odds Ratio estimator closely approximates the true permutation variance. In the more extensive simulations regarding 24 clusters and larger sample sizes, the methods perform similarly, suggesting that 24 clusters may be sufficient when sample size is large.

Chapter 3 expands on the work in Chapter 2 by considering CRT estimators that only rely on the case counts. Case count-based analyses are standard practice in the evaluation of CRTs, where counts are commonly aggregated at the cluster level and a cluster-level population offset is used to adjust for differences in cluster size. When case ascertainment is nondifferential by treatment arm, these methods perform well across all metrics considered. However, when case ascertainment is differential by treatment arm, an example of which may be differential health-care seeking behavior in an unblinded CRT, these standard approaches are biased and unreliable. We further show that this bias cannot be resolved by a population offset, but can be mitigated when negative controls are used to approximate the underlying health-care seeking populations. This negative control adjusted estimator is equivalent to the simple TND[27] and aggregate Odds Ratio estimator proposed in Chapter 2.

Chapter 4 considers the ITS design for the evaluation of infectious diseases with patterns that are highly variable and only surveiled for a short window of time, as is common in prospective trials. As ITS designs for public health interventions grow in popularity,[5] research on simulating such data also grows in order to provide guidance in the *a priori* selection of statistical methods, determine an appropriate follow-up length, and inform other critical trial decisions. This chapter aims to contribute a flexible parametric modeling approach for simulating data when extensive, reliable historical data is available. The proposed simulation approach is applied to a real data example to demonstrate a handful of the many questions that simulations may help answer in an endemic infectious disease ITS setting.

Future work on this topic includes the development of per-protocol and individual-level analyses for the CR-TND that can account for interference and individual-level mobility between clusters of varying intervention and control statuses. Additionally, when the CR-TND is infeasible, another design receiving considerable attention is the stepped-wedge quasi-experiment in which each cluster receives the intervention according to a staggered treatment schedule. Considering the complications in inference estimation when there are few clusters available for randomization as well as the threats to unbiased point estimates observed by differential case ascertainment, there is a pressing need for statistical methods that incorporate negative controls and permutation-based inference for stepped wedge designs. Recent literature has examined permutation-based inference methods,[29, 61] but the incorporation of negative controls would be a novel contribution to the growing literature.

Bibliography

[1] Katherine L Anders et al. “Cluster-randomized test-negative design trials: a novel and efficient method to assess the efficacy of community-level dengue interventions”. In: *American Journal of Epidemiology* 187.9 (2018), pp. 2021–2028.

[2] Katherine L Anders et al. “The AWED trial (Applying Wolbachia to Eliminate Dengue) to assess the efficacy of Wolbachia-infected mosquito deployments to reduce dengue incidence in Yogyakarta, Indonesia: Study protocol for a cluster randomised controlled trial”. In: *Trials* 19.1 (2018), pp. 1–16. ISSN: 17456215. DOI: 10.1186/s13063-018-2670-z.

[3] Douglas Bates et al. “Fitting Linear Mixed-Effects Models Using lme4”. In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48. DOI: 10.18637/jss.v067.i01.

[4] Scarlett L Bellamy et al. “Analysis of dichotomous outcome data for community intervention studies”. In: *Statistical Methods in Medical Research* 9.2 (2000), pp. 135–159.

[5] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. “Interrupted time series regression for the evaluation of public health interventions: a tutorial”. In: *International Journal of Epidemiology* 46.1 (2017), pp. 348–355.

[6] Samir Bhatt et al. “The global distribution and burden of dengue”. In: *Nature* 496.7446 (2013), pp. 504–507.

[7] Leigh R Bowman, Sarah Donegan, and Philip J McCall. “Is dengue vector control deficient in effectiveness or evidence?: Systematic review and meta-analysis”. In: *PLoS Neglected Tropical Diseases* 10.3 (2016).

