RANDOMIZATION INFERENCE AND PRINCIPAL STRATIFICATION IN HIV PREVENTION STUDIES

Tracy L. Nolen, MStat

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics.

Chapel Hill
2012

Approved by:

Michael G. Hudgens, Ph.D.
Stephen Cole, Ph.D.
Gary Koch, Ph.D.
Pranab K. Sen, Ph.D.
Dennis Wallace, Ph.D.

Abstract

TRACY L. NOLEN, MSTAT: RANDOMIZATION INFERENCE AND PRINCIPAL STRATIFICATION IN HIV PREVENTION STUDIES
(Under the direction of Michael G. Hudgens, Ph.D.)

Infectious disease prevention studies often aim to test or estimate the "causal effect" of a preventive measure on outcomes by comparing the potential outcomes individuals would have under treatment versus control. Examples of outcomes of interest include infection incidence or post-infection outcomes (e.g., disease severity, death). Two analytical challenges of interest exist for these studies. First, analyses on post-infection outcomes are subject to selection bias as only a subset of the randomized population become infected (i.e., infection status is a post-randomization measure on which analyses are often conditioned). Treatment comparisons conditional on post-randomization measures using standard analytic methods do not have a causal interpretation in that the estimates obtained are not unbiased estimates of the contrast between potential outcomes. The principal stratification framework provides estimates of treatment causal effect in the presence of potential selection bias due to post-randomization measures; however, existing methods comprise only Bayesian or large-sample frequentist approaches. To date a general approach to randomization inference within principal strata has not been developed. Furthermore, while principal stratification approaches are abundant in statistical literature, their presence as an applied analytic approach within infectious disease journals is limited. The second challenge of prevention studies involves the analysis of repeated low-dose mucosal challenge preclinical studies of potential vaccines which are becoming more prevalent in an attempt to conduct studies that better mirror 'real life' human transmission. Current statistical literature exploring the analysis of these studies is somewhat limited and simulation results for certain proposed analytic approaches have demonstrated an inflated type I error. Therefore, in this dissertation we 1) develop methods for exact randomization-based causal inference within principal strata in the presence selection bias due to post-randomization measures, 2) present a discussion of selection bias in randomized studies and the use of principal stratification analytic approaches for handling such bias targeted at subject-matter investigators and 3) present a discussion of appropriate analytic approaches for repeated low-dose challenge preclinical vaccine studies.

Acknowledgments

I owe immeasurable gratitude to my husband, Kurt Nolen and parents, Nancy and David Robinson for their unwavering support and strength as well as my dissertation advisor, Dr. Michael Hudgens for his patience and guidance.

Table of Contents

List of Tables

List of Figures

Chapter 1

Literature Review

1.1 Motivating Examples

1.1.1 Mother-to-Child Transmission of HIV

In sub-Saharan Africa, approximately 40% of new human immunodeficiency virus (HIV) infections occur via breast milk transmission from mother to child (Mofenson 2009). Options such as replacement feeding in place of breastfeeding that have made HIV mother-to-child transmission (MTCT) through breast milk extremely uncommon in developed countries are simply not viable solutions in resource limited settings such as sub-Saharan Africa. Specifically, the cost of replacement feeding and increased risk of other diseases such as diarrhea, pneumonia and malnutrition associated with replacement feeding in these settings make replacement feeding an impractical approach to reducing MTCT in resource-limited settings (Mofenson 2009). As such, an important area of research in MTCT reduction involves identifying an effective prophylactic treatment for use on the mother and/or the infant prior to and/or during breastfeeding. Examples include the Petra trial which assessed the efficacy of prophylactic therapy given to women during pregnancy and 1 week postpartum, the ZEB study which assessed the efficacy of early weaning, the Mma Bana study which assessed the efficacy of prophylactic therapy given to women during pregnancy and breastfeeding and the

BAN study which assessed the efficacy of prophylactic therapy given to women and infants during breastfeeding (Petra Study Team 1999; Kuhn, Aldrovandi and Sinkala et al. 2008; Shapiro, Hughes and Ogwu et al. 2010; Chasela, Hudgens and Ogwu et al. 2010).

HIV MTCT prevention studies pose interesting analytical challenges. As evidenced by the examples provided above, the prophylactic treatments of interest are typically started during or shortly after birth; therefore, randomization must occur before or soon after birth. Additionally, HIV MTCT can occur *in utero*, perinatal and postpartum (Mock 1999). When the effect of interest is the ability of the treatment to reduce postpartum breast milk transmission, some of the randomized infants infected *in utero* or perinatal are not relevant for the analyses of interest. As such, *in utero* and perinatal infections must be accounted for when estimating treatment effect during the breastfeeding period (Mofenson 2010).

An additional challenge of MTCT prevention studies is that post-infection endpoints are often of interest for secondary analyses. For example, a secondary outcome of interest in the ZEB study was post-infection survival (Kuhn, Aldrovandi and Sinkala et al. 2008). Analysis of such outcomes are challenging because such outcomes are only observed for a portion of the randomized infants, the subset that become infected, and infected infants typically comprise a small proportion of the study population (Mofenson 2010). As such, approaches that appropriately control for infection status and do not require large sample frequentist assumptions are required.

1.1.2 HIV Vaccine Development in Macaques

Preclinical proof-of-concept vaccine trials using animal models limit the risk, time and cost of clinical trials involving human subjects by providing preliminary evidence of potential safety and efficacy of an investigational vaccine (Koff 2006; Shedlock, Silvestri and Weiner 2009). While chimpanzees are the only non-human primate that can be infected with HIV-1, research on chimpanzees is limited due to ethical and financial considerations as chimpanzees are endangered and expensive to maintain (Shedlock et al. 2009; Smith 2009). Therefore a large portion of the preclinical studies of HIV vaccines have been conducted using macaques and viral surrogates of HIV, simian immunodeficiency viruses (SIVs), as the disease progression of SIVs in macaques mirrors that of HIV in humans (Shedlock et al. 2009).

The virus challenge in these preclinical trials has historically been administered via a single high-dose intravenous or mucosal inoculation which often resulted in near guaranteed infection of all animals (Hudgens et al. 2009). Although single high-dose challenge studies are appealing in that high infection rates allow for a greater chance of observing an effect of vaccine assuming the vaccine is completely protective against infection, the vaccine efficacy in these trials may not translate directly to vaccine efficacy in 'real life'. For example, the high infection rates of these challenge studies do not mirror the low per heterosexual probability of HIV transmission per sexual act, estimated at < 0.01 in various studies of US, European, Thai and African populations (Gray et al. 2001; Boily et al. 2009) . Likewise, per month probability of late postnatal HIV transmission via breastfeeding is estimated at 0.01 (WHO 2007). Additionally, it is unlikely that vaccines are equally efficacious against high-dose and low-dose challenges. Therefore potential vaccines that would be efficacious against low-dose challenges may be discarded because they were not observed to be efficacious in the high-dose challenge studies (Regoes, Longini, Feinberg, and Staprans 2005).

As an alternative, repeated low-dose mucosal challenge studies that attempt to more accurately mirror 'real life' human transmission have been proposed and are becoming more prevalent (McDermott et al. 2004; Hessell et al. 2009; Hudgens et al. 2009). For these studies, each individual is challenged multiple times (i.e., until infected or until a maximum number of challenges C_{max} has been performed) such that each individual has the potential for multiple outcomes corresponding to the number of times the individual was challenged. Currently literature exploring the design of these repeated low-dose challenge studies is somewhat limited (Regoes et al. 2005; Hudgens and Gilbert 2009; Hudgens et al. 2009). Specifically, Regoes et al. (2005) performed a statistical power analysis of repeated low-dose challenge studies that involved summarizing outcome data using a 2×2 contingency table of infection status by treatment assignment where the cell counts are the number of challenge events falling in each category, and analyzing the results using a one-tailed Fisher's exact test. Hudgens et al. (2009) performed additional power analyses using simulations which showed that in certain settings, this approach had an inflated type I error (rates as high as 0.20 for scenarios simulated). Therefore additional research into the appropriate analytic approaches for these designs is warranted.

1.2 Causal Inference

1.2.1 Introduction

The primary goal of most randomized and non-randomized studies is to estimate the effect of an intervention on an outcome of interest. This effect is often described as the 'causal effect' of the intervention and a framework for inference about this effect is Rubin's causal model (Holland 1986). For example, the causal effect of aspirin on the outcome of resolution of a headache is the measure of the aspirin's ability to cure a headache. This can be quantified as an odds ratio (e.g., the ratio of the odds of the headache resolving when aspirin is taken versus the odds of the headache resolving when no aspirin is taken). The causal effect of vaccine on disease incidence is often measured by vaccine efficacy, defined as the reduction in disease incidence among vaccinated individuals compared to the incidence in unvaccinated individuals (Hudgens and Halloran 2006).

For Rubin's model, consider a setting where a population of n individuals (units) are observed at a specific time and place and each individual receives a particular treatment (intervention). For ease of discussion, assume there are only two potential treatments and let Z_i represent treatment assignment for the i^{th} individual such that $Z_i = 0$ for control and $Z_i = 1$ for treatment. When discussing causal effects, it should be hypothetically possible for each individual to receive any of the potential interventions; a concept often referred to as "No causation without manipulation" (Rubin 1978, Holland 1986). Denoting the outcome measure of interest as $Y_i(Z_i)$, there are then 2 potential outcomes, $Y_i(0)$ for the true outcome under control and $Y_i(1)$ for the true outcome under treatment. The fundamental problem of causal inference is that it is impossible to observe both potential outcomes on any one individual (Holland 1986). Therefore, define the observed outcome for each individual as $Y_i^{obs} \equiv Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ (Rosenbaum 2010). Because of this fundamental problem, potential outcomes are often described in terms of counterfactuals (Robins and Greenland 1992). Specifically, the counterfactual circumstance is the potential outcome that would have been observed for an individual under the treatment scenario that individual did not receive (e.g., what would have happened to an individual that was treated had they received control instead).

Under this framework, an example of a measure of the causal effect of treatment on an individual is the difference in potential outcomes, $Y_i(1) - Y_i(0)$. Because both potential outcomes cannot be observed for an individual, it is not possible to make direct inference about $Y_i(1) - Y_i(0)$ for each individual. Therefore, Neyman (1923) proposes estimating or testing the 'typical' causal effect of the intervention such as the average causal effect for the observed population of individuals, $n^{-1} \sum_i (Y_i(1) - Y_i(0))$. To make inference about the average causal effect of an intervention, the treatment assignment mechanism must be specified or modeled (Rubin 2005). A significant characteristic of Rubin's framework is that, in the setting of a randomized trial, the treatment assignment mechanism is "ignorable" in that it is independent of the potential outcomes for each individual (Rosenbaum and Rubin 1983, Little and Rubin 2000). In other words, $Pr(Z_i|Y_i(1), Y_i(0)) = Pr(Z_i)$ for all i. Therefore, assuming $n/2$ individuals are randomly assigned to each treatment group, an unbiased estimate of the average causal effect is simply the difference in the observed treatment group means of our population, $(n/2)\sum_i(Y_i^{obs}(Z_i) - Y_i^{obs}(1 - Z_i))$ (Neyman 1923).

Analyses using the potential outcome model have been employed in a variety of settings including observational and randomized studies in the health and social science and econometric fields (Winship and Morgan 1999, Heckman and Vytlacil 1999, Greenland 2000, Sobel 2000).

1.2.2 Randomization-Based Inference

Randomized studies are the gold standard in clinical trials for evaluating the effects of biomedical interventions because randomization (i) produces in expectation comparable groups with respect to measured and unmeasured covariates and (ii) provides a basis for statistical inference. Regarding (i) balance of covariates between treatment groups allows any observed difference between treatment groups to be attributed to treatment assignment and therefore, allows for estimation and testing of the causal effect of treatment (Rubin 2008, Foster 2010). Regarding (ii) randomization-based inference of treatment effects is based on distributions created from the randomization process rather than assuming random sampling from an infinite population or that particular parametric distributions hold (Koch, Gillings and Stokes 1980; Rubin 1991; Rosenbaum 2002a). Specifically when comparing the effect of two treatments, randomization allows for the calculation of the exact probability of how unusual the observed difference in effect between the treatment groups is under a specified hypothesis where the calculation of these probabilities relies solely only on the randomization design implemented for the study, the specified hypothesis and the observed data (Rubin 1974, Rosenbaum 2010). Therefore, randomization-based inference allows for inference in small to intermediate sample size settings where methods based on asymptotic approximations may not be appropriate.

To further explore randomization-based inference, consider the sample of n individuals introduced in Section 1.2.1 to be a finite, fixed population that are randomly assigned treatment using a fixed randomization scheme such that $n/2$ individuals are randomly assigned to each treatment group. For randomization-based inference, the set of potential outcomes for each individual is considered a fixed feature of the population and therefore is denoted as $(y_i(0), y_i(1))$ (Rosenbaum2010). As treatment assignment is random and Y_i^{obs} is a function of treatment assignment, Y_i^{obs} is considered to be random. Assume the null hypothesis of interest is that there is no effect of treatment on the outcome of interest, $H_0 : y_i(0) = y_i(1)$ for $i = 1, \ldots, n$. Additionally, assume the one-sided alternative hypothesis where the response under control is greater than that under treatment (e.g., disease incidence is more likely or disease severity is greater in the control group compared to the treated group). The primary question of randomization-based inference is "To what extent do the data from the randomized experiment provide evidence against the null hypothesis of no treatment effect?" (Rosenbaum 2010). The null hypothesis specified here is referred to as Fisher's sharp null hypothesis of no effect and implies that each individual would experience the same outcome regardless of whether they received treatment or control (Rubin 2005). This implication is important as it means both potential outcomes for each individual are observed under the null and therefore, the distribution of any test statistic can be obtained by calculating the value of the test statistic for each possible treatment assignment combination using the observed outcome data (Rubin 2005, Foster 2010).

Specifically recall that like any pretreatment covariates, $y_i(0)$ and $y_i(1)$ are fixed features of the population while treatment assignment is assumed to be random. Therefore, randomized trials are "unconfounded" in that the potential outcomes are independent of treatment assignment (i.e., treatment is ignorable, $\Pr(Z_i|y_i(0), y_i(1)) = Pr(Z_i)$) and as such there are $\binom{n}{n/2}$ possible treatment assignment combinations that are all equally likely to be selected (Rubin 2008, Rosenbaum 2010). Since $y_i(0) = y_i(1)$ under the null hypothesis, H_0 provides a unique setting where both potential outcomes are observed for each individual and therefore, Y_i^{obs} is fixed regardless of treatment assignment. For this simple setting a randomization-based definition of the p-value is the probability of obtaining a treatment assignment combination that results in a distribution of outcomes as or more extreme than what was observed when assuming all treatment assignments are equally likely.

1.2.3 Intermediate Post-Randomization Outcomes

In many situations, assessing the causal effects of treatment on the outcome of interest in randomized studies is complicated by the presence of an intermediate post-randomization outcome. Rubin (2005) used the term "concomitant variable" and describes it as"an outcome variable that is not the outcome of primary outcome of interest, but may be on the causal pathway of the treatment affecting the primary outcome variable" (Rubin 2005). An example of interest for this research is HIV prevention studies where an outcome of interest is disease severity (Hudgens, Hoering and Self 2003). For this example, the intermediate post-randomization outcome is occurrence of disease as only individuals who become diseased will have a measurable level of disease severity.

Other examples include assessing causal treatment effect in the presence of competing risks or informative censoring such as truncation by death or loss-to-follow and in the presence of treatment noncompliance. For studies with the competing risk of truncation by death, the intermediate post-randomization outcome is occurrence or death as the outcome of interest may only be observed or be of interest for individuals who survive (Robins 1995). In randomized studies where there is treatment noncompliance, compliance to randomized treatment assignment can be considered an intermediate post-randomization outcome as compliance status impacts the outcome of interest (Angrist, Imbens and Rubin 1996). Various statistical approaches have been used to conduct analyses conditional on these intermediate post-randomization outcomes; however, not all of these approaches truly allow for estimation of the causal effects of treatment on the outcome of interest. For example, analyses of disease severity often condition on infection status such that only infected individuals are included in the analyses or are performed via a composite analysis of both disease incidence and severity (Mick 2010; Chang, Guess and Heyse 1994). Analyses in the presence of non-compliance are often performed using an intention-to-treat approach where all individuals are classified by their assigned treatment group regardless of actual compliance (Lachin 2000).

To explore this topic further, amend the scenario described in Section 1.2.2 by denoting the intermediate post-randomization outcome as $s_i(Z_i)$. Now each individual has four potential outcomes $(s_i(1), s_i(0), y_i(1), y_i(0))$ where all four of these potential outcomes are considered fixed features of the finite population of individuals and Z_i is still considered a random variable. Thus define the observed intermediate post-randomization outcome, S_i^{obs} analogously to Y_i^{obs} such that $S_i^{obs} \equiv Z_i s_i(1)+(1-Z_i)s_i(0)$. Hence, S_i^{obs} and Y_i^{obs} are both random variables since they depend on Z_i.

One standard method that adjusts for intermediate post-randomization outcomes is referred to as the 'net-treatment effect adjusting for the intermediate post-randomization outcome' (Frangakis and Rubin 2002). This approach is a comparison of the distribution of the observed outcome of interest under each treatment assignment among individuals with the same observed outcome for the intermediate post-randomization outcome:

$$Pr[Y_i^{obs}|S_i^{obs} = s, Z_i = 0] \text{ compared to } Pr[Y_i^{obs}|S_i^{obs} = s, Z_i = 1] \tag{1.1}$$

which, in our randomized study setting, is equivalent to the comparison of:

$$Pr[Y_i^{obs}|s_i(0)] \text{ compared to } Pr[Y_i^{obs}|s_i(1)] \tag{1.2}$$

If treatment has any effect on the intermediate post-randomization outcome, then the net-treatment effect conditioning on the intermediate post-randomization outcome does not equal the causal effect of treatment on the outcome of interest (Robins and Greenland 1992). Specifically, if treatment has an effect on $s_i(Z_i)$, the net-treatment effect estimate is subject to post-randomization selection bias as the individuals with $S_i^{obs} = s$ under control are not necessarily the same as those with $S_i^{obs} = s$ under treatment (Rosenbaum 1984). Due to the potential introduction of selection bias, the analytical benefits of conducting a randomized study are lost when conditioning on an intermediate post-randomization outcome. Halloran and Struchiner (1995) and Hernán, Hernández-Díaz and Robins (2004) provide further details about the selection bias present in standard analytic methods when assessing a treatment's effect on the outcome of interest in the presence of intermediate post-randomization outcomes.

Alternative analytic approaches have been proposed for assessing causal effects in these types of settings. For example, Angrist, Imbens and Rubin (1996) propose the use of instrumental variables in the presence of treatment non-compliance in order to estimate the causal effect of treatment within the subgroup of individuals considered compliers. Robins and Greenland (1992) describe a process where a covariate that 'controls/explains' the intermediate variable is identified and incorporated into a G-computation algorithm in order to obtain estimates of the direct/causal effects of treatment on the outcome of interest. Of particular interest, Frangakis and Rubin (2002) describe the framework of principal stratification as an alternative analytic approach for obtaining estimates of causal effects of treatment on the outcome of interest.

1.2.4 Principal Stratification

Frangakis and Rubin (2002) formally introduced the concept of principal stratification in causal inference as an alternative to the standard approaches that do not allow for a causal interpretation. While the framework proposed by Frangakis and Rubin is applicable to a wide range of settings including observational studies; for our purposes, we restrict our focus to the setting of randomized studies. As such, what we refer to as the "intermediate post-randomization outcome" is more generally described in Frangakis and Rubin as the "post-treatment variable for which adjustment is required". Frangakis and Rubin define principal stratification as a "cross-classification of the units based on their joint potential values of [the intermediate post-randomization outcome] under each of the treatments being compared" and subsequently define principal effects as the "comparison of treatments within principal strata". The attraction of principal stratification is that, because the potential outcomes are not affected by treatment assignment then membership in the principal strata is also not affected by treatment assignment and therefore principal strata membership can be treated in a manner similar to pretreatment covariates. For example, by comparing potential outcomes for a common set of people (i.e., those within a particular principal strata), this approach allows for causal inference.

Though principal stratification was not formally defined until 2002, the approach was used previously in specific settings (e.g., Baker, Wax and Patterson 1993; Frangakis and Rubin 1999; Rubin 2000). For example, Frangakis and Rubin (1999) show that the standard intent-to-treat (ITT) analytic approach can be biased in the presence of noncompliance and missing outcome data and proposed an alternate test using the framework of potential outcomes where principal strata are defined as compliers (those that only take experimental therapy when assigned) and never-takers (those that never take experimental therapy regardless of treatment assignment). Theoretically, the principal strata defined by compliance also includes always takers (those that always take experimental therapy regardless of treatment assignment) and defiers (those that only take experimental therapy when not assigned); however, under the construct of a random trial, it is assumed that only subjects assigned experimental therapy have access to the therapy and therefore, these latter two strata do not exist.

Frangakis and Rubin (2002) provided the following formal definitions and properties:
Definition: The "basic principal stratifications" P_0 with respect to post-treatment variable s is the partition of units $i = 1, \ldots, n$ such that, within any set of P_0, all units have the same vector $(s_i(1), s_i(0))$.
Definition: A "principal stratification" P with respect to post-treatment variable s is a partition of the units whose sets are unions of sets in the basic principal stratification P_0.
Definition: Let s_i^P indicate the stratum of P to which unit i belongs. The "principal effect" with respect to that principal stratification is defined as a comparison of potential outcomes under the two treatments within a principal stratum c in P, i.e., a comparison between the ordered sets $y_i(1) : s_i^P = c$ and $y_i(0) : s_i^P = c$
Property: The stratum s_i^P to which individual i belongs is unaffected by treatment for any principal stratification P.
Property: Any principal effect is a causal effect

Therefore, conditioning on principal stratum membership allows for inference of the causal effect of treatment on the outcome of interest. However, because only a subset of the potential outcomes are observed, either $(s_i(1), y_i(1))$ or $(s_i(0), y_i(0))$ depending on treatment assignment additional assumptions about individual membership to the principal strata and missing outcomes are required in order to make inference about principal effects (Frangakis and Rubin 2002).

Methods for inference within principal strata often appeal to large sample frequentist or Bayesian theory. Assumptions typically used to aid in the identification of principal strata membership and draw inference within strata include the stable unit treatment value assumption, independent treatment assignment, and monotonicity. For example, in the infectious disease setting, the principal strata of interest are the always infected (those infected regardless of treatment), protected (those infected under control but not under treatment), harmed (those infected under treatment but not under control) and immune (those not infected regardless of treatment) strata. Monotonicity implies a person who is infected when treated would also become infected if not treated and therefore is a member of the always infected group. However, additional assumptions are needed in order to completely identify always infected strata membership in the control group. Assumptions in the form of selection bias models have been suggested to attain identifiability (e.g., see Gilbert et al. 2003; Shephard et al. 2006). These models are helpful if one can elicit prior information regarding the selection bias model parameter (Scharfstein, Halloran, Chu and Daniels 2006; Shepherd, Gilbert and Mehrotra 2007). Alternatively, large sample bounds of the distribution of the outcome of interest in the control group and therefore of treatment effect can be obtained assuming maximum possible levels of positive and negative selection bias (Zhang and Rubin 2003; Hudgens et al. 2003; Imai 2008). These upper and lower bound estimates of treatment effect provide the full range of estimates consistent with the observed data. To draw inference about these estimates, large sample frequentist methods such as profile likelihood CIs (Hudgens and Halloran 2006) or bootstrap tests (Gilbert et al. 2003; Mehrotra et al.

2006) have been employed.

For HIV prevention studies with an objective of understanding the effects of a preventive treatment (e.g. vaccine administered prior to infection) on post-infection events, methods have been developed to assess causal treatment effects on post-infection outcomes in the always infected principal strata (Hudgens, Hoering and Self 2003; Gilbert, Bosch and Hudgens 2003; Mehrotra, Li and Gilbert 2006; Shepherd, Gilbert, Jemiai and Rotnitzky 2006; Jemiai, Rotnitzky, Shepherd and Gilbert 2007; Shepherd, Gilbert and Lumley 2007). Similarly, in studies to prevent MTCT of HIV infection the outcome of interest is long term HIV infection status among infants not infected at or shortly after birth (Chasela et al. 2010). When infants are randomly assigned treatment at birth, the principal stratum of interest is individuals who would not be infected shortly after birth regardless of treatment assignment. Other settings where principal stratification has been applied include treatment noncompliance (Angrist, Imbens and Rubin 1996; Baker, Frangakis and Lindeman 2007; Jin and Rubin 2009), truncation by death (Robins 1995; Zhang and Rubin 2003; Rubin 2006), and evaluation of surrogate endpoints (Gilbert and Hudgens 2008; Joffe and Greene 2009).

1.3 Summary

In summary, assessing the causal effect of treatment on infection occurrence and post-infection outcomes in infection-prevention studies is an important goal. However, interest in post-infection outcomes, small sample sizes and characteristics of certain study designs can make addressing this goal challenging even when randomization is employed. Specifically, while large sample frequentist principal stratification causal inference approaches are available, a randomization-based approach for inference within principal strata has not been developed and the presence of principal stratification approaches in the applied literature is limited. Additionally, published approaches for analyzing data from repeated low-dose mucosal challenge preclinical vaccine studies have been shown to have inflated type I error.

Accordingly, we first developed methods for exact randomization-based causal inference within principal strata in the presence selection bias due to post-randomization measures. Characteristics of the new method are supported by mathematical proofs and simulations (e.g., proof that resulting p-value is a valid p-value and simulations to assess power) as well as comparisons to ITT-based approaches using composite outcomes. The developed approach was expanded by augmenting the test to allow for adjustment for covariates and creating confidence intervals for the causal effect by inverting the proposed test. Second, we present a general discussion of selection bias in randomized studies and principal stratification approaches for handling such bias targeted at subject-matter investigators in order to expand knowledge and use of these analytic approaches. Lastly, we present a discussion of appropriate analytic approaches for repeated low-dose challenge preclinical vaccine studies aimed at preclinical vaccine researchers.

Chapter 2

Randomization-Based Inference within Principal Strata

2.1 Introduction

2.1.1 Principal Stratification

Sometimes in randomized studies, treatment comparisons conditional on intermediate post-randomization outcomes are of interest. For example, in vaccine studies, a common question of interest is whether infections in vaccinated individuals are more or less severe than infections in unvaccinated individuals (Hudgens and Halloran 2006). Unfortunately, the estimands underlying standard methods typically employed for these comparisons do not have a causal interpretation (Rosenbaum 1984). To address this deficiency, Frangakis and Rubin (2002) proposed a general framework for comparing treatments adjusting for the intermediate post-randomization outcomes. In particular, they defined causal effect estimands within strata determined by a cross-classification of individuals defined by the joint potential intermediate post-randomization outcomes under each of the treatments being compared. Since these "principal strata" are not affected by treatment assignment, they can be conditioned on just as any pretreatment covariate. Accordingly, causal effect estimands within principal strata do not suffer from the complications of standard post-randomization adjusted estimands.

The simple framework of principal stratification has a wide range of applications. For example, in human immunodeficiency virus (HIV) prevention studies an objective is understanding the effects of a preventive treatment (e.g., vaccine administered prior to infection) on post-infection events, such as severe disease or death. Assessing a treatment's effect on post-infection outcomes is challenging since such outcomes may only be defined for infected individuals and standard comparisons between infected treated individuals and infected controls are subject to selection bias (Halloran and Struchiner 1995; Hernán, Hernández-Díaz and Robins 2004). Moreover, because the set of individuals who would become infected if assigned treatment is likely not identical to the set of those who would become infected if not assigned treatment, comparisons that condition on infection do not have a causal interpretation. Recently, methods have been developed to assess causal treatment effects on post-infection outcomes in the principal strata of individuals who would be infected regardless of treatment assignment (Hudgens, Hoering and Self 2003; Gilbert, Bosch and Hudgens 2003; Mehrotra, Li and Gilbert 2006; Shepherd, Gilbert, Jemiai and Rotnitzky 2006; Shepherd, Gilbert and Lumley 2007). Similarly, in studies to prevent mother-to-child HIV transmission the outcome of interest is long term HIV infection status among infants not infected at or shortly after birth (Chasela et al. 2010). When infants are randomly assigned treatment at birth, the principal stratum of interest is individuals who would not be infected shortly after birth regardless of treatment assignment. Other settings where principal stratification has been applied include treatment noncompliance (Angrist, Imbens and Rubin 1996; Baker, Frangakis and Lindeman 2007), truncation by death (Robins 1995; Zhang and Rubin 2003) and evaluation of surrogate endpoints (Gilbert and Hudgens 2008; Joffe and Greene 2009).

Methods for inference within principal strata often appeal to large sample frequentist or Bayesian theory. Assumptions typically used to aid in the identification of principal strata membership and draw inference within strata include the stable unit treatment value assumption, independent treatment assignment and monotonicity. However, additional assumptions are needed in order to completely identify principal strata membership in the both treatment groups. For example, assumptions in the form of selection bias models have been suggested to attain identifiability (e.g., see Gilbert et al. 2003; Shephard et al. 2006). These models are helpful if one can elicit prior information regarding the selection bias model parameter (Scharfstein, Halloran, Chu and Daniels 2006; Shepherd, Gilbert and Mehrotra 2007). Alternatively, large sample bounds of the distribution of the outcome of interest in the control group and therefore of treatment effect can be obtained assuming maximum possible levels of positive and negative selection bias (Zhang and Rubin 2003; Hudgens et al. 2003; Imai 2008). These upper and lower bound estimates of treatment effect provide the full range of estimates consistent with the observed data. To draw inference about these estimates, large sample frequentist methods such as profile likelihood CIs (Hudgens and Halloran 2006) or bootstrap tests (Gilbert et al. 2003; Mehrotra et al. 2006) have been employed.

2.1.2 Randomization-Based Inference

Randomized studies are the clinical trial gold standard for evaluating treatment effects because randomization (*i*) produces in expectation comparable groups with respect to measured and unmeasured covariates and (*ii*) provides a basis for statistical inference. Regarding (*ii*) randomization inference is based on distributions created from the randomization process rather than assuming random sampling of individuals from an infinite population (Koch, Gillings and Stokes 1980; Rubin 1991; Rosenbaum 2002a). Unfortunately, the benefits of conducting a randomized study are lost when conditioning on an intermediate post-randomization outcome, as the treatment and control groups are no longer comparable. Ideally one would like to conduct randomization-based inference within principal strata determined by the set of intermediate potential outcomes. However, while randomization inferential methods have been proposed in the instrumental variable setting (Rosenbaum 1996; Rosenbaum 2002a; Imbens and Rosenbaum 2005; Hansen and Bowers 2009), to date a general approach to randomization inference within principal strata has not been developed.

Another benefit of randomization-based inference is that the methods are exact, allowing for inference in small to intermediate sample size settings where methods based on asymptotic approximations may be inappropriate (Imbens and Rosenbaum 2005). In the HIV vaccine setting, small trials are often employed to screen possible vaccines for larger Phase III efficacy studies (Rida, Fast, Hoff and Fleming 1997). For instance, Mehrotra et al. (2006) describe a proof-of-concept (POC) efficacy trial where the study is ceased after just 50 HIV infections are observed in the vaccine and placebo arms combined. In these small sample settings, Bayesian inference about treatment effects within principal strata may not be ideal if investigators are hesitant to make assumptions regarding prior distributions. On the other hand, large sample frequentist methods may lead to incorrect inferences in such settings. For example, simulation studies have demonstrated inflated type I error of bootstrap tests and under-coverage of bootstrap and Wald based confidence intervals (CIs) when the principal stratum of interest is small (Hudgens et al. 2003; Gilbert et al. 2003; Shepherd et al. 2007; Jemiai, Rotnitzky, Shepherd, and Gilbert 2007). It will be seen that the proposed method lifts these limitations.

2.1.3 Outline

This paper considers randomization-based methods for inference within principal strata. The main development is an exact test for a causal treatment effect within principal strata. In Section 2.2, the principal stratum exact test (PSET) is developed. Section 2.3 presents simulation results comparing the PSET to a large sample frequentist approach for testing a treatment effect within principal strata. Section 2.4 includes examples of applications of the PSET. Section 2.5 includes some empirical comparisons between the PSET and intent-to-treat (ITT) based tests. Section 2.6 describes an extension of the PSET to allow for adjustments for covariates. In Section 2.7 exact CIs for treatment effect are derived by inverting the PSET.Areas for future work are discussed in Section 2.8 and proofs are provided in Section 2.9.

2.2 Principal Stratum Exact Test

2.2.1 Assumptions and Notation

Suppose there are n individuals assigned to treatment or control. Assume:

A.1 Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980): Treatment assignment of one individual does not affect another individual's outcomes (no interference) and there are not multiple versions of treatment.

Under SUTVA, let $s_i(z)$ denote the potential intermediate post-randomization outcome and $y_i(z)$ denote the outcome of interest of the i^{th} individual given treatment assignment z, where $z = 0$ for control and $z = 1$ for treatment. Assume the intermediate post-randomization outcome is binary. For ease of presentation, assume the intermediate outcome represents infection status. As such, $s_i(z) = 1$ if the i^{th} individual is infected when assigned treatment z and $s_i(z) = 0$ if uninfected. The principal strata are formed by classifying individuals according to their pair of infection potential outcomes $(s_i(1), s_i(0))$. The always-infected (AI) principal stratum is defined as the individuals with $s_i(0) = s_i(1) = 1$, i.e., individuals who would be infected regardless of treatment assignment. Similarly the harmed stratum is defined as those individuals with $s_i(0) = 0, s_i(1) = 1$; the protected stratum by $s_i(0) = 1, s_i(1) = 0$; and the immune (never-infected) stratum by $s_i(0) = s_i(1) = 0$.

The goal of this paper is to develop a principal stratum exact test of treatment effect on a post-infection outcome, y, among individuals within a principal stratum. Assume the stratum of interest is the AI stratum such that the desired comparison is between $\{y_i(1) \colon s_i(0) = s_i(1) = 1\}$ and $\{y_i(0) \colon s_i(0) = s_i(1) = 1\}$. While motivated by infectious disease settings where the AI stratum is of interest, the PSET is applicable to alternative strata such as the immune stratum as well as other settings. For example, if the intermediate variable represents compliance status, the principal strata of interest might be those that are always compliant regardless of assigned treatment (Angrist et al. 1996). Likewise, if the intermediate variable is survival status, the principal strata of interest might comprise those that always survive regardless of treatment assignment (Zhang and Rubin 2003).

To develop a PSET of treatment effect on the post-infection outcome, consider testing the sharp null hypothesis

$$H_0 \colon y_i(1) = y_i(0) \text{ for all } i \in \mathscr{A}\mathscr{I}, \tag{2.1}$$

where $\mathscr{A}\mathscr{I} \equiv \{i \colon s_i(1) = s_i(0) = 1\}$ is the set of individuals in the AI stratum. Using terminology of VanderWeele (2008), the null (2.1) corresponds to no principal stratum direct effect. An exact test requires the resulting p-value, p, be exact in the sense that $\Pr[p \leq \alpha] \leq \alpha$ for each $\alpha \in [0, 1]$ under the null (Casella and Berger 2002).

While each individual has four potential outcomes $(s_i(1), s_i(0), y_i(1), y_i(0))$, only two of these outcomes are observed dependent on treatment assignment, either $(s_i(1), y_i(1))$ or $(s_i(0), y_i(0))$. Let Z_i denote the treatment assignment for individual i and let $\boldsymbol{Z} = (Z_1, ..., Z_n)$. To make inference about treatment effect, the treatment assignment mechanism must be specified or modeled. This paper uses randomization inference whereby the randomization distribution induced by the experimental design forms the basis for statistical inference (Rubin 1991). In particular, the potential outcomes are considered fixed features of the finite population of individuals while Z_i is considered a random variable. Let $S_i^{obs} \equiv Z_i s_i(1) + (1 - Z_i)s_i(0)$ denote the observed intermediate post-randomization outcome and define Y_i^{obs} analogously. Both S_i^{obs} and Y_i^{obs} are random variables since they depend on Z_i. To develop a test of (2.1), assume independent treatment assignment:

A.2 Independent treatment assignment: $\Pr[\boldsymbol{Z} = \boldsymbol{z}] = \Pr[\boldsymbol{Z} = \boldsymbol{z}']$ for any $\boldsymbol{z}, \boldsymbol{z}'$ such that $\sum_{i=1}^n z_i = \sum_{i=1}^n z_i'$ where $\boldsymbol{z} = (z_1, ..., z_n)$, $\boldsymbol{z}' = (z_1', ..., z_n')$ are treatment assignment vectors.

If principal stratum membership was known, for the AI stratum in particular, the development of an exact test of (2.1) would be straight forward. As assumptions A.1 and A.2 are generally not sufficient to identify principal stratum membership, an additional assumption often made is that treatment does not cause infections:

A.3 Monotonicity: $s_i(1) \leq s_i(0)$ for all $i \in \{1, \ldots, n\}$.

Assumption A.3 identifies AI stratum membership for individuals assigned to treatment. Specifically, A.3 implies infected treated individuals (i.e., $S_i^{obs} = Z_i = 1$) would have become infected if assigned control (i.e., $s_i(0) = 1$) and are therefore members of the AI stratum i.e., $\{i\colon S_i^{obs} = 1, Z_i = 1\} \subseteq \mathscr{A}\mathscr{I}$. Unfortunately, A.1-A.3 do not identify AI stratum membership for individuals in the control group because infected control individuals are a mixture of members of the AI and protected strata.

In Section 2.2.3 the PSET of (2.1) is developed under A.1-A.3. In infectious disease settings A.1 may be violated due to interference between individuals, although this is unlikely in certain settings such as mother-to-child transmission studies. A.2 generally holds in randomized studies. While A.3 cannot be verified from the observable data, it has testable implications. For example, A.3 implies a non-negative average causal treatment effect on infection, i.e., $n^{-1}\sum_{i=1}^{n}\{s_i(0)-s_i(1)\}\geq 0$. Should the data provide evidence to the contrary, A.3 can be rejected. Even if the proportion infected is not higher in the treated arm, the veracity of A.3 may be questionable in some settings. For example, results from a recent HIV vaccine trial (Buchbinder et al. 2008) suggest certain vaccine recipients were more likely to be infected than placebo recipients. Similar concerns arise in vaccine development for other viruses (Greenwood 1997, Tirado and Yoon 2003). A vaccine that causes many infections is likely of no utility, making inference about post-infection endpoints moot. However, if a vaccine causes a few infections but prevents many more, then effects on post-infection endpoints are of interest but invoking A.3 may be dubious. Violations of A.3 are discussed further in Section 2.3.

2.2.2 Example with Binary Outcome

Suppose for now that $y_i(z)$ is a binary variable where $y_i(z)$=1 if the event of interest occurs (e.g., death or severe disease), 0 otherwise. To test (2.1), first imagine we know exactly which individuals are in $\mathscr{A}\mathscr{I}$. Then the following 2×2 table can be constructed

	Event	No Event	
Treatment	$\sum_{i\in\mathscr{A}\mathscr{I}} Z_i Y_i^{obs}$	$\sum_{i\in\mathscr{A}\mathscr{I}} Z_i(1-Y_i^{obs})$	$\sum_{i\in\mathscr{A}\mathscr{I}} Z_i$
Control	$\sum_{i\in\mathscr{A}\mathscr{I}}(1-Z_i)Y_i^{obs}$	$\sum_{i\in\mathscr{A}\mathscr{I}}(1-Z_i)(1-Y_i^{obs})$	$\sum_{i\in\mathscr{A}\mathscr{I}}(1-Z_i)$
	$\sum_{i\in\mathscr{A}\mathscr{I}} Y_i^{obs}$	$\sum_{i\in\mathscr{A}\mathscr{I}}(1-Y_i^{obs})$	m

(2.2)

where $m \equiv \sum_{i \in \mathscr{A}\mathscr{I}} 1$ is the number of individuals in $\mathscr{A}\mathscr{I}$. Under the sharp null, $Y_i^{obs} = y_i(0)$ implying (2.2) can equivalently be written as

	Event	No Event	
Treatment	$\sum_{i \in \mathscr{A}\mathscr{I}} Z_i y_i(0)$	$\sum_{i \in \mathscr{A}\mathscr{I}} Z_i(1 - y_i(0))$	$\sum_{i \in \mathscr{A}\mathscr{I}} Z_i$
Control	$\sum_{i \in \mathscr{A}\mathscr{I}}(1 - Z_i)y_i(0)$	$\sum_{i \in \mathscr{A}\mathscr{I}}(1 - Z_i)(1 - y_i(0))$	$\sum_{i \in \mathscr{A}\mathscr{I}}(1 - Z_i)$
	$\sum_{i \in \mathscr{A}\mathscr{I}} y_i(0)$	$\sum_{i \in \mathscr{A}\mathscr{I}}(1 - y_i(0))$	m

(2.3)

For randomization-based inference, the potential outcomes are fixed features of the finite population. The column totals of (2.3) depend only on the potential outcomes and thus can be considered fixed. Therefore, conditional on the row totals, (2.1) can be tested by applying Fisher's exact test to (2.2) where the p-value is obtained by calculating the probability of each possible table using the hypergeometric distribution.

Because principal strata membership is not completely known, we cannot construct (2.2). Instead the following table of infected individuals is observable

	Event	No Event	
Treatment	$\sum Z_i S_i^{obs} Y_i^{obs}$	$\sum Z_i S_i^{obs}(1 - Y_i^{obs})$	$\sum Z_i S_i^{obs}$
Control	$\sum(1 - Z_i)S_i^{obs} Y_i^{obs}$	$\sum(1 - Z_i)S_i^{obs}(1 - Y_i^{obs})$	$\sum(1 - Z_i)S_i^{obs}$
	$\sum S_i^{obs} Y_i^{obs}$	$\sum S_i^{obs}(1 - Y_i^{obs})$	$\sum S_i^{obs}$

(2.4)

where here and in the sequel $\sum$ denotes the summation over $i = 1, \ldots, n$.

To develop an exact test of (2.1), information from (2.4) can be used to make inference about the unobservable table (2.3). Under A.3, $Z_i S_i^{obs} = 1$ implies $i \in \mathscr{A}\mathscr{I}$.

Thus assuming A.3, under (2.1) the observable table (2.4) can be written as

$$\begin{array}{c|cc|c}
 & \text{Event} & \text{No Event} & \\
\hline
\text{Treatment} & \sum_{i \in \mathscr{A}\mathscr{I}} Z_i y_i(0) & \sum_{i \in \mathscr{A}\mathscr{I}} Z_i(1 - y_i(0)) & \sum_{i \in \mathscr{A}\mathscr{I}} Z_i \\
\text{Control} & \sum (1 - Z_i) S_i^{obs} y_i(0) & \sum (1 - Z_i) S_i^{obs} (1 - y_i(0)) & \sum (1 - Z_i) S_i^{obs} \\
\hline
 & \sum S_i^{obs} y_i(0) & \sum S_i^{obs} y_i(0) & \sum S_i^{obs}
\end{array} \tag{2.5}$$

Table (2.5) differs from (2.3) only in that the principal stratum membership of control recipients who become infected is unknown. This problem is analogous to conducting a test in the presence of nuisance parameters. The following section will detail how an exact p-value for testing (2.1) can be obtained by conducting exact tests over a range of plausible values of the nuisance parameters and defining the exact p-value as a function of the largest p-value from this set of exact tests.

2.2.3 PSET Development

Now assume that $y_i(1)$ and $y_i(0)$ are any type of event and not necessarily binary. Let $\boldsymbol{Y}_1^{ai} \equiv \{Y_i^{obs} \colon Z_i = 1, i \in \mathscr{A}\mathscr{I}\}$ and $\boldsymbol{Y}_0^{ai} \equiv \{Y_i^{obs} \colon Z_i = 0, i \in \mathscr{A}\mathscr{I}\}$ and define $p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai})$ as the p-value for an exact randomization-based test of (2.1) assuming AI membership were known. For example, if no ties exist in $(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai})$ then the usual exact Wilcoxon rank sum test could be employed to compute $p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai})$.

As illustrated in Section 2.2.2, the set of AI stratum membership indicators for the infected control individuals can be viewed as unknown nuisance parameters. Analogous to Barnard's test, an exact test can be constructed by conducting a test for each possible AI subset of the infected control individuals and reporting the largest p-value (Barnard 1947). Unfortunately, this approach is overly conservative because it almost always fails to reject (2.1). Specifically, the set of possible AI subsets from infected control individuals includes subsets comprising only one individual. Provided there is at least one infected control individual with a post-infection outcome $y_i(0)$ that is not significantly different from $\{y_i(1)\colon s_i(1) = 1, Z_i = 1\}$, (2.1) will not be rejected. Furthermore, this approach ignores information available about the AI stratum. While the observed data do not identify which infected control individuals are in the AI stratum, the data do provide some information about the number of control individuals in the AI stratum. Thus an alternative approach is to view the number of control individuals in the AI stratum as the nuisance parameter and to obtain bounds for possible values of this nuisance parameter based on the observed data.

Let $M_0 \equiv \sum_{i \in \mathscr{A}\mathscr{I}} (1 - Z_i)$ and $M_1 \equiv \sum_{i \in \mathscr{A}\mathscr{I}} Z_i$ be the number of individuals in $\mathscr{A}\mathscr{I}$ assigned control and treatment such that $M_0 + M_1 = m$. Since the number of individuals in $\mathscr{A}\mathscr{I}$ does not depend on $\boldsymbol{Z}$, m is fixed, whereas M_0 and M_1 are random variables. Under A.3, $M_1 = \sum I[Z_i = S_i^{obs} = 1]$ is observable. In contrast, M_0 is not observable.

Suppose contrary to fact that M_0 is observed. Then an exact p-value could be obtained by performing an exact test for all possible selections of M_0 individuals from $\{i\colon Z_i = 0, S_i^{obs} = 1\}$ and taking the maximum of the resulting p-values. Define this p-value as

$$p^{ai}(M_0) \equiv \max\{p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0)\colon \boldsymbol{Y}_0 \in \Omega(M_0)\} \tag{2.6}$$

where $\Omega(M_0)$ equals the set of subsets of $\{Y_i^{obs}\colon Z_i = 0, S_i^{obs} = 1\}$ of size M_0.

Although M_0 is not observed, it is bounded above by $\sum I[Z_i = 0, S_i^{obs} = 1]$. Moreover, the observed M_1 provides information about m and thus M_0. Specifically, conditional on the total number assigned treatment $\sum Z_i$, under assumption A.2 an exact $100(1-\gamma)\%$ CI for m, say $C_\gamma \equiv [L_m, U_m]$, can be computed based on $\sum Z_i S_i^{obs} = \sum_{i \in \mathscr{A}\mathscr{I}} Z_i$ using standard results about simple random sampling (e.g., Thompson 2002). Then, following Berger and Boos (1994) , define

$$p_\gamma^{ai} \equiv \max\{p^{ai}(\tilde{m} - M_1)\colon \tilde{m} \in C_\gamma\} + \gamma. \tag{2.7}$$

The following proposition indicates that p_γ^{ai} is an exact p-value for testing (2.1).

Proposition 1: For any $\gamma \in [0,1]$, $\Pr[p_\gamma^{ai} \leq \alpha] \leq \alpha$ for all $\alpha \in [0,1]$ under H_0 (2.1).

The choice of γ should be made prior to looking at the data in a formal hypothesis testing scenario as the proposition holds only assuming γ is fixed. Section 2.3 presents simulation studies which provide empirical evidence suggesting $\gamma = \alpha/2$ may be recommended in certain settings. For tests where $p^{ai}(\tilde{m} - M_1)$ tend to decrease as $\tilde{m}$ increases, letting $U_m = \sum S_i^{obs}$ and computing a one-sided $(1-\gamma)\%$ CI for m to obtain L_m should result in a test with higher power compared to using a two-sided CI.

2.2.4 Computations

Calculating $p^{ai}(M_0)$ can be computationally intensive because it requires performing $\binom{\sum(1-Z_i)S_i^{obs}}{M_0}$ exact tests corresponding to $\Omega(M_0)$. In many settings, the computation requirements can be reduced by implicitly determining $\max\{p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0)\colon \boldsymbol{Y}_0 \in \Omega(M_0)\}$ without having to calculate $p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0)$ for each $\boldsymbol{Y}_0 \in \Omega(M_0)$. The proposition below shows how $p^{ai}(M_0)$ can be implicitly determined for a particular class of test statistics.

First consider the situation where AI membership is known such that the exact p-value $p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai})$ can be calculated. Let $j_1, j_2, ..., j_m$ denote the labels of individuals in AI such that $\mathscr{A}\mathscr{I} = \{j_1, ..., j_m\}$ and $\boldsymbol{Y}_1^{ai} \cup \boldsymbol{Y}_0^{ai} = \{Y_{j1}^{obs}, Y_{j2}^{obs}, ..., Y_{jm}^{obs}\}$. Let $\boldsymbol{y}^{ai}$ denote the vector $(Y_{j1}^{obs}, Y_{j2}^{obs}, ..., Y_{jm}^{obs})$, which is fixed under the null (2.1), and correspondingly let $\boldsymbol{Z}^{ai} = (Z_{j1}, Z_{j2}, ..., Z_{jm})$. Following Rosenbaum (2002a), let $t(\boldsymbol{Z}^{ai}, \boldsymbol{y}^{ai})$ denote the test statistic corresponding to $p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai})$. Assuming large values of $t(\boldsymbol{Z}^{ai}, \boldsymbol{y}^{ai})$ are considered evidence against the null (2.1), the exact one-sided p-value is calculated as

$$p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai}) = \frac{|\{\boldsymbol{z}^{ai} \in \Omega_m^{ai} \colon t(\boldsymbol{z}^{ai}, \boldsymbol{y}^{ai}) \geq t(\boldsymbol{Z}^{ai}, \boldsymbol{y}^{ai})\}|}{\binom{m}{M_1}} \tag{2.8}$$

where $|A|$ denotes the number of elements in the set A and Ω_m^{ai} denotes the set of possible treatment assignment vectors of length m with M_1 ones and M_0 zeros.

Define the test statistic $t(\boldsymbol{z}^{ai}, \boldsymbol{y}^{ai})$ to be *effect increasing* (Rosenbaum 2002a) if $t(\boldsymbol{z}^{ai}, \boldsymbol{y}^{ai1}) \geq t\ (\boldsymbol{z}^{ai}, \boldsymbol{y}^{ai2})$ for two possible response vectors $\boldsymbol{y}^{ai1}$ and $\boldsymbol{y}^{ai2}$ whenever $(y_j^{ai1} - y_j^{ai2})(2z_j^{ai} - 1) \geq 0$ for $j = 1, ..., m$ where in general u_j denotes the j^{th} element of vector $\boldsymbol{u}$. Informally, t is effect increasing if the value of the statistic increases when responses for the treated group are increased and the responses for the control group are decreased. Next define $t(\boldsymbol{z}^{ai}, \boldsymbol{y}^{ai})$ to be *invariant* if $t(\boldsymbol{z}^{ai}, \boldsymbol{y}^{ai}) = t(\boldsymbol{z}_{jk}^{ai}, \boldsymbol{y}_{jk}^{ai})$ for all j, k, where in general $\boldsymbol{u}_{jk}$ denotes the vector formed by interchanging the j^{th} and k^{th} elements of $\boldsymbol{u}$. In words, $t(\boldsymbol{z}^{ai}, \boldsymbol{y}^{ai})$ is invariant if permuting the labels of individuals does not change the value of the statistic. Many common statistics such as Fisher's exact test statistic and the Wilcoxon rank sum statistic are invariant and effect increasing (Rosenbaum 2002a). According to the proposition below, for invariant and effect increasing statistics $\max\{p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0) \colon \boldsymbol{Y}_0 \in \Omega(M_0)\}$ can be determined by calculating a single p-value.

Proposition 2: If $t(\boldsymbol{z}^{ai}, \boldsymbol{y}^{ai})$ is invariant as well as effect increasing, then $p^{ai}(M_0) = p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai}[1{:}M_0])$ where $\boldsymbol{Y}_0^{ai}[1{:}M_0]$ is the set of M_0 largest values of $\{Y_i^{obs} \colon Z_i = 0, S_i^{obs} = 1\}$.

2.2.5 Positive Effect

The choice of test statistic used for conducting the PSET of (2.1) will be dictated by the type of post-infection outcome (e.g., whether y is binary, ordinal, continuous, etc) and alternative hypothesis of interest. One possible alternative hypothesis is that treatment has a *positive effect* (Rosenbaum 2002a; see also Lehmann 1998) in the AI stratum, i.e.,

$$H_A : y_i(1) \geq y_i(0) \text{ for all } i \in \mathscr{A}\mathscr{I} \tag{2.9}$$

where the inequality in (2.9) is strict for at least one $i \in \mathscr{A}\mathscr{I}$. In words, treatment has a positive effect if it increases y for at least one individual and does not decrease y for any individual in AI. The additivity model $y_i(1) - y_i(0) = \delta$ for all $i \in \mathscr{A}\mathscr{I}$ and constant $\delta > 0$ is a special case of (2.9). If the test statistic t is effect increasing, the following proposition shows the PSET is an unbiased test of (2.1) against (2.9), i.e., the PSET is at least as likely to reject H_0 at the α significance level when H_A holds as compared to when H_0 holds.

Proposition 3: If $t(\boldsymbol{z}^{ai}, \boldsymbol{y}^{ai})$ is effect increasing, then $\Pr[p_\gamma^{ai} < \alpha | H_A] \geq \Pr[p_\gamma^{ai} < \alpha | H_0]$.