[8] Donald Thomas Campbell and Thomas D Cook. *Quasi-experimentation: Design & analysis issues for field settings.* Rand McNally College Publishing Company Chicago, 1979.

[9] Adam Chaffee. “Comparative Analysis of SEIR and Hawkes Models for the 2014 West Africa Ebola Outbreak”. PhD thesis. UCLA, 2017.

[10] Huiying Chua et al. “The use of test-negative controls to monitor vaccine effectiveness”. In: *Epidemiology* in press (2019).

[11] J Cornfield. “Randomization by group: a formal analysis”. In: *American Journal of Epidemiology* 108.2 (1978), pp. 100–102.

[12] Maricela Cruz, Miriam Bender, and Hernando Ombao. “A robust interrupted time series model for analyzing complex health care intervention data”. In: *Statistics in Medicine* 36.29 (2017), pp. 4660–4676.

[13] *Dengue and severe dengue*. Mar. 2020. URL: `https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue`.

[14] Heverton Leandro Carneiro Dutra et al. “Wolbachia blocks currently circulating Zika virus isolates in Brazilian Aedes aegypti mosquitoes”. In: *Cell Host & Microbe* 19.6 (2016), pp. 771–774.

[15] Jill M Ferdinands et al. “Re:“Invited Commentary: Beware the Test-Negative Design””. In: *American Journal of Epidemiology* 185.7 (2017), pp. 613–613.

[16] Colin B Fogarty et al. “Randomization inference and sensitivity analysis for composite null hypotheses with binary outcomes in matched observational studies”. In: *Journal of the American Statistical Association* 112.517 (2017), pp. 321–331.

[17] Antonio Gasparrini, Giuseppe Gorini, and Alessandro Barchielli. “On the relationship between smoking bans and incidence of acute myocardial infarction”. In: *European Journal of Epidemiology* 24.10 (2009), pp. 597–602.

[18] Alan J Girling and Karla Hemming. “Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models”. In: *Statistics in medicine* 35.13 (2016), pp. 2149–2166.

[19] M Haber et al. “A probability model for evaluating the bias and precision of influenza vaccine effectiveness estimates from case-control studies”. In: *Epidemiology & Infection* 143.7 (2015), pp. 1417–1426.

[20] Margaret A Handley et al. “Selecting and improving quasi-experimental designs in effectiveness and implementation research”. In: *Annual Review of Public Health* 39 (2018), pp. 5–25.

[21] Richard J Hayes and S Bennett. “Simple sample size calculation for cluster-randomized trials”. In: *International Journal of Epidemiology* 28.2 (1999), pp. 319–326.

[22] Richard J Hayes and Lawrence H Moulton. *Cluster Randomised Trials*. Second. CRC Press, 2017.

[23] Kirsten Hilgenboecker et al. “How many species are infected with Wolbachia?–a statistical analysis of current data”. In: *FEMS Microbiology Letters* 281.2 (2008), pp. 215–220.

[24] Søren Højsgaard, Ulrich Halekoh, and Jun Yan. “The R Package geepack for Generalized Estimating Equations”. In: *Journal of Statistical Software* 15/2 (2006), pp. 1–11.

[25] Michael A Hussey and James P Hughes. “Design and analysis of stepped wedge cluster randomized trials”. In: *Contemporary Clinical Trials* 28.2 (2007), pp. 182–191.

[26] Citra Indriani et al. “Reduced dengue incidence following deployments of *Wolbachia*-infected *Aedes aegypti* in Yogyakarta, Indonesia”. In: *Lancet Global Health* (submitted).

[27] Michael L Jackson and Jennifer C Nelson. “The test-negative design for estimating influenza vaccine effectiveness”. In: *Vaccine* 31.17 (2013), pp. 2165–2168.

[28] Nicholas P Jewell et al. “Analysis of cluster-randomized test-negative designs: cluster-level methods”. In: *Biostatistics* 20.2 (2019), pp. 332–346.