2.2.6 Plug-in P-value Alternative

An alternative testing approach that has been proposed for addressing the presence of nuisance parameters entails conditioning on estimates of the unknown parameters; the resulting p-value is sometimes referred to as the "plug-in p-value" (Bayarri and Berger 2000). For example, a plug-in p-value approach has been advocated as an alternative to Fisher's exact test (Storer and Kim 1990). Plug-in p-values are computationally straight forward and asymptotically exact under certain assumptions about the form of the test statistic (Robins, van der Vaart, and Ventura 2000). Considering m to be a nuisance parameter when testing (2.1), a plug-in type p-value can be defined by conditioning on an unbiased estimate $\hat{M} = nM_1/\sum Z_i$ (Thompson 2002) of m:

$$p_{plug}^{ai} \equiv p^{ai}(\hat{M} - M_1) \equiv \max\{p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0) \colon \boldsymbol{Y}_0 \in \Omega(\hat{M} - M_1)\}$$

Unfortunately such an approach does not take into account uncertainty about m and therefore p^{ai}_{plug} is not guaranteed to be exact. For example, consider the following population of 8 individuals with potential infection and post-infection outcomes $s_i(z)$ and $y_i(z)$ where $y_i(z) = *$ indicates the post-infection outcome is not defined if $s_i(z) = 0$. Suppose by design $\Pr[\sum Z_i = 4] = 1$.

i	$s_i(0)$	$s_i(1)$	$y_i(0)$	$y_i(1)$	i	$s_i(0)$	$s_i(1)$	$y_i(0)$	$y_i(1)$
1	1	1	8	8	5	1	1	4	4
2	1	1	7	7	6	1	0	3	$*$
3	1	1	6	6	7	1	0	2	$*$
4	1	1	5	5	8	1	0	1	$*$

Suppose a one-sided Wilcoxon rank sum test is used to test (2.1), where $t(\boldsymbol{Z}^{ai}, \boldsymbol{Y}^{ai}) = \sum_{i \in \mathscr{A}\mathscr{I}} R^{ai}_i Z^{ai}_i$ with R^{ai}_i denoting the rank of Y^{obs}_i among $\{Y^{obs}_i : i \in \mathscr{A}\mathscr{I}\}$ and the p-value is computed by (2.8). Then for $\alpha = 0.05$, $\Pr[p^{ai}_{plug} \leq \alpha] = 0.07 > \alpha$ because $p^{ai}_{plug} \leq \alpha$ for 5 of the $\binom{8}{4} = 70$ possible treatment assignment permutations.

2.3 Simulation Study

A primary objective of preventive HIV vaccine trials is to assess whether vaccination has an effect on viral load in individuals who become infected. Gilbert et al. (2003) and Hudgens et al. (2003) developed bootstrap tests of the null hypothesis that vaccination has no effect on viral load in the AI principal stratum. To evaluate the operating characteristics (type I error and power) of these proposed tests, they conducted simulation studies of HIV vaccine trials with 2000 HIV negative individuals randomized 1:1 to either vaccine or placebo under various assumptions regarding the rates of infection in the vaccine and placebo arms. In settings where the expected number of observed infections was moderate or large, the proposed tests preserved the nominal type I error probability. However, in settings where the expected number of observed infections was small (45 infections in the placebo arm, 31.5 in the vaccine arm), the bootstrap tests demonstrated inflated type I error. Therefore we conducted a simulation study under identical assumptions to assess how the PSET performs in comparison.

Let $z = 0$ for placebo individuals and $z = 1$ for vaccinated individuals. For assignment z, $s_i(z) = 1$ if an individual is infected and $y_i(z)$ is the log-transformed viral load when $s_i(z) = 1$. Thus, the null (2.1) corresponds to the vaccine having no effect on viral load in the AI principal stratum. It is of interest to test the null against the one-sided alternative that viral load is higher when vaccinated, given concerns that an HIV vaccine may actually increase viral load in breakthrough infections (Hudgens et al. 2003).

The following steps were performed for each trial simulation. First, $s_i(0)$ was set equal to 1 for $i = 1, ..., 90$ and $s_i(0) = 0$ for $i = 91, ..., 2000$. For $i = 1, ..., 90$, $y_i(0)$ was randomly generated from a normal distribution with mean 4.5 and standard deviation 0.6. Next, selection bias was simulated by setting $s_i(1) = 1$ for the 63 individuals with the largest values of $y_i(0)$ and $s_i(1) = 0$ otherwise. Thus vaccination caused a 30% reduction in the number of infections, with only individuals who would have low viral load if not vaccinated being protected from infection by vaccine. Vaccine effect on viral load was simulated by letting $y_i(1) = y_i(0) + \delta$ for $i \in \mathscr{A}\mathscr{I}$. Finally, 1000 individuals were randomly assigned placebo, the remaining 1000 assigned vaccine and the observed outcomes were selected from $(s_i(1), s_i(0), y_i(1), y_i(0))$ accordingly.

For each simulated dataset, the PSET, nonparametric mean bootstrap test of Hudgens et al. (2003) and plug-in p-value from Section 2.2.6 were calculated. For the PSET, we used a one-sided $100(1 - \gamma)\%$ CI of m to obtain L_m and a Wilcoxon rank sum test to compute the conditional p-value $p^{ai}(M_0)$ in (2.6). Simulations using a two-sided $100(1 - \gamma)\%$ CI to obtain L_m and U_m resulted in reduced power compared to the one-sided approach (results not shown). Table 2.1 gives the empirical type I error and power of the PSET for various values of γ and significance level α, based on 10,000 simulations per combination of γ and α. As expected, in all scenarios the empirical type I error of the PSET was less than α. In contrast, for $\alpha = 0.05$ the empirical type I errors of the bootstrap test and plug-in p-value were 0.11 and 0.19 respectively, i.e., over twice the nominal level. For $\alpha = 0.05$ the power of the PSET was highest for $\gamma = 0.030$. For $\alpha = 0.10$, $\gamma = 0.05$ yielded the greatest power. Thus, choosing $\gamma = \alpha/2$ may be recommended in this setting.

Additional simulation studies were conducted to compare the power of the PSET to the bootstrap test when the AI stratum sample size was increased. Specifically, the simulations studies described above were repeated twice, but with 90 and 135 expected observed infections in the placebo arm and vaccination causing a 30% reduction in the number of infections in both scenarios. For 90 expected placebo arm infections, the empirical type I error and power for $\delta = 1/3$ and $2/3$ were 0.004, 0.426, and 0.990 and 0.057, 0.742, and 0.999 for the PSET and bootstrap tests respectively. For 135 expected placebo arm infections, the empirical type I error and power for $\delta = 1/3$ and $2/3$ were 0.005, 0.664, and 0.999 and 0.039, 0.908, and 1.00 for the PSET and bootstrap tests respectively. Thus for larger AI stratum the bootstrap test controlled the type I error and had greater power than the PSET for small δ.

As discussed in Section 2.2.1, the veracity of A.3 may be of concern in some settings. To assess the robustness of the PSET when A.3 is violated, additional simulations were conducted where some individuals infected under vaccine belong to the harmed stratum. Data were simulated as described in the original scenario (where 90 individuals were infected if not vaccinated), except that 6 (10%) of the 63 individuals infected if vaccinated were from the harmed stratum. In particular, for each trial, $s_i(0)$ and $y_i(0)$ were generated as described above. Selection bias was simulated by setting $s_i(1) = 1$ for a random selection of 57 of the 63 individuals with the largest values of $y_i(0)$. Vaccine effect on viral load in the AI stratum was simulated by letting $y_i(1) = y_i(0) + \delta$ for these individuals. Harmed individuals were then simulated by setting $s_i(1) = 1$ for $i = 91, ..., 96$ and generating $y_i(1)$ from the same normal distribution used to generate $y_i(0)$. All subsequent steps of the simulation were the same as before. The PSET empirical type I error and power for $\delta = 1/3$ and $2/3$ were 0.003, 0.108, and 0.649. That is, the PSET type I error was less than the nominal α despite A.3 not holding, and the PSET power was slightly diminished relative to simulations where A.3 holds. Similar results were obtained when 20% and 30% of individuals infected when vaccinated were members of the harmed stratum.

2.4 Application

2.4.1 Zambia Exclusive Breastfeeding (ZEB) Study

The ZEB study was a randomized study to evaluate whether abrupt weaning at 4 months as compared to continued breastfeeding increases survival of children of HIV infected mothers (Kuhn et al. 2008). The trial was conducted in 958 HIV-infected women and their infants in Lusaka, Zambia with 481 children randomized to the intervention and 477 randomized to standard practice of continued breastfeeding. Randomization occurred at one month postpartum to allow for sufficient preparation time for weaning at 4 months. Kuhn et al. present an analysis of the effect of weaning on survival through 24 months based on a log-rank test comparing survival between randomization groups for the subset of infants who became HIV-infected prior to 4 months but survived more than 4 months. A total of 62 individuals in the intervention arm were HIV-infected and alive at 4 months, 39 (63%) who died prior to 24 months. Likewise, 70 in the standard practice arm were HIV-infected and alive at 4 months, 32 (46%) who died prior to 24 months. The log-rank p-value was 0.007, leading Kuhn et al. to conclude that there is evidence of a harmful effect of weaning on survival among HIV positive infants alive at 4 months.

Because the reported analysis conditions on infection and survival status at 4 months, the results do not necessarily have a causal interpretation and could be due to selection bias. Specifically, any differences between the study arms during months 1-4 could affect infection and survival status at 4 months. For instance, at month 2 women in the intervention group were counseled on techniques for weaning and given a three month supply of infant formula and fortified weaning cereal. This may have caused women in the intervention group to wean earlier than had they been randomized to the control group, in turn perhaps impacting HIV acquisition. In fact, more women in the intervention arm weaned by 4 months (37 versus 18) and, possibly because of this, fewer infants in the intervention arm became HIV positive at or before 4 months (71 versus 81).

The principal stratum of interest is the AI stratum, defined as all individuals who would be HIV-infected and alive at 4 months regardless of randomization assignment. The PSET was used to test the null hypothesis of no effect of the intervention on death in the AI stratum. To compute (2.6), a one-sided Fisher's exact test was used where each individual was classified as having died or not. An exact log-rank test might be preferable for calculating the conditional p-values, however the individual death and censoring times were not reported by Kuhn et al. For $\gamma = 0.025$, the PSET resulted in $p_{\gamma}^{ai} = 0.98$ suggesting no evidence of a harmful effect of weaning on survival for the AI stratum. The one-sided CI for m, i.e., the total number of infants in the AI stratum, is 104 to 132. Figure 1 plots the p-values from Fisher's exact test conditional on $m = \tilde{m}$ for each $\tilde{m} \in C_{\gamma} = [104, 132]$. While a Fisher's exact test using all available data and ignoring the potential for selection bias is less than $\alpha = 0.05$ (p-value$= 0.0355$), 27 of the 29 conditional p-values are greater than 0.05. Additionally, even testing the hypothesis using the plug-in p-value (i.e. $m = \hat{M}$) results in $p^{ai}_{plug} = 0.1611 > \alpha$. In order reject to (2.1) based on the PSET for $\alpha = 0.05$ and $\gamma = 0.025$, 58 of 62 individuals would have had to die in the intervention arm compared to the 32 of 70 in the standard practice arm ($p^{ai}_{\gamma} = 0.0375$).

2.4.2 Breastfeeding, Antiretroviral and Nutrition (BAN) Study

The BAN study was a randomized trial of infants of HIV infected mothers to evaluate whether daily administration of nevirapine (NVP) to the infant through 28 weeks decreased risk of HIV transmission via breastfeeding to infants when compared to a control arm receiving no antiretroviral therapy (Chasela et al. 2010). A total of 668 mothers and their infants were randomized to control while 852 were randomized to infants receiving NVP. Fewer mother-infant pairs were randomized to control because the data and safety monitoring board (DSMB) stopped enrollment in this arm early. The effect of NVP on infection status through 28 weeks was assessed using a log-rank test that compared infection between treatment groups for all infants who were not infected at two weeks. Of the 632 infants not infected at two weeks in the control arm, 32 (5.1%) were infected by 28 weeks. Likewise, of the 815 infants not infected at two weeks in the NVP arm, 12 (1.5%) were infected by 28 weeks. The log-rank p-value was < 0.001, suggesting NVP prevents breast milk transmission of HIV. However, these results do not have an immediate causal interpretation and could be subject to selection bias because the analysis conditions on a post-randomization outcome: HIV infection status at two weeks. To guard against this Chasela et al. also reported results from an ITT analysis of all infants randomized, including those infected before two weeks. However, a primary objective of BAN was to investigate the effect of NVP to prevent breast milk transmission. Thus the investigators were primarily interested only in infections occurring after two weeks, as infants who were HIV positive by two weeks may have been infected *in utero* or during birth. Because daily NVP from birth could potentially effect infection status at two weeks, the groups of infants not infected at two weeks in each study arm may not be comparable.

The principal stratum of interest is the never-infected (NI) stratum, defined as individuals who would be HIV uninfected at two weeks regardless of randomization assignment. The PSET can be used to test the null of no treatment effect on infection by 28 weeks in the NI stratum. In contrast to the AI stratum, membership in the NI stratum is known for control individuals not infected at two weeks since by A.3 infants not infected at two weeks when assigned control would also not be infected at two weeks when assigned NVP. On the other hand, membership in the NI stratum is unknown for infants assigned NVP not infected at two weeks. Thus the PSET can be conducted in the NI stratum with the roles of the treated and control individuals reversed relative to conducting the PSET in the AI stratum. For given γ, denote the PSET p-value for the test of no principal stratum direct effect in the NI stratum by p_{γ}^{ni}, which is computed analogous to (2.7). Because enrollment was stopped early in one arm, we compute p_{γ}^{ni} using only data available prior to the DSMB decision. These data include all the control arm infants described above but only 670 of the infants in the NVP arm, 639 of who were not infected at two weeks. Of these 639 infants, 10 (1.6%) were infected by 28 weeks. Because the BAN study was a multi-arm trial, the analysis plan stipulated that tests between the NVP and control arms be conducted at the $\alpha = 0.025$ significance level. Letting $\gamma = 0.0125$ and using a one-sided Fisher's exact test, the PSET resulted in $p_{\gamma}^{ai} = 0.0131$ indicating a benefit of NVP among infants who were immune to infection at two weeks. Using an exact log-rank test with Monte Carlo sampling (Mehta and Patel 2007) yielded a similar result with $p_{\gamma}^{ai} = 0.0127$.

2.4.3 Sensitivity Analysis

As discussed in Sections 2.2 and 2.3, A.3 is a key assumption of the PSET. In the BAN study, comparison of infection rates at two weeks provides no evidence that monotonicity is violated, with the proportion infected at two weeks slightly lower in the NVP arm. Additionally, multiple other studies (Mofenson 2009) have shown daily administration of NVP protects infants from breast milk transmission of HIV and to date there is no evidence suggesting NVP may cause HIV transmission. Nonetheless, when A.3 may be of concern, a sensitivity analysis can be conducted.

To illustrate one possible sensitivity analysis, suppose there was concern about A.3 in the BAN study. Without A.3 the 632 control arm infants uninfected at two weeks are not all necessarily members of the NI stratum. Rather, some of these infants may belong to the harmed stratum, i.e., they may have been infected by two weeks if randomized to NVP. Suppose h of the 632 are from the harmed stratum. If these h infants could be identified, the PSET could be conducted based on the remaining $632 - h$ control arm infants uninfected at two weeks. Because the h infants cannot be identified without additional assumptions, the sensitivity analysis entails considering different scenarios. Specifically, divide the h infants into h_1 from the 600 control arm infants not infected by 28 weeks and $h_2 = h - h_1$ from the 32 control arm infants infected by 28 weeks. Then conduct the PSET for different combinations of (h_1, h_2). For the BAN study the PSET p-value (based on a one-sided Fisher's exact test) is more sensitive to changes in h_2 than h_1. For example, for γ=0.0125, $h_1 = 8$ and $h_2 = 0$ yields $p_\gamma^{ni} = 0.013$, while $h_1 = 0$ and $h_2 = 8$ yields $p_\gamma^{ni} = 0.027$. Holding $h_1 = 0$ fixed, $p_\gamma^{ni} < 0.025$ for $h_2 = 0, 1, 2, \ldots, 7$ and $p_\gamma^{ni} > 0.025$ for $h_2 > 7$. In words, NVP was beneficial in the NI stratum at the 0.025 significance level provided no more than 7 of the 32 control arm infants infected by week 28 were from the harmed stratum.

2.5 Comparisons with ITT Approaches

Principal stratification provides a method for dealing with possible selection bias induced by conditioning on an intermediate post-randomization outcome. Alternatively, an ITT based approach can be employed. The ITT principle generally refers to analyzing all individuals according to randomization assignment. ITT has become the gold standard in clinical trials as it ensures the validity of testing the null hypothesis of no treatment effect (assuming perfect compliance) and helps minimize bias such that observed differences in outcomes between the groups can be attributed to the treatment under study. The ITT approach does however have some potential drawbacks. For instance, in the infectious disease setting, unlike principal stratification the ITT approach does not clearly differentiate treatment effects on infection and post-infection outcomes. Also, it is conceptually challenging to define post-infection outcomes for non-infected individuals. Similarly, quality of life outcomes may be considered undefined in individuals not alive (Rubin 2006).

To obviate the latter problem, Chang, Guess and Heyse (1994) proposed an ITT-based burden of illness (BOI) test for assessing treatment effect on disease severity by assigning burden of illness scores to each incident infection, with individuals who escape infection receiving a score of zero. Denote the observed disease severity scores by W_i^{obs} where $W_i^{obs} = Y_i^{obs}$ if $S_i^{obs} = 1$ and $W_i^{obs} = 0$ if $S_i^{obs} = 0$. Then define $\boldsymbol{W}_1^{ITT} \equiv \{W_i^{obs} \colon Z_i = 1\}$, $\boldsymbol{W}_0^{ITT} \equiv \{W_i^{obs} \colon Z_i = 0\}$ and $p(\boldsymbol{W}_1^{ITT}, \boldsymbol{W}_0^{ITT})$ as the p-value for an exact randomization-based test comparing $\boldsymbol{W}_1^{ITT}$ and $\boldsymbol{W}_0^{ITT}$. As opposed to (2.1), the null hypothesis of the randomization BOI test is $H_0 \colon w_i(1) = w_i(0)$ for all $i \in \{1, \ldots, n\}$. Assuming $Y_i^{obs} > 0$ whenever $S_i^{obs} = 1$, it follows that the null hypothesis of the BOI test is equivalent to testing the composite hypothesis:

$$H_0 \colon s_i(1) = s_i(0) \text{ and } y_i(1) = y_i(0) \text{ for all } i \in \{1, \ldots, n\}, \tag{2.10}$$

Because the BOI test may have poor power when infections are rare, Follmann, Fay, and Proschan (2009) proposed the chop-lump test as an alternative ITT test of (2.10). For this method, a test statistic is calculated based on a subset of the data obtained by removing (or "chopping") $\min\{\sum(1-Z_i)(1-S_i^{obs}), \sum Z_i(1-S_i^{obs})\}$ observations where $W_i^{obs} = 0$ from each randomization group such that the remaining data from at least one of the groups has no observations where $W_i^{obs} = 0$. The test statistic (e.g., difference in means between groups) is computed based on this subset. Randomization-based p-values are obtained in the usual fashion, i.e., by considering all possible randomization assignments of the n individuals and computing the test statistic for each possibility.

For the simulation scenario of Section 2.3, the power was < 0.05 for the BOI for all δ and 0.181, 0.370 and 0.483 for the chop-lump for $\delta = 1/3$, 2/3 and 1 respectively. For these tests, let $W_i^{obs} = 0$ for uninfected individuals ($S_i^{obs} = 0$) and $W_i^{obs} = Y_i^{obs}$ for infected individuals ($S_i^{obs} = 1$). Then for both tests the Wilcoxon rank sum test statistic was used to compare W_i^{obs} and one-sided p-values were computed corresponding to the vaccine causing higher viral load. The lack of power for the ITT tests in this setting is partially due to the opposite direction of vaccine effects on infection and viral load, i.e., for $\delta > 0$ the vaccine is protecting some individuals but causing a higher viral load in the AI stratum.

To compare the BOI, chop-lump and PSET when the vaccine only effects the post-infection outcome, additional simulations were conducted similar to that described in Section 2.3 except we let $s_i(1) = s_i(0)$ for all i such that the expected number of observed infections was 45 for each arm. For $\alpha = 0.05$, the empirical type I error and power for $\delta = 1/3$, 2/3 and 1 were 0.049, 0.061, 0.062 and 0.063 for the BOI test; 0.048, 0.365, 0.747 and 0.898 for the chop-lump; and 0.002, 0.100, 0.644 and 0.978 for the PSET (with $\gamma = 0.025$). That is, the PSET is markedly more powerful test than the BOI approach for all δ and comparable in power to the chop-lump for larger values of δ. Mehrotra et al. (2006) presented similar findings when comparing large-sample frequentist based principal stratification tests with a BOI test.

The PSET is unambiguously better for testing principal stratum direct effects than the BOI, chop-lump and other ITT-based tests in settings where treatment z has an effect on infection s but not on the post-infection outcome y. For then the ITT-based tests may reject (2.10) even though the null hypothesis of interest (2.1) is true (i.e., treatment has no effect on the post-infection outcome y). For example, consider the scenario described in Section 2.3 where vaccine causes a 30% reduction in the number of infections, there are 45 expected infections in the placebo arm, and $\delta = 0$ (i.e., (1) is true). Suppose the alternative hypothesis of interest is that the vaccine reduces viral load. In this scenario, one-sided BOI and chop-lump tests reject (2.10) at the $\alpha = 0.05$ level of significance for over 50% of the simulated data sets. In other words, the BOI and chop-lump tests do not have the correct size for testing (2.1) when there is a treatment effect on s.

2.6 Adjusting for Covariates

Covariate adjustment is often used in analysis of randomized experiments to account for chance imbalances that may exist between study arms, thereby allowing for more precise inference. Following Rosenbaum (2002b), in this section we consider extending the PSET to incorporate baseline (i.e., pre-randomization) covariates. This approach entails first regressing the outcomes of interest on covariates and then conducting an exact test on the residuals. Ideally, the residuals obtained from the regression model are less variable than the original outcomes of interest, resulting in increased power of the PSET. The appeal of this approach is no distributional assumptions about the response nor the selected regression model are required. As before, randomization inference is employed such that the potential outcomes as well as the covariates are assumed to be fixed features of the finite population and not affected by treatment assignment.

2.6.1 PSET Development Adjusting for Covariates

Let x_i represent the baseline covariate value for individual i. Denote the function that creates residuals from $y_i(z)$ and x_i by g such that $g(y_i(z), x_i) = e_i(z)$ where $e_i(z)$ is the residual for the i^{th} individual when assigned treatment z. Under the null hypothesis (2.1), $y_i(1) = y_i(0)$ and therefore $e_i(1) = e_i(0)$ for all $i \in \mathscr{A}\mathscr{I}$. Therefore, a test of (2.1) can be constructed using the residuals.

The covariate-adjusted PSET is constructed in a similar fashion to the PSET from Section 2.2.3 except Y_i^{obs} is replaced with the observed residuals, $E_i^{obs} = Z_i e_i(1) + (1 - Z_i)e_i(0)$. The exact randomization-based test used to obtain $p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai})$ need not be the same test used on the residuals as the choice of tests depends on the characteristics of $y_i(z)$, x_i and $g(,)$. For example, consider the logistic regression setting where $y_i(z)$ is binary, x_i has no ties and $g(y, x) = y - \exp(\hat{\beta}_0 + \hat{\beta}_1 x)/\{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)\}$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by maximum likelihood estimation. Resulting values for $e_i(z)$ will typically have no ties. Accordingly, Fisher's exact test could be used to obtain $p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai})$ while a Wilcoxon rank sum test could be used on the residuals.

2.6.2 Simulations

To assess the power of the covariate adjusted PSET, the simulation scenario described in Section 2.3 was updated to include baseline CD4 count, a measure of immune function. Baseline CD4 count x_i was assumed to be normally distributed with mean 850 and standard deviation 300. For all individuals who would be infected if assigned control, CD4 count x_i and post-infection log viral load when receiving control $y_i(0)$ were simulated under various levels of correlation ($\rho = 0.0, 0.1, ..., 0.9$). Residuals were obtained using $g(Y_i^{obs}, x_i) = Y_i^{obs} - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are solutions to the normal equations for the linear regression model of Y_i^{obs} on x_i for all i with $S_i^{obs} = 1$. A one-sided $100(1-\gamma)\%$ CI was computed to obtain L_m and a one-sided Wilcoxon rank sum test of the residuals was used to obtain the conditional p-values $p^{ai}(M_0)$ in (2.6).

Results of the simulations for $\alpha = 0.05$ and $\gamma = 0.025$ are displayed in Table 2.2. Comparing to Table 2.1, adjusting for x_i increased the power of the PSET when $\rho > .5$, markedly so for the larger value of δ. For weak levels of correlation, an increase in power was not observed; for ρ=0 there was a slight loss of power when adjusting for the covariate. While the covariate-adjusted PSET is not guaranteed to increase power compared to the unadjusted PSET, it is still guaranteed to be exact.

2.7 Confidence Intervals

The PSET can be used to form an exact CI for a principal stratum direct effect. Suppose treatment effect in the AI stratum is additive such that $y_i(1) - y_i(0) = \delta_0$ for all $i \in \mathscr{A}\mathscr{I}$. Then a CI for the principal stratum direct effect δ_0 can be obtained by inverting a generalized version of the PSET developed in Section 2.2.3. The CI is constructed by conducting the generalized PSET for all possible values of δ_0 and forming the set of values where the test is not rejected (Lehmann 1959, Rosenbaum 2002a).

The first step is to adapt the PSET to allow for testing a more general null hypothesis. For some constant δ not necessarily equal to zero, consider testing:

$$H_0\colon y_i(1) - y_i(0) = \delta \text{ for all } i \in \mathscr{A}\mathscr{I}, \tag{2.11}$$

Note under (2.11) that $(Y_i^{obs} - \delta)Z_i + Y_i^{obs}(1-Z_i) = (y_i(1)-\delta)Z_i + y_i(0)(1-Z_i) = y_i(0)$ is constant. Thus the PSET of (2.11) can be constructed as in Section 2.2.3 except Y_i^{obs} is replaced with $Y_i^{obs} - \delta$ for individuals where $Z_i = 1$ (Rosenbaum 2002a). A one-sided

$100(1-\alpha)\%$ CI for the true δ_0 is formed by the set of all δ where a one-sided test of (2.11) is not rejected.

More specifically, let $p_\delta(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai}) = p(\boldsymbol{Y}_1^{ai} - \delta, \boldsymbol{Y}_0^{ai})$ denote the one-sided p-value from an exact randomization-based test of (2.11) using $Y_i^{obs} - \delta$ for individuals where $Z_i = 1$, and let $p_\delta^{ai}(M_0) = \max\{p_\delta(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0) \colon \boldsymbol{Y}_0 \in \Omega(M_0)\}$ and $p_{\gamma,\delta}^{ai} = \max\{p_\delta^{ai}(\tilde{m} - M_1) \colon \tilde{m} \in C_\gamma\} + \gamma$. Let δ_{min} and δ_{max} denote the lower and upper limits of the range of possible values for δ_0. Following Mehta and Patel (2007), define the lower bound of the $100(1-\alpha)\%$ one-sided CI of δ_0 by $\Delta_\gamma^\alpha = \sup\{\delta \colon p_{\gamma,\tilde{\delta}}^{ai} \leq \alpha \text{ for all } \tilde{\delta} < \delta\}$. If there does not exist δ such that $p_{\gamma,\tilde{\delta}}^{ai} \leq \alpha$ for all $\tilde{\delta} < \delta$ then Δ_γ^α is set to δ_{min}. The upper bound of the CI is set to δ_{max}. According to the following proposition, the interval $[\Delta_\gamma^\alpha, \delta_{max}]$ is an exact one-sided $100(1-\alpha)\%$ CI of the principal stratum direct effect δ_0 because the probability of covering the true value of δ_0 is at least $(1-\alpha)$.

Proposition 4: For $\gamma \in [0,1]$ and $\alpha \in [0,1]$, $\Pr[\delta_0 \in [\Delta_\gamma^\alpha, \delta_{max}]] \geq 1 - \alpha$.

2.8 Discussion

2.8.1 Summary

In randomized studies, comparisons between randomized groups that condition on intermediate post-randomization outcomes generally do not have a causal interpretation. An alternate approach entails comparisons within principal strata defined by the intermediate potential outcomes that would be observed under each randomization assignment. In this paper, we develop exact, randomization-based methods for inference about the treatment effect within a principal stratum. The three key assumptions for the PSET are SUTVA (A.1), random treatment assignment (A.2), and monotonicity (A.3); no assumptions are required about random sampling or that particular parametric distributions hold. Simulation studies indicate the PSET can be as or more powerful than ITT approaches when treatment has no impact on the intermediate post-randomization outcome. The power of the PSET can be increased by adjusting for baseline covariates and exact CIs for the principal stratum direct effect can be obtained by inverting the PSET.

2.8.2 Other applications

This work is motivated by infectious disease prevention studies where the treatment is some preventive measure (such as a vaccine), the intermediate variables s is infection, and the outcome of interest y is a post-infection endpoint (such as death). Two other settings where principal stratification methods are typically employed include truncation by death (Zhang and Rubin 2003) and non-compliance (Angrist et al. 1996). For the truncation by death problem, the PSET can readily be employed. In this setting, the intermediate variable s is death (0 for alive, 1 for death), the outcome of interest y is some measurement such as quality of life that is only well defined when individuals are alive, and the principal stratum of interest is the set of individuals who would be alive under either treatment assignment. The stratum of interest $\{i\colon s_i(0) = s_i(1) = 0\}$ is directly analogous to the never infected principal stratum discussed in Section 2.4.2. Here the monotonicity assumption A.3 indicates no individuals would die due to treatment.

In the non-compliance setting, the intermediate variable s is compliance to randomization assignment z and the goal is to make inference about the outcome of interest y for those individuals who would comply under either randomization assignment. Following Angrist et al. (1996), let $s_i(z) = 1$ if individual i actually receives treatment and $s_i(z) = 0$ if individual i receives control when assigned z. If individual i always complies with their randomization assignment then $s_i(z) = z$. Thus the principal stratum of interest, the compliers, is $\{i\colon s_i(0) = 0, s_i(1) = 1\}$. Typically a form of monotonicity is assumed such that there are no individuals who always defy their randomization assignment, i.e., $\{i\colon s_i(0) = 1, s_i(1) = 0\}$ is empty. Additionally, often it is assumed that randomization assignment has no effect on individuals who ignore treatment assignment, i.e., $y_i(0) = y_i(1)$ if $s_i(0) = s_i(1)$. Under this exclusion restriction and assuming monotonicity, the principal stratum direct effect null

$$H_0\colon y_i(0) = y_i(1) \text{ for all } i \in \{j\colon s_j(0) = 0, s_j(1) = 1\} \tag{2.12}$$

will be true if and only if the ITT null

$$H_0\colon y_i(0) = y_i(1) \text{ for } i \in \{1, \ldots, n\}, \tag{2.13}$$

is true, so that the usual randomization tests of (2.13) can be used to test (2.12), as suggested by Rosenbaum (1996).

If one is not willing to make the exclusion restriction assumption above (e.g., see Jo 2002), then (2.12) and (2.13) are not equivalent and thus the usual randomization (ITT) tests will generally not have the correct size for testing (2.12), since effects of randomization on y in non-compliers can lead to rejection of (2.13) even though (2.12) is true. In certain settings treatment may not be available to individuals randomized to control (e.g., see Ten Have et al. 2003, Little et al. 2009), so that $s_i(0) = 0$ always. In this setting and assuming monotonicity, the PSET applies; individuals with $Z_i = S_i^{obs} = 1$ must be compliers (just as infected treated individuals must be in the AI stratum) and individuals with $Z_i = S_i^{obs} = 0$ are a mixture of compliers and never takers (just as infected controls are a mixture of individuals from the protected and AI strata).

2.8.3 Future Directions

We close by mentioning five possible avenues of future research. (i) The development of the PSET as an exact test of (2.1) arose from viewing individuals' principal stratum memberships as partially unknown nuisance parameters and then employing the approach developed by Berger and Boos (1994). Other approaches to testing in the presence of nuisance parameters might be adapted to the principal stratification setting, giving rise to exact tests of (2.1) different from the PSET in this paper. (ii) Extensions to observational settings where assumption A.2 does not necessarily hold could be considered. For examples of permutation inference in observational studies see Rosenbaum (1984, 2002). (iii) A method for obtaining an exact CI of the principal stratum direct effect by inverting the PSET was presented in Section 2.7. This method assumes the treatment effect is additive (i.e., constant) within the principal stratum of interest. Future research could entail relaxing this assumption. Similar to Rosenbaum (2001), one approach might entail extending the PSET to the more general null hypothesis $H_0\colon y_i(1) - y_i(0) = \delta_{0i}$ for $i \in \mathscr{A}\mathscr{I}$ where the individual treatment effects δ_{0i} may differ between individuals; this extended PSET could then, perhaps, be inverted to obtain a CI for the average principal stratum direct effect (VanderWeele 2008). (iv) As discussed in Sections 2.2, 2.3 and 2.4, the monotonicity assumption may be dubious in certain settings. Additional investigation is needed into weakening assumption A.3. (v) Covariate adjustment was considered in Section 2.6 as a method for possibly increasing the power of the PSET. Alternatively, baseline covariates could be used to predict principal stratum membership (e.g., see Roy et al. 2008) and perhaps this information could somehow be incorporated within the randomization-based inference framework.

2.9 Proof of Propositions

Proof of Proposition 1

Proof: Suppose H_0 (2.1) is true. Fix $\gamma \in [0,1]$ and $\alpha \in [0,1]$. If $\gamma > \alpha$, then $p^{ai}_{\gamma} = \max\{p^{ai}(\tilde{m} - M_1)\colon \tilde{m} \in C_{\gamma}\} + \gamma > \alpha$. Therefore $\Pr[p^{ai}_{\gamma} \leq \alpha] = 0 \leq \alpha$. If $\gamma \leq \alpha$, then

$$\begin{aligned}
\Pr[p^{ai}_{\gamma} \leq \alpha] &= \Pr[p^{ai}_{\gamma} \leq \alpha, m \in C_{\gamma}] + \Pr[p^{ai}_{\gamma} \leq \alpha, m \in \overline{C}_{\gamma}] \\
&\leq \Pr[\max\{p^{ai}(\tilde{m} - M_1)\colon \tilde{m} \in C_{\gamma}\} + \gamma \leq \alpha, m \in C_{\gamma}] + \Pr[m \in \overline{C}_{\gamma}] \\
&\leq \Pr[p^{ai}(m - M_1) + \gamma \leq \alpha, m \in C_{\gamma}] + \gamma \\
&\leq \Pr[p^{ai}(m - M_1) \leq \alpha - \gamma] + \gamma \\
&\leq \alpha - \gamma + \gamma = \alpha
\end{aligned}$$

where the 2^{nd} inequality holds since $\max\{p^{ai}(\tilde{m} - M_1)\colon \tilde{m} \in C_{\gamma}\} \geq p^{ai}(m - M_1)$ when $m \in C_{\gamma}$ and the 4^{th} inequality holds due to the following *Lemma.*

Lemma: $p^{ai}(m - M_1)$ is an exact p-value, i.e., $\Pr[p^{ai}(m - M_1) \leq \alpha] \leq \alpha$ for each $\alpha \in [0,1]$ under the null (2.1).

Proof of Lemma: Suppose H_0 (2.1) is true.

$$\begin{aligned}
\Pr[p^{ai}(m - M_1) \leq \alpha] &= \Pr[\max\{p(\boldsymbol{Y}^{ai}_1, \boldsymbol{Y}_0)\colon \boldsymbol{Y}_0 \in \Omega(M_0)\} \leq \alpha] \\
&\leq \Pr[p(\boldsymbol{Y}^{ai}_1, \boldsymbol{Y}^{ai}_0) \leq \alpha] \leq \alpha
\end{aligned}$$

where the 1^{st} inequality holds because $\max\{p(\boldsymbol{Y}^{ai}_1, \boldsymbol{Y}_0)\colon \boldsymbol{Y}_0 \in \Omega(M_0)\} \geq p(\boldsymbol{Y}^{ai}_1, \boldsymbol{Y}^{ai}_0)$ and the 2^{nd} inequality holds because $p(\boldsymbol{Y}^{ai}_1, \boldsymbol{Y}^{ai}_0)$ is an exact p-value under (2.1).

Proof of Proposition 2

Proof: Assume t is an invariant and effect increasing statistic. The proposition is proved if we can show $p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai}[1{:}M_0]) \geq p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0)$ for all $\boldsymbol{Y}_0 \in \Omega(M_0)$. Let $\boldsymbol{Y}_0$ be an element of $\Omega(M_0)$ and define the labels $k_1, \ldots, k_m$ such that $\boldsymbol{Y}_1^{ai} \cup \boldsymbol{Y}_0 = \{Y_{k_1}^{obs}, \ldots, Y_{k_m}^{obs}\}$. Let $\boldsymbol{Y}^{ai1} = (Y_{k_1}^{obs}, Y_{k_2}^{obs}, \ldots, Y_{k_m}^{obs})$ and $\boldsymbol{Z}^{ai1} = (Z_{k_1}, Z_{k_2}, \ldots, Z_{k_m})$.

Since t is invariant, we can assume without loss of generality that the labels $k_1, \ldots, k_m$ are defined such hat $Z_j^{ai1} = 1$ for $j = 1, \ldots, M_1$; $Z_j^{ai1} = 0$ otherwise; and $Y_j^{ai1} \geq Y_k^{ai1}$ if $j \leq k$ and $Z_j^{ai1} = Z_k^{ai1}$. In other words, the elements of $\boldsymbol{Z}^{ai1}$ are in descending order and the elements of $\boldsymbol{Y}^{ai1}$ are in descending order within fixed levels of $\boldsymbol{Z}^{ai1}$. Similarly define the labels $l_1, \ldots, l_m$ such that $\boldsymbol{Y}_1^{ai} \cup \boldsymbol{Y}_0^{ai}[1{:}M_0] = \{Y_{l_1}^{obs}, \ldots, Y_{l_m}^{obs}\}$. Let $\boldsymbol{Y}^{ai2} = (Y_{l_1}^{obs}, Y_{l_2}^{obs}, \ldots, Y_{l_m}^{obs})$ and define $\boldsymbol{Z}^{ai2}$ analogously. Assume the labels $l_1, \ldots, l_m$ are defined similar to $k_1, \ldots, k_m$ such that $\boldsymbol{Z}^{ai2} = \boldsymbol{Z}^{ai1}$ and $Y_j^{ai2} \geq Y_k^{ai2}$ if $j \leq k$ and $Z_j^{ai2} = Z_k^{ai2}$.

Note the first M_1 elements of $\boldsymbol{Y}^{ai1}$ and $\boldsymbol{Y}^{ai2}$ are the same, such that if $Z_i^{ai1} = 1$, then $Y_i^{ai2} = Y_i^{ai1}$. On the other hand, if $Z_i^{ai} = 0$, then $Y_i^{ai2} \geq Y_i^{ai1}$ because (i) $\boldsymbol{Y}^{ai2}$ and $\boldsymbol{Y}^{ai1}$ are both in descending order within fixed levels of $\boldsymbol{Z}^{ai2} = \boldsymbol{Z}^{ai1}$ and (ii) $\boldsymbol{Y}^{ai2}$ contains the M_0 largest values of $\{Y_i^{obs} \colon Z_i = 0, S_i^{obs} = 1\}$. This implies $(Y_j^{ai1} - Y_j^{ai2})(2Z_j^{ai} - 1) \geq 0$ for $j = 1, \ldots, m$. Since t is an effect increasing statistic, it follows $t(\boldsymbol{Z}^{ai1}, \boldsymbol{Y}^{ai1}) \geq t(\boldsymbol{Z}^{ai2}, \boldsymbol{Y}^{ai2})$, which implies $p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0^{ai}[1{:}M_0]) \geq p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0)$.

Proof of Proposition 3

Proof: Fix γ and α. Let $\boldsymbol{Z}$ denote a randomly selected treatment assignment vector and let p_γ^{ai} denote the resulting PSET p-value. By Proposition 2, there exists an $\tilde{m} \in C_\gamma$ such that $p_\gamma^{ai} - \gamma = p(\boldsymbol{Y}_1^{ai}, \boldsymbol{Y}_0[1{:}(\tilde{m} - M_1)])$. Similar to the proof of Proposition 2, define the labels $l_1, \ldots, l_{\tilde{m}}$ such that $\boldsymbol{Y}_1^{ai} \cup \boldsymbol{Y}_0[1{:}(\tilde{m} - M_1)] = \{Y_{l_1}^{obs}, \ldots, Y_{l_{\tilde{m}}}^{obs}\}$. Let $\boldsymbol{Y}^{ai\tilde{m}} \equiv (Y_{l_1}^{obs}, Y_{l_2}^{obs}, \ldots, Y_{l_{\tilde{m}}}^{obs})$ and define $\boldsymbol{Z}^{ai\tilde{m}}$ analogously. Let $\boldsymbol{y}_{\boldsymbol{0}}^{ai\tilde{m}} \equiv (y_{l_1}(0), \ldots, y_{l_m}(0))$ be the vector of potential outcomes if individuals $l_1, \ldots, l_m$ were all assigned control. Note if $z_{l_k} = 1$ then $l_k \in \mathscr{A}\mathscr{I}$ and thus $Y^{obs}_{l_k} = y_{l_k}(1) \geq y_{l_k}(0)$ under (2.1) or (2.9). On the other hand, by construction the l_k element of $\boldsymbol{Y}^{ai\tilde{m}}$ and the l_k element of $\boldsymbol{y}^{ai\tilde{m}}_{\boldsymbol{0}}$ are equal if $z_{l_k} = 0$. Because t is effect increasing, it follows $t(\boldsymbol{Z}^{ai\tilde{m}}, \boldsymbol{Y}^{ai\tilde{m}}) \geq t(\boldsymbol{Z}^{ai\tilde{m}}, \boldsymbol{y}^{ai\tilde{m}}_{\boldsymbol{0}})$ when either (2.1) or (2.9) hold. Therefore

$$\frac{\sum_{\boldsymbol{z} \in \Omega^{ai}_{\tilde{m}}} I[t(\boldsymbol{z}, \boldsymbol{Y}^{ai\tilde{m}}) \geq t(\boldsymbol{Z}^{ai\tilde{m}}, \boldsymbol{Y}^{ai\tilde{m}})]}{\binom{\tilde{m}}{M_1}} \leq \frac{\sum_{\boldsymbol{z} \in \Omega^{ai}_{\tilde{m}}} I[t(\boldsymbol{z}, \boldsymbol{Y}^{ai\tilde{m}}) \geq t(\boldsymbol{Z}^{ai\tilde{m}}, \boldsymbol{y}^{ai\tilde{m}}_{\boldsymbol{0}})]}{\binom{\tilde{m}}{M_1}} \tag{2.14}$$

The left side of (2.14) equals $p(\boldsymbol{Y}^{ai}_1, \boldsymbol{Y}_0[1{:}(\tilde{m} - M_1)])$ under H_A where as the right side of (2.14) equals $p(\boldsymbol{Y}^{ai}_1, \boldsymbol{Y}_0[1{:}(\tilde{m} - M_1)])$ under H_0. Therefore

$$\Pr[p(\boldsymbol{Y}^{ai}_1, \boldsymbol{Y}_0[1{:}(\tilde{m} - M_1)]) < \alpha - \gamma | H_A] \geq \Pr[p(\boldsymbol{Y}^{ai}_1, \boldsymbol{Y}_0[1{:}(\tilde{m} - M_1)]) < \alpha - \gamma | H_0]$$

which implies $\Pr[p^{ai}_\gamma < \alpha | H_A] \geq \Pr[p^{ai}_\gamma < \alpha | H_0]$, i.e., the probability of rejecting the null is at least as likely given (2.9) as compared to given (2.1).

Proof of Proposition 4

Proof: Let δ_0 be the true (unknown) value of the principal stratum direct effect. Fix $\gamma \in [0, 1]$ and $\alpha \in [0, 1]$. If $\gamma > \alpha$ then the PSET does not reject (2.11) for any choice of δ. Therefore, $\Pr[\delta_0 \notin [\Delta^\alpha_\gamma, \delta_{max}]] = \Pr[\delta_0 \notin [\delta_{min}, \delta_{max}]] = 0 \leq \alpha$. If $\gamma \leq \alpha$ then

$$\begin{aligned}
\Pr[\delta_0 \notin [\Delta^\alpha_\gamma, \delta_{max}]] &= \Pr[\delta_0 \notin [\Delta^\alpha_\gamma, \delta_{max}], m \in C_\gamma] + \Pr[\delta_0 \notin [\Delta^\alpha_\gamma, \delta_{max}], m \in \overline{C}_\gamma] \\
&\leq \Pr[\delta_0 \notin [\Delta^\alpha_\gamma, \delta_{max}], m \in C_\gamma] + \Pr[m \in \overline{C}_\gamma] \\
&\leq \Pr[\delta_0 < \Delta^\alpha_\gamma, m \in C_\gamma] + \gamma \\
&\leq \Pr[p^{ai}_{\gamma,\delta_0} \leq \alpha, m \in C_\gamma] + \gamma \\
&\leq \alpha - \gamma + \gamma = \alpha
\end{aligned}$$

where the 3^{rd} inequality follows from the definition of Δ_γ^α (i.e., $\delta_0 < \Delta_\gamma^\alpha$ implies $p_{\gamma,\delta_0}^{ai} \leq \alpha$) and the 4^{th} inequality follows for reasons analogous to the proof of Proposition 1.

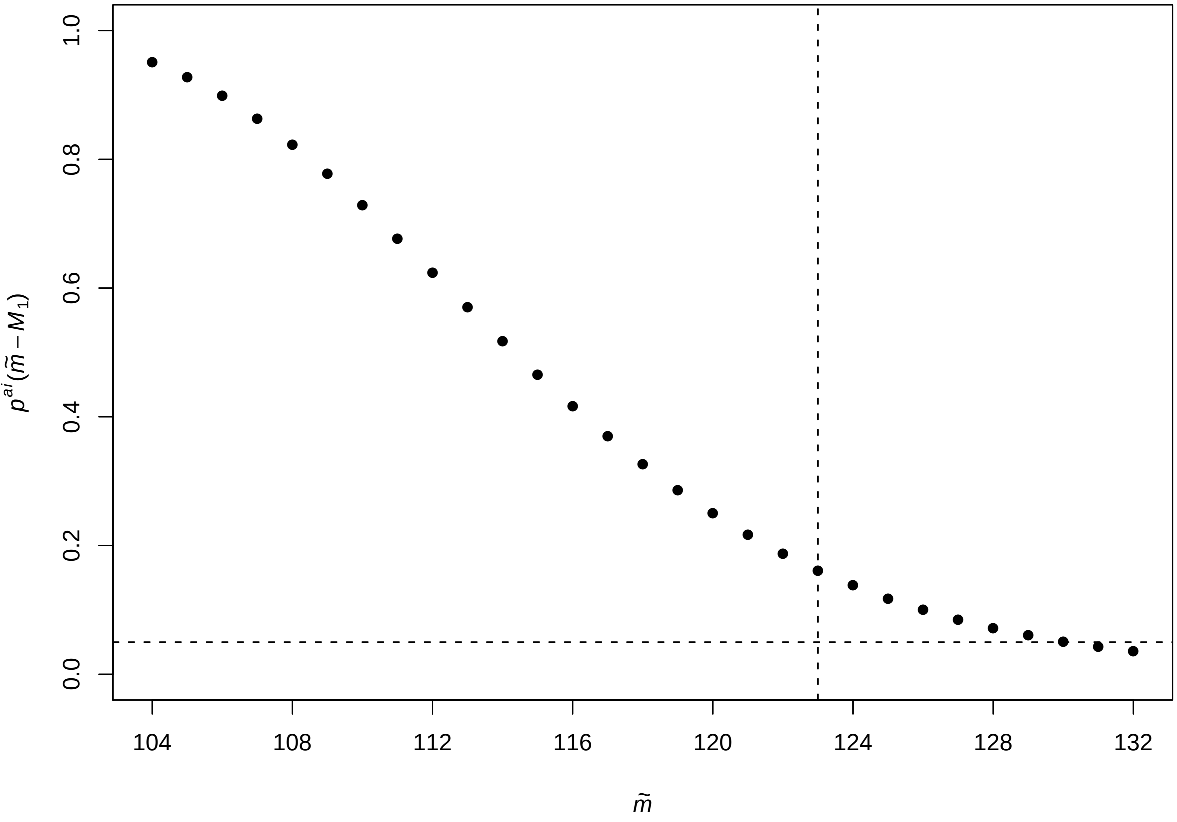
Table 2.1: Empirical type 1 error and power for α significance level, $100(1-\gamma)\%$ CI for m, and δ increase in log_{10} viral load in the AI stratum, where $\delta = 0$ under the null hypothesis (2.1)

α	γ	$\delta = 0$	$\delta = 1/3$	$\delta = 2/3$	α	γ	$\delta = 0$	$\delta = 1/3$	$\delta = 2/3$
0.05	0.005	0.001	0.09	0.65	0.05	0.045	0.002	0.13	0.73
0.05	0.010	0.002	0.12	0.72	0.10	0.010	0.004	0.18	0.79
0.05	0.020	0.003	0.16	0.77	0.10	0.050	0.009	0.29	0.89
0.05	0.025	0.004	0.16	0.77	0.10	0.090	0.009	0.25	0.86
0.05	0.030	0.005	0.17	0.78	0.10	0.095	0.005	0.21	0.84
0.05	0.040	0.003	0.15	0.77					

Table 2.2: Empirical type 1 error and power for $\alpha = 0.05$, $\gamma = 0.025$, and δ increase in log_{10} viral load in the AI stratum, where $\delta = 0$ under the null hypothesis (2.1), when adjusting for baseline CD4 count at various levels of ρ between viral load and CD4 count

ρ	$\delta = 0$	$\delta = 1/3$	$\delta = 2/3$	ρ	$\delta = 0$	$\delta = 1/3$	$\delta = 2/3$
0.0	0.004	0.16	0.76	0.5	0.002	0.17	0.82
0.1	0.004	0.16	0.76	0.6	0.002	0.18	0.85
0.2	0.004	0.16	0.77	0.7	0.001	0.21	0.90
0.3	0.003	0.16	0.78	0.8	0.001	0.27	0.96
0.4	0.003	0.16	0.80	0.9	< 0.001	0.50	> 0.99

Figure 2.1: Plot of the 29 conditional p-values, $p^{ai}(\tilde{m} - M_1)$, for mother-to-child HIV transmission weaning study. Horizontal reference line indicates significance level $\alpha = 0.05$. Vertical reference line indicates p^{ai}_{plug} where $\tilde{m} = \hat{M} = nM_1/\sum Z_i = 123$



Chapter 3

Addressing Selection Bias In Infectious Disease Prevention Studies

3.1 Introduction

Randomized trials are commonly used to estimate treatment effects. Such trials are often preferred over using observational data because the observational data may provide biased estimates of treatment effect if a subject characteristic that impacts treatment choice is associated with the outcome. For example, if women have a higher risk of infection (whether or not vaccinated) but are also more likely to choose vaccination then the vaccine effect will be underestimated as the beneficial effect will be hidden by the higher proportion of women observed in the vaccine group. Randomization removes this source of bias as it produces comparable groups with respect to baseline subject characteristics. This benefit is characterized as the randomized trial being "unconfounded" and implies that any observed difference between treatment groups can be attributed to treatment assignment (Rubin 2008).

Even in randomized studies bias can be introduced if analyses are conducted on a subset of the randomized population where the subset is defined by an event occurring post-randomization but prior to the outcome. These intermediate post-randomization events may cause the outcome to be missing for a subset of subjects or limit research interest to a subset of subjects (Rubin 2005). For example, one aim of prevention studies is to determine the treatment effect on post-infection outcomes including disease severity or death (Fitzgerald et al. 2011). Analyses of these endpoints often condition on the intermediate post-randomization event of infection since outcomes are only observed for infected individuals. If treatment reduces probability of infection, then the set of individuals with observed post-infection outcomes in the treated group is not comparable to the set in the control group. Therefore estimates conditional on infection status (e.g. difference between treatment groups in proportion of infected subjects who survive) are subject to selection bias. Likewise, analyses restricted to uninfected individuals in prevention studies suffer from the same potential bias (Gray et al. 2010). Other examples of intermediate post-randomization events include early infections in analysis of postpartum infections due to mother to child transmission (MTCT) via breast milk, presence of competing risks such as mortality that lead to truncation by death (i.e. missing outcomes due to death) and treatment non-compliance (Thior et al. 2006; Kuhn et al. 2008; Chasela et al. 2010; Grant et al. 2010; Gilbert et al. 2011).

One approach for addressing selection bias is to use intent-to-treat (ITT) methods that include all randomized individuals. For analyses of disease severity one ITT approach comprises assigning all uninfected individuals the lowest disease severity (i.e. burden of illness analyses) (Chang, Guess and Hayes 1994). However, such approaches are a joint test of the infection and post-infection outcomes and therefore fail to separate treatment effect on infection from that on post-infection outcomes. Such approaches are not ideal when the goal is to isolate the treatment effect on the post-infection outcome. An alternate approach that allows for testing the treatment effect solely on the outcome of interest is principal stratification (PS). While PS techniques appear in methodological literature and are implemented in other fields, these approaches have rarely been implemented in published infectious disease research likely due researchers being unfamiliar with these techniques. This paper introduces PS techniques to infectious disease research by implementing PS approaches to relevant study examples.

3.2 Principal Stratification Methodology

Within the potential outcomes framework each individual has a set of potential outcomes for any outcome of interest, the outcome the individual would have under treatment and that under control. A measure of the "causal effect" of an infectious disease preventative for an individual is the difference in the potential outcome the individual would have under treatment versus control (Holland 1986). This difference is written as $Y_i(0) - Y_i(1)$ where $Y_i(Z)$ is the potential outcome for the i^{th} individual under treatment Z (0 for control, 1 for active treatment). For example, vaccine effect on disease severity can be defined as decrease in severity for an individual when vaccinated compared to when not vaccinated. Because both potential outcomes cannot be observed for an individual, PS focuses on estimating average treatment effect.

PS categorizes individuals into principal strata based on their potential outcomes for the intermediate post-randomization event and then treatment effect on the outcome of interest is estimated within a stratum (Frangakis and Rubin 2002). For post-infection outcomes, the intermediate post-randomization event of infection is denoted as $S_i(Z)$ (1 if infected under treatment Z, 0 if not). The four principal strata are immune (never infected [NI]), harmed (infected under vaccine, not control), protected (infected under control, not vaccine) and doomed (always infected [AI]) individuals (Table 3.1). A randomized trial being unconfounded is also described as treatment assignment being "ignorable" in that treatment assignment is independent of the individual potential outcomes (Rosenbaum 1983). Due to ignorability, strata membership defined by potential outcomes for the intermediate post-randomization event is not affected by observed treatment assignment and therefore strata membership can be conditioned on without introducing bias. For example, analysis of post-infection outcomes within the AI stratum is unbiased and of interest as post-infection outcomes occur under both treatment assignments only for these individuals.

The PS approaches applied assume ignorability (obtained via randomization) and no interference (i.e. treatment assignment of one individual does not affect another individual's outcome). Because only one potential outcome is observed per subject, $S_i(1)$ or $S_i(0)$ depending on treatment assignment, additional assumptions are needed to determine strata membership (Frangakis and Rubin 2002). Monotonicity ($S_i(1) \leq S_i(0)$ for all individuals) can be assumed which implies in the post-infection outcome setting that treatment does no harm such that it does not cause infection. Accordingly, the harmed stratum is empty and infected active treated individuals belong to the AI stratum. To identify infected control individuals in the AI stratum, assumptions about possible selection bias are required. The most conservative approach assumes the maximum amount of selection bias.

For example, assume a one-sided test of the null hypothesis of no difference between treatment groups in the post-infection outcome among AI individuals versus the one-sided alternative that vaccine increases the post-infection outcome in AI individuals (e.g. vaccine is harmful with respect to disease severity such that it increases severity among subjects that would become infected regardless of treatment assignment). Assuming maximum selection bias means assuming the infected control individuals with the worst post infection outcomes are AI stratum members. Suppose N_Z individuals are randomized to the Z^{th} treatment group, randomization is $1:1$ such that $N_0 = N_1$, and n_Z individuals are infected in the Z^{th} group. A potential test statistic is simply the mean difference for individuals assumed to be in the AI stratum when assuming the maximum amount of selection bias, $T_M = {n_0}^{-1}\{\sum_{i=1}^{n_1} Y_i(1) - \sum_{i=n_0-n_1+1}^{n_0} Y_{(i)}(0)\}$

where $Y_{(i)}(0)$ is the i^{th} order statistic of the control group post-infection outcomes $\{Y_1(0), ..., Y_{n_c}(0)\}$ (Hudgens, Hoering and Self 2003). The distribution of the test statistic under the null and a corresponding p-value for the observed data can be obtained using bootstrap procedures. Large values of T_m are evidence against the null. All PS techniques employed in this paper are conducted using the sensitivityPStrat package in R.

3.3 Examples and Applications

3.3.1 Post-Infection Outcomes and Truncation by Death

The ZEB trial evaluated whether abrupt weaning at 4 months (intervention) versus continued breastfeeding (control) increases survival of children of HIV infected mothers (Kuhn et al. 2008). The trial included 958 HIV-infected women and their infants with 481 children randomized to intervention and 477 to control. Randomization occurred at one month postpartum to allow for preparation time for weaning at 4 months. Kuhn presents analyses of the weaning effect on 24 month survival based on a log-rank test for the subset of infants HIV-infected and alive at 4 months. Sixty-two individuals assigned to intervention were HIV-infected and alive at 4 months, 39 (63%) who died prior to 24 months. Likewise, 70 assigned to control were HIV-infected and alive at 4 months, 32 (46%) who died prior to 24 months. The reported p-value was 0.007, and the authors concluded there is evidence of a harmful effect of weaning on survival among HIV positive infants alive at 4 months.

Because the analysis conditioned on infection and survival status at 4 months, the findings could be due to selection bias. Specifically, any differences between groups during months 1-4 could affect infection and survival status at 4 months. For instance, at month 2 women assigned intervention were counseled on weaning techniques and provided infant formula and fortified weaning cereal. This may have caused women assigned intervention to wean earlier than had they been assigned control, in turn perhaps impacting HIV acquisition. In fact, more women assigned intervention weaned by 4 months (37 versus 18) and, possibly because of this, fewer infants in the intervention arm became HIV positive at or before 4 months (71 versus 81).

The AI stratum includes all infants who would be HIV-infected and alive at 4 months regardless of randomization assignment. The Hudgens, Hoering and Self (2003) PS approach was used to test the null hypothesis of no intervention effect on death in the AI stratum (2003). While the reported analyses used time-to-event data, the PS analyses use the binomial outcome of 24 month survival as individual death and censoring times were not reported. Applying the conditional approach of the published results to a binomial outcome also results in a significant p-value (0.036). The PS p-value was 0.22 suggesting no evidence of a harmful effect of weaning on survival for the AI stratum (Table 3.2).

3.3.2 Exclusion of Early Infections

The BAN trial evaluated whether daily administration of nevirapine (NVP) to infants of HIV infected mothers through 28 weeks decreased breast milk MTCT risk versus no antiretroviral therapy (control) (Chasela et al. 2010). A total of 668 mother-infant pairs were randomized to control and 852 to NVP. Fewer mother-infant pairs were randomized to control because the data monitoring committee (DMC) stopped enrollment in this arm early. Treatment effect on infection through 28 weeks was assessed using a log-rank test that compared infection between groups among infants not infected at two weeks. Of the 632 control infants not infected at two weeks , 32 (5.1%) were infected by 28 weeks. Of the 815 NVP infants not infected at two weeks, 12 (1.5%) were infected by 28 weeks. The log-rank p-value was < 0.001, suggesting NVP prevents breast milk MTCT.

Because the analyses conditioned on 2-week infection status, the findings could be due to selection bias. To guard against this bias, an ITT analysis including infants infected before two weeks was reported. However, given a primary objective of BAN was to investigate the NVP effect on breast milk transmission, the investigators were primarily interested only in infections occurring after two weeks, as infections prior to two weeks may have occurred *in utero* or during birth. Because NVP from birth could potentially effect infection status at two weeks, the infants not infected at two weeks in each group may not be comparable.

The NI stratum includes all infants who would be HIV uninfected at two weeks regardless of randomization assignment. A PS technique developed by Shepherd, Gilbert and Lumley was used to test the null of no NVP effect on infection by 28 weeks in the NI stratum (2007). In contrast to the AI stratum, NI stratum membership is known for control individuals not infected at two weeks since by monotonicity infants not infected at two weeks under control would also not be infected at two weeks under NVP. Membership in the NI stratum is unknown for NVP infants not infected at two weeks. Thus the PS technique is applied to the NI stratum with the roles of the treated and control individuals reversed relative to conducting the test in the AI stratum. Because enrollment was stopped early in one group, we use only data available prior to the DMC decision. These data include all control infants but only 670 NVP infants, 639 of who were not infected at two weeks. Of these 639 infants, 10 (1.6%) were infected by 28 weeks. The PS p-value was < 0.001 indicating a benefit of NVP among infants who were immune to infection at two weeks (Table 3.3).

3.4 Discussion

PS techniques are appropriate for addressing potential selection bias in analyses using only a subset of the randomized population defined by a post-randomization event. While the examples presented above consider post-randomization events of mortality and infection status, the approach is also applicate to other post-randomization events including compliance.

For the compliance setting, consider the iPrEx trial which evaluated whether antiretroviral chemoprophylaxis (intervention) decreased HIV transmission in men who have sex with men versus placebo (control) (Grant et al. 2010). A log-rank test that compared infection between treatment groups resulted in a p-value of 0.005, suggesting chemoprophylaxis prevents transmission. Because pill use was lower in the intervention group versus control, the authors also conducted analyses adjusting for treatment compliance ($\geq$ 90% pill use). However, these analyses are subject to selection bias since they incorporate the post-randomization event of observed compliance. In this setting, the principal strata are defined by whether or not subjects take more than 90% of the active chemoprophylaxis pills such that the four strata are always compliers (take under intervention, not control), never takers, always takers and defiers (take under control, not intervention) (Table 3.4). Monotonicity holds as subjects did not have access to chemoprophylaxis unless assigned intervention and therefore the always takers strata is empty. As such, any individuals who take more than 90% of the chemoprophylaxis pills when assigned intervention will not take any chemoprophylaxis pills under control and so are always compliers. However, individuals in the control group are a mixture of always compliers and never takers. Thus this scenario is similar to conducting analyses for the AI stratum of our previous examples and therefore the methods applied above are applicable to this scenario. Of note regarding the above methods is that compliance with respect to the placebo pill is ignored in these analyses. While the exclusion of this information may be considered a limitation of the approach, it could also be argued that compliance with respect to placebo should be ignored as it does not necessarily provide information regarding compliance with respect to active therapy as subjects are likely to comply or not comply with active therapy for different reasons than placebo (e.g., failure to comply due to perceived lack of efficacy on placebo versus failure to comply due to side effects on active therapy).

As with all analytic approaches, there are limitations with the PS techniques. Primarily, PS methods are only valid when the underlying assumptions hold. For the employed approaches, the key assumptions that may not always be valid are monotonicity and no interference. For example, monotonicity fails in the post-infection setting if the prophylactic treatment can cause infection or in the treatment non-compliance setting if control subjects can receive treatment. Additionally, conclusions based on PS techniques are limited to the stratum to which the analyses are restricted. For example, a test of disease severity in the AI stratum will not detect the scenario in which there is no difference in the disease severity in the AI stratum but all infected individuals in the harmed stratum have severe disease and all individuals in the protected stratum have mild disease (Joffe 2011). However, ITT approaches could be used to test for this type of effect. In general, while PS techniques are not intended to answer all questions in infectious disease prevention research, PS is an appropriate alternative to the often used but inappropriate approach of conditioning on observed post-randomization events when such events may introduce selection bias.

Table 3.1: Potential outcomes for post-infection endpoint example

	Infection		Post-Infection	
Principal Strata	$S(0)$	$S(1)$	$Y(1)$	$Y(0)$
Immune (Never infected)	0	0	$*$	$*$
Protected	1	0	$Y(1)$	$*$
Harmed	0	1	$*$	$Y(0)$
Doomed (Always Infected)	1	1	$Y(1)$	$Y(0)$

Table 3.2: ZEB trial analysis results

	Control	Intervention
Randomized	477	481
Infected and Alive at 4 Months	70	62
Post-infection Deaths at 24 Months	32(46%)	39(63%)
Published p-value	0.007	
- Using Binary Outcome	0.036	
AI p-value	0.215	

Table 3.3: BAN trial analysis results

	Control	Intervention	
		All	pre-DSMB
Randomized	668	852	670
Not Infected at 2 Weeks	632	815	639
Infections among Non-infectees at 28 Weeks	32(5.1%)	12(1.5%)	10(1.6%)
Published p-value: All		< 0.001	
NI p-value: pre-DSMB		< 0.001	

Table 3.4: Principal strata for compliance example ($S(Z) = 1$ for $> 90\%$ chemoprophylaxis and 0 otherwise)

Principal Strata	$S(0)$	$S(1)$
Never Takers	0	0
Defiers	1	0
Always Compliers	0	1
Always Takers	1	1

Chapter 4

Analysis of Repeated Low-Dose Challenge Studies

4.1 Introduction

Preclinical proof-of-concept vaccine trials using animal models limit the risk, time and cost of clinical trials involving human subjects by providing preliminary evidence of potential safety and efficacy of an investigational vaccine (Koff 2006; Shedlock, Silvestri and Weiner 2009). A large portion of the preclinical studies of HIV vaccines have been conducted using macaques as the disease progression of simian immunodeficiency viruses in macaques mirrors that of HIV in humans (Shedlock et al. 2009). The virus challenge in these preclinical trials has historically been administered via a single high-dose intravenous or mucosal inoculation which often resulted in a high probability of infection for all unprotected macaques (McDermott et al. 2004; Straprans, Feinberg, Shiver and Casimiro 2010).

Although single high-dose challenge studies are appealing in that high infection rates allow for a greater chance of observing a vaccine effect assuming a vaccine is completely protective against infection, the vaccine efficacy in these trials may not translate to 'real life' (Straprans et al. 2010). For example, high infection rates in challenge studies do not mirror the low probability of heterosexual HIV transmission per sexual act or low per month probability of late postnatal HIV transmission via breastfeeding (Gray et al. 2001; Boily et al. 2009; WHO 2007). Vaccines may not be equally efficacious against high-dose and low-dose challenges, such that vaccines efficacious against low-dose challenges may be discarded due to not demonstrating efficacy in high-dose challenge studies (Regoes, Longini, Feinberg, and Straprans 2005).

As an alternative, repeated low-dose mucosal challenge studies have been employed more recently (McDermott et al. 2004; Hessell et al. 2009; Hudgens et al. 2009). In these studies, each macaque is challenged multiple times (i.e., until infected or a maximum number of challenges C_{max} are performed) such that each macaque has numerous observed outcomes corresponding to the number of times the macaque was challenged. Regoes, et al. (2005) performed a power analysis of repeated low-dose challenge (RLC) studies demonstrating that these trials are viable alternatives to traditional single high-dose challenge studies.

A standard analysis used for the single-dose challenge study entails performing Fisher's exact test on a 2×2 contingency table of infection status by treatment assignment where cell counts are the number of the macaques in each category (e.g., vaccinated/infected, vaccinated/not-infected, control/infected, control/not-infected). For the RLC setting, Regoes et al. proposed the Fisher's exact test be conducted on a similar 2×2 contingency table of infection status by treatment assignment except that the cell counts are the number of challenge events falling in each category. Hudgens et al. (2009) showed via simulations that in certain settings this approach has an inflated type I error. Alternative analytic approaches implemented in this setting include the exact log-rank test implemented using StatXact as well as non-exact (i.e., large sample) approaches including the log-rank test and Cox proportional hazard modeling (Cytel Software Corporation 2007; García-Lerma et al. 2008; Hessell et al. 2009; Parikh et al. 2009; Reynolds et al. 2010). Unfortunately, some of these analytic approaches employed in the RLC setting are not appropriate in that assumptions required for these tests are not met while other analysis methods may not be easy to implement using software available to clinical researchers.

This paper describes appropriate analytic approaches for RLC studies. The important aspects of the data collected from RLC studies are that 1) infection status is a time to event endpoint subject to censoring, and 2) sample sizes are often very small such that large sample, frequentist analytic approaches may yield incorrect inference. Given that macaques are challenged at discrete time-points with determination of infection and censoring status occurring within the interval between each challenge, the resulting data are naturally modeled using discrete time-to-event survival methods.

The outline of this paper is as follows. Section 4.2 introduces notation used in the remainder of the paper and Section 4.3 reviews randomization-based inference, a mode of inference appropriate for randomized studies with small sample sizes. Section 4.4 provides details regarding why Fisher's exact test is invalid for RLC studies. Valid analytic approaches for RLC studies are provided in Section 4.5 and compared in Section 4.6. In Section 4.7 the various methods are compared using data from a recent RLC study. Section 4.8 includes a discussion and Section 4.9 provides SAS and R code for the various analytic approaches.

4.2 Notation

Notation will be defined for both single and repeated challenge studies to allow for discussion and comparison. Suppose there are n macaques. In the single challenge study, let $x_i(z)$ denote the potential infection outcome when macaque i is assigned z, where z=0 denotes control and z=1 denotes vaccine. Let $x_i(z) = 1$ if the macaque becomes infection and $x_i(z) = 0$ otherwise. Prior to the study each macaque has two potential outcomes, only one of which is observed during the trial, which we denote by

$X_i^{obs} = x_i(0)(1 - Z_i) + x_i(1)Z_i = X_i(Z_i)$ where Z_i is the treatment randomly assigned for the i^{th} macaque.

For RLC studies, let $t_i(z)$ denote the potential outcome for the number of challenges each macaque receives when assigned z, where $t_i(z) \leq C_{max}$. Similarly, let $d_i(z)$ denote the potential outcome for infection indicator, where $d_i(z) = 1$ if the macaque would become infected when assigned z and $d_i(z) = 0$ otherwise. Only either $(t_i(1), d_i(1))$ or $(t_i(0), d_i(0))$ are observed, which we denoted by (T_i^{obs}, D_i^{obs}) defined in a similar manner to X_i^{obs}.

4.3 Randomization-Based Inference

Since preclinical challenge studies typically randomize a small number of macaques, randomization-based statistical methods are ideal for inference about vaccine effect. Randomization-based inference is based on distributions created from the randomization process rather than assuming random sampling from an infinite population or that particular parametric distributions hold (Koch, Gillings and Stokes 1980; Rubin 1991; Rosenbaum 2002a). Under randomization-based inference the potential outcomes, (i.e., $(x_i(1), x_i(0))$ or $(t_i(1), t_i(0), d_i(1), d_i(0)))$ are considered fixed, discrete features of the finite population of n macaques and Z_i is considered a random variable. As observed outcomes are functions of treatment assignment, they are also considered random.

Consider the null hypotheses that vaccine has no effect on any of the n macaques for a single challenge study:

$$H_0 : x_i(0) = x_i(1) \text{ for } i = 1, \ldots, n. \qquad (4.1)$$

Likewise, the null hypothesis in a RLC study is:

$$H_0 : d_i(0) = d_i(1) \text{ and } t_i(0) = t_i(1) \text{ for } i = 1, \ldots, n. \tag{4.2}$$

This type of null is referred to as Fisher's sharp null hypothesis of no effect and states potential outcomes under both treatment groups are the same (Rubin 2005).

Under the null (4.1 or 4.2) all potential outcomes are observed for each macaque and therefore, the observed outcomes are fixed regardless of treatment assignment. Assuming a permutation randomization scheme is employed such that exactly m macaques are randomly assigned to vaccine and the remaining to control, there are $\binom{n}{m}$ possible treatment assignment combinations that are all equally likely to be selected. Therefore, a simple approach for calculating a p-value for testing H_0 is to define the p-value as the probability of obtaining a treatment assignment combination that results in a distribution of outcomes as or more extreme than what was observed when assuming all treatment assignments are equally likely. Specifically, a test statistic is calculated for all permutations of treatment assignment and the p-value is simply the proportion of test statistics as or more extreme than the test statistic value calculated based on the observed data and treatment assignment. Since the p-value in randomization-based inference is obtained by permuting treatment assignment, the tests are often referred to as permutation tests. The resulting p-values are considered exact in that they are based on calculating the exact permutation distribution of the test statistic as opposed to relying on asymptotic properties of the statistic.

4.4 Fisher's Exact Test

Fisher's exact test is a randomization-based test for 2 × 2 tables that calculates p-values based on the probability of the observed table under Fisher's sharp null hypothesis that the vaccine has no effect on the number of challenges until infection for any macaque. While Regoes et al. proposed using Fisher's exact test for RLC studies, this test is not valid for testing (4.2) in that it is not guaranteed to protect against inflated type I error. To further detail the inappropriateness of the test, we first illustrate its justification for single challenge studies.

4.4.1 Single Challenge Study

Suppose a single challenge study is conducted where four macaques are randomized such that 2 receive vaccine and 2 receive control. The $\binom{4}{2} = 6$ possible treatment assignment combinations are all equally likely, each with probability of 1/6. Assume the observed treatment assignments are $Z_1 = Z_2 = 1$ and $Z_3 = Z_4 = 0$. Further assume only one macaque, $i = 1$ is uninfected ($X_1^{obs} = 0$ and $X_2^{obs} = X_3^{obs} = X_4^{obs} = 1$). Under the null where potential outcomes are observed for all macaques, the potential treatment assignments include:

Comb	$Z_1Z_2Z_3Z_4$	$\sum_i Z_i X_i^{obs}$	$\sum_i Z_i(1-X_i^{obs})$	$\sum_i(1-Z_i)X_i^{obs}$	$\sum_i(1-Z_i)(1-X_i^{obs})$
1	1100	1	1	2	0
2	1010	1	1	2	0
3	1001	1	1	2	0
4	0110	2	0	1	1
5	0101	2	0	1	1
6	0011	2	0	1	1

(4.3)

where combination 1 is observed. The Fisher's exact p-value for H_0 (4.1) versus a one-sided alternative that vaccine has a protective effect is calculated as the proportion of combinations that result in an outcome distribution at least as extreme as the observed data (in the direction of the alternative). This set includes combinations where the number of infected vaccine macaques is ≤ 1 (combinations 1-3). Therefore the p-value is $3/6 = 0.5$.

Since multiple treatment assignment combinations result in the same 2×2 table the calculations for Fisher's exact test only consider unique tables (e.g., treatment permutation combinations 1-3 all result in the same 2×2 table). As such, the p-value is calculated by summing the probabilities of each unique table that is as or more extreme than the observed table (in the direction of the alternative). The table based on observed data is constructed as:

	Infected	Non-infected	
Vaccine	$\sum_i Z_i X_i^{obs}$	$\sum_i Z_i(1 - X_i^{obs})$	$\sum_i Z_i$
Control	$\sum_i(1 - Z_i)X_i^{obs}$	$\sum_i(1 - Z_i)(1 - X_i^{obs})$	$\sum_i(1 - Z_i)$
	$\sum_i X_i^{obs}$	$\sum_i(1 - Xi^{obs})$	n

(4.4)

For example (4.3), this table is:

	Infected	Non-infected	
Vaccine	1	1	2
Control	2	0	2
	3	1	4

(4.5)

Fisher's exact test assumes row and column margins are considered fixed. As such, only one other table in addition to (4.5) is possible for example (4.3). This other table switches the cells for vaccine and control macaques such that there is one non-infected vaccine macaque and zero non-infected control macaques. The probability for each table under the null (4.1) is obtained by realizing the first table cell, $\sum_i Z_i X_i^{obs}$ has a hypergeometric distribution under the fixed margins assumption. For example (4.3), the probability for each table is 0.5 and therefore, the final p-value is 0.5.

In single challenge studies, the fixed margins assumption holds as the number of macaques assigned each treatment is fixed by design and the number of infected and non-infected macaques are observed, fixed features of the finite population under (4.1).

4.4.2 Repeated Low-Dose Challenge Studies

Expand the example such that $C_{max} = 2$ and all infected macaques are infected on the first challenge. Thus, $(T_i^{obs}, D_i^{obs}) = (2, 0)$ for $i = 1$ and $(T_i^{obs}, D_i^{obs}) = (1, 1)$ for $i = 2, 3, 4$. Regoes et al. propose conducting a Fisher's exact test of the following table:

	Infected	Non-infected	
Vaccine	$\sum_i Z_i D_i^{obs}$	$\sum_i Z_i (T_i^{obs} - D_i^{obs})$	$\sum_i Z_i T_i^{obs}$
Control	$\sum_i (1 - Z_i) D_i^{obs}$	$\sum_i (1 - Z_i)(T_i^{obs} - D_i^{obs})$	$\sum_i (1 - Z_i) T_i^{obs}$
	$\sum_i D_i^{obs}$	$\sum_i (T_i^{obs} - D_i^{obs})$	$\sum_i T_i^{obs}$

(4.6)

For the expanded example, this table is:

	Infected	Non-infected	
Vaccine	1	2	3
Control	2	0	2
	3	2	5

(4.7)

To apply Fisher's exact test, the probabilities of the table given in (4.7) and 2 other tables are calculated. The first table (a) has 1 non-infection event each in the vaccine and control groups and the second table (b) has both non-infection events in the control group. However, it is not possible to observe table (a) as both non-infection events occur within one macaque and therefore the 2 'non-infected' events cannot be divided between the vaccine and control rows of the table. In general, Fisher's exact test in this setting uses the incorrect set of potential tables such that non-zero probabilities of observation are assigned to tables that are not actually observable when permuting treatment assignments among macaques and zero probabilities are assigned to tables that are observable.

More formally, the assumption of fixed margins required for Fisher's exact test is not valid in this setting because the number of challenges per treatment group (row margins) are not fixed. This violation is the cause for the inflated type I error. For example, assume a trial with $n = 6$ macaques where 5 are infected after the first challenge and the remaining macaque remains uninfected. Regardless of treatment assignment, the probability of rejecting the null (4.2) in favor or a two-sided alternative for all $\alpha > 0.012$ is 1 as the Fisher exact test two-sided p-value is always 0.012.

4.5 Analytic Approaches

Both parametric and nonparametric survival analysis methods are available for analyzing right-censored discrete time to event data. Methods frequently employed include the nonparametric log-rank test statistic as well as parametric-based approaches that typically use large-sample likelihood methods based on the asymptotic distribution of one of three test statistics: the likelihood ratio test (LRT), score and Wald's test statistics. These statistics and associated large sample p-values can be obtained via a variety of statistical packages. Randomization-based p-values for these test statistics can be obtained using standard packages but options are limited. This section details these tests as implemented in discrete time settings including asymptotic distributions used for large-sample p-vales and methods for obtaining randomization-based p-values.

4.5.1 Nonparametric Log-Rank Tests

The log-rank test is nonparametric in the sense that no assumptions are required about the distribution of the survival time random variable. Since time is discrete, macaques censored at challenge t are accounted for by excluding these macaques from the number of macaques at risk for all subsequent time-points.

The conditional probability of infection at challenge t for randomized treatment group z, p_t is estimated as $D_t(z)/N_t(z)$ where $N_t(z) = \sum_{i:Z_i=z} I(T_i^{obs} \geq t)$ is the number of macaques at risk and $D_t(z) = \sum_{i:Z_i=z} I(T_i^{obs} = t, D_i^{obs} = 1)$ is the number infected. The joint probability function of infection at challenge t in the vaccine and control group is expressed as the product of independent binomial terms, one for each treatment group where p_t is common to both treatment groups under the null (4.2). As such the conditional distribution of $D_t(1)$ given the total number of infections observed for challenge t, $D_t = D_t(0) + D_t(1)$ has a hypergeometric distribution. The log-rank test statistic is then defined as $LR = \sum_{t=1}^{C_{max}} \{D_t(1) - E(D_t(1))\}$ where $E(D_t(1)) = N_t(1)D_t/N_t$, $N_t = N_t(0) + N_t(1)$, and $E(D_t(1))$ is the expected number of infections in the vaccine group under the null based on the hypergeometric distribution.

P-values are obtained by dividing LR by its variance assuming a hypergeometric distribution $LR_{CMH} = LR/\sum_{t=1}^{C_{max}} V(D_t(1))$ where $V(D_t(1)) = N_t(1)N_t(0)/\{D_t(N_t - D_t)N_t^2(N_t - 1)\}$. The LR_{CMH} is equivalent to the test statistic obtained stratified Cochran-Mantel-Haenszel test where strata are defined by time. As such, LR_{CMH} would have an asymptotically chi-square distribution if the C_{max} tables were independent (Mantel 1966). While these tables are clearly not independent, in the discrete time setting LR_{CMH} is also equivalent to the partial likelihood-based score test statistic (detailed in Section 4.5.2) and therefore asymptotic p-values for the LR_{CMH} are obtained using a chi-square distribution (Kalbfleisch and Prentice 1980).

4.5.2 Parametric Tests

For parametric approaches, the probability of the i^{th} macaque being infected at challenge t is defined as $f_{it}(z) = Pr(t_i(z) = t)$. The corresponding survival function and hazard rate are defined as $S_{it}(z) = Pr(t_i(z) \geq t) = \sum_{j=1}^{t} f_{it}(z)$ and $p_{it}(z) = Pr(t_i(z) = t | t_i(z) \geq t) = f_{it}(z)/S_{it}(z)$. The survival function can be expressed as $S_{it}(z) = (1 - p_{i1}(z))(1 - p_{i2}(z)) \dots (1 - p_{i(t-1)}(z))$.

A model for the conditional odds of infection and its dependence on time and treatment often assumed for discrete time is the semi-parametric Cox odds model which is also referred to as a relative risk model

$$\frac{p_{it}(z)}{1-p_{it}(z)} = \frac{p_{0t}}{1-p_{0t}} \exp(z_i\beta) \tag{4.8}$$

which can be also written as $\text{logit}(p_{it}(z)) = \alpha_t + z_i\beta$ where $\alpha_t = \text{logit}(p_{0t})$ is a function of the baseline hazard at each time-point and β is the shift in the hazard caused by treatment (Cox 1972). Typically interest is primarily focused on the treatment effect β and not the baseline log odds α_t, in which case αs are treated as unknown nuisance parameters and the basis of inference is the partial likelihood function:

$$L_{partial} = \prod_{t=1}^{C_{max}} \frac{\exp(\beta)\sum_{i\in\mathscr{D}_t} Z_i}{\sum_{q\in\mathscr{Z}_t} \exp(\beta)\sum_{l\in q} Z_i} \tag{4.9}$$

where $\mathscr{D}_t$ is the set of D_t macaques infected at challenge t and $\mathscr{Z}_t$ is the set of all subsets of macaques of size D_t chosen from the set of N_t macaques at risk at t without replacement. Inference about β then proceeds by applying the usual large sample maximum likelihood methods based on (4.9).

Alternatively model (4.8) can be fit using a logistic regression model with standard parametric maximum likelihood approaches providing estimates of both β and the αs (Brown 1975, Allison 1982, Singer and Willett 1993). These parameter estimates are used to obtained estimates of odds ratios (as opposed to relative risks). Estimates of β obtained by maximizing the full likelihood compared to maximizing the partial likelihood (4.9) will tend to be similar but not identical. Benefits of the binomial likelihood approach include that restrictions can be placed on the αs and inclusion of additional covariates (both fixed and time-varying) in the binomial likelihood approach is less computationally demanding than in the partial likelihood approach (Allison 1982, Singer and Willett 1993, Allison 2010).

Specifically, under the assumption that macaques are independent the likelihood function can be written as

$$L_{full} = \prod_{i=1}^{n}\prod_{t=1}^{T_i^{obs}} p_{it}(z)^{X_{it}^{obs}}(1-p_{it}(z))^{1-X_{it}^{obs}} \tag{4.10}$$

where X_{it}^{obs} for $t = 1, ..., T_i^{obs}$ are the observed outcomes for each macaque ($X_{it}^{obs} = 0$ if the macaque is not infected and 1 if infected at the t^{th} challenge), $Pr(t_i(z) = T_i^{obs}) = p_{it}(z)\prod_{t=1}^{T_i^{obs}-1}(1 - p_{it}(z))$ is the probability of an uncensored macaque becoming infected at T_i^{obs} and $Pr(t_i(z) > T_i^{obs}) = \prod_{t=1}^{T_i^{obs}}(1 - p_{it}(z))$ is the probability of a censored macaque remaining uninfected at the last challenge (Singer and Willett 1993). Since this likelihood is equivalent to the likelihood of $N_t = T_1^{obs} + ... + T_n^{obs}$ independent Bernoulli trials with probability parameters $p_{it}(z)$, maximizing (4.10) using the logistic parameterization in (4.8) provides maximum likelihood estimates for the αs, β and $p_{ij}(z)$.

Assuming the probability of infection is constant within treatment groups and across challenges (i.e., $p_{it}(z) = p_z$ for all t for all macaques receiving z), the likelihood is

$$L_{full} = \prod_{i=1}^{n}\{p_z(1-p_z)^{X_i^{obs}-1}\}^{I(D_i^{obs}=1)}[(1-p_z)^{X_i^{obs}}]^{I(D_i^{obs}=0)} \tag{4.11}$$

This likelihood represents a vaccine that has a leaky effect in that infection susceptibility is changed by a constant factor at each challenge for vaccinated macaques.

The partial (4.9) or full likelihood (4.11) can be used to test for a vaccine effect by computing the LRT, score, or Wald statistics given, respectively, by:

$$-2\log\{L(0)/L(\hat{\beta})\} \tag{4.12}$$

$$\left\{\frac{\partial}{\partial\beta}\log L(0)\right\}^2 / \left\{\frac{\partial^2}{\partial\beta^2}\log L(0)\right\} \tag{4.13}$$

$$\hat{\beta}^2 / \left\{\frac{\partial^2}{\partial\beta^2}\log L(\hat{\beta})\right\} \tag{4.14}$$

where L is either the partial or full likelihood. These test statistics are all asymptotically chi-square distributed with one degree of freedom for a test of vaccine versus control; however, the LRT generally most closely follows the asymptotic distribution in small to moderate sample size settings (Kalbfleisch and Prentice 2002).

4.5.3 Implementation

Large sample p-values for the discrete time setting LRT, score/log-rank and Wald statistics can be obtained via a variety of software packages. For example, p-values for the LR_{CMH} log-rank test statistic can be obtained using the LIFETEST procedure from SAS® software and the SURVDIFF function from the R SURVIVAL package (SAS Institute Inc. 2008; R Development Core Team 2010). Since this statistic is equivalent to the partial likelihood-based score test, it can also be obtained using the FREQ and PHREG procedures in SAS or corresponding CMH_EST and COXPH functions in R. The LRT and Wald statistics from the SAS PHREG and LOGISTIC procedures will differ slightly as PHREG uses the partial likelihood and LOGISTIC uses the full likelihood. In R, the COXPH function obtains the LRT and Wald statistics based on the partial likelihood while the GLM function obtains these statistics using the full likelihood.

Exact randomization-based p-values using these statistics can be obtained by calculating the probability of obtaining a value for the test statistic as or more extreme than the observed test statistic under the sharp null (4.2). This probability is simply the proportion of test statistic values arising from permuting treatment assignment that are greater than or equal to the observed test statistic. Built-in procedures for calculating randomization-based p-values based on these test statistics are limited.

StatXact® includes a built in procedure for obtaining randomization-based p-values for the log-rank test statistic (Cytel Software Corporation 2007). While the log-rank test statistic employed in StatXact is a valid randomization-based test, the test statistic is not equivalent to LR or LR_{CMH} in the discrete time setting (Callaert 2003). Specifically, the exact 'log-rank test' in StatXact is based on Savage scores and therefore is only equivalent to the LR when there are no ties or censoring. An exact p-value based on LR can be obtained in StatXact by calculating log-rank scores for each individual and then conducting a general permutation test (Callaert 2003). Both approaches are valid and the difference in the resulting p-values are typically minimal. Because StatXact is a specialized licensed software package for performing exact inference it is less frequently available in comparison to SAS or R. Accordingly alternate approaches that are easier to employ are preferred. The SURVTEST function from the R COIN package also produces randomization-based p-values for a log-rank test statistic that converges to a z-test statistic (as opposed to the chi-square distribution). The square of the resulting statistic will be close to LR_{CMH} but not equivalent. While the numerator of the squared SURVTEST log-rank test statistic is LR, the denominator is a variance estimate conditional on integer-valued log-rank scores instead of the variance estimate used in LR_{CMH} (Hothorn and Lausen 2003).

Alternatively, exact conditional logistic regression models using the SAS LOGISTIC procedure with an EXACT option provide a randomization-based p-value based on the score statistic. Exact conditional logistic regression comprises generating the permutation or exact distribution for the parameter of interest based on the likelihood conditional on sufficient statistics for all other parameters in the model which are considered nuisance parameters (Cox 1970). In the simplest case where time is the only covariate in the model besides treatment, conditioning on sufficient statistics for time is equivalent to conditioning on the margins of the challenge life-tables. Thus the resulting full likelihood score statistic is equivalent to LR_{CMH} and the partial likelihood score statistic.The procedure uses multivariate shift algorithm for processing