[29] Xinyao Ji et al. “Randomization inference for stepped-wedge cluster-randomized trials: an application to community-based health insurance”. In: *The Annals of Applied Statistics* 11.1 (2017), pp. 1–20.

[30] Karyn N Johnson. “The impact of Wolbachia on virus infection in mosquitoes”. In: *Viruses* 7.11 (2015), pp. 5705–5717. ISSN: 19994915. DOI: 10.3390/v7112903.

[31] Luke Keele, Dylan Small, and Richard Grieve. “Randomization-based instrumental variables methods for binary outcomes with an application to the fffdfffdfffdIMPROV-Efffdfffdfffdtrial”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.2 (2017), pp. 569–586.

[32] Ariel Linden. “Combining synthetic controls and interrupted time series analysis to improve causal inference in program evaluation”. In: *Journal of Evaluation in Clinical Practice* 24.2 (2018), pp. 447–453. ISSN: 13652753. DOI: 10.1111/jep.12882.

[33] Marc Lipsitch, Eric Tchetgen Tchetgen, and T Cohen. “Negative controls: a tool for detcting confounding and bias in observational studies”. In: *Epidemiology* 21.3 (2010), pp. 383–388.

[34] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. “The use of controls in interrupted time series studies of public health interventions”. In: *International journal of epidemiology* 47.6 (2018), pp. 2082–2093.

[35] John M Marshall and Omar S Akbari. *Gene Drive Strategies for Population Replacement.* Elsevier Inc., 2015, pp. 169–200. ISBN: 9780128004050. DOI: 10.1016/B978-0-12-800246-9.00009-0. URL: http://dx.doi.org/10.1016/B978-0-12-800246-9.00009-0.

[36] Daniel McNeish and Laura M Stapleton. “Modeling clustered data with very few clusters”. In: *Multivariate Behavioral Research* 51.4 (2016), pp. 495–518.

[37] Ai Milojevic et al. “Health effects of flooding in rural Bangladesh”. In: *Epidemiology* (2012), pp. 107–115.

[38] Jorge G Morel, MC Bokossa, and Nagaraj K Neerchal. “Small sample correction for the variance of GEE estimators”. In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 45.4 (2003), pp. 395–409.

[39] Kevin Mortimer et al. “A cleaner burning biomass-fuelled cookstove intervention to prevent pneumonia in children under 5 years old in rural Malawi (the Cooking and Pneumonia Study): a cluster randomised controlled trial”. In: *The Lancet* 389.10065 (2017), pp. 167–175.

[40] *Mosquito-borne diseases*. URL: `https://www.who.int/neglected_diseases/vector_ecology/mosquito-borne-diseases/`.

[41] Wei Pan and Melanie M Wall. “Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations”. In: *Statistics in Medicine* 21.10 (2002), pp. 1429–1441.

[42] Roger D Peng and with contributions from Aidan McDermott. *tsModel: Time Series Modeling for Air Pollution and Health*. R package version 0.6. 2013. URL: `https://CRAN.R-project.org/package=tsModel`.

[43] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: `https://www.R-project.org/`.

[44] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: `https://www.R-project.org/`.

[45] Stephanie M Rainey et al. “Understanding the Wolbachia-mediated inhibition of arboviruses in mosquitoes: progress and challenges”. In: *Journal of General Virology* 95.3 (2014), pp. 517–530.

[46] Alex Reinhart et al. “A review of self-exciting spatio-temporal point processes and their applications”. In: *Statistical Science* 33.3 (2018), pp. 299–318.

[47] Alex Reinhart et al. “Rejoinder: A review of self-exciting spatio-temporal point processes and their applications”. In: *Statistical Science* 33.3 (2018), pp. 330–333.

[48] Marian-Andrei Rizoiu et al. “SIR-Hawkes: linking epidemic models and Hawkes processes to model diffusions in finite populations”. In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 419–428.

[49] Franklin E Satterthwaite. “An approximate distribution of estimates of variance components”. In: *Biometrics Bulletin* 2.6 (1946), pp. 110–114.

[50] Donald S Shepard, Eduardo A Undurraga, and Yara A Halasa. “Economic and disease burden of dengue in Southeast Asia”. In: *PLoS neglected tropical diseases* 7.2 (2013).

[51] Donald S Shepard et al. “Economic impact of dengue illness in the Americas”. In: *The American journal of tropical medicine and hygiene* 84.2 (2011), pp. 200–207.

[52] Cameron P Simmons et al. “Dengue”. In: *New England Journal of Medicine* 366.15 (2012), pp. 1423–1432.

[53] DM Skowronski et al. "Effectiveness of vaccine against medical consultation due to laboratory-confirmed influenza: results from a sentinel physician pilot project in British Columbia, 2004–2005". In: *Canada Communicable Disease Report* 31.18 (2005), pp. 181–91.

[54] Dylan S Small, Thomas R Ten Have, and Paul R Rosenbaum. "Randomization inference in a group–randomized trial of treatments for depression: covariate adjustment, noncompliance, and quantile effects". In: *Journal of the American Statistical Association* 103.481 (2008), pp. 271–279.

[55] Rebecca Steinbach et al. "The effect of reduced street lighting on road casualties and crime in England and Wales: controlled interrupted time series analysis". In: *Journal of Epidemiology and Community Health* 69.11 (2015), pp. 1118–1124.

[56] Sheldon Paul Stone et al. "Evaluation of the national Cleanyourhands campaign to reduce Staphylococcus aureus bacteraemia and Clostridium difficile infection in hospitals in England and Wales by improved hand hygiene: four year, prospective, ecological, interrupted time series study". In: *BMJ* 344 (2012).

[57] Sheena G Sullivan, Shuo Feng, and Benjamin J Cowling. "Influenza vaccine effectiveness: potential of the test-negative design. A systematic review". In: *Expert Review of Vaccines* 13.12 (2014), p. 1571.

[58] Sheena G Sullivan, Eric J Tchetgen Tchetgen, and Benjamin J Cowling. "Theoretical basis of the test-negative study design for assessment of influenza vaccine effectiveness". In: *American Journal of Epidemiology* 184.5 (2016), pp. 345–353.

[59] SM Tam. "On covariance in finite population sampling". In: *The Statistician* (1985), pp. 429–433.

[60] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: `http://www.stats.ox.ac.uk/pub/MASS4`.

[61] Rui Wang and Victor De Gruttola. "The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials". In: *Statistics in Medicine* 36.18 (2017), pp. 2831–2843.

[62] Bernard L Welch. "The generalization of student's' problem when several different population variances are involved". In: *Biometrika* 34.1/2 (1947), pp. 28–35.

[63] Bernard L Welch. "The significance of the difference between two means when the population variances are unequal". In: *Biometrika* 29.3/4 (1938), pp. 350–362.

[64] Daniel Westreich and Michael G Hudgens. "Invited commentary: beware the test-negative design". In: *American Journal of Epidemiology* 184.5 (2016), pp. 354–356.

[65] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: `https://ggplot2.tidyverse.org`.

[66] Anne L Wilson et al. "Evidence-based vector control? Improving the quality of vector control trials". In: *Trends in Parasitology* 31.8 (2015), pp. 380–390.

[67] Jun Yan. "geepack: Yet Another Package for Generalized Estimating Equations". In: *R-News* 2/3 (2002), pp. 12–14.

[68] Jun Yan and Jason P Fine. "Estimating Equations for Association Structures". In: *Statistics in Medicine* 23 (2004), pp. 859–880.

[69] Fang Zhang, Anita K Wagner, and Dennis Ross-Degnan. "Simulation-based power calculation for designing interrupted time series analyses of health policy interventions". In: *Journal of Clinical Epidemiology* 64.11 (2011), pp. 1252–1261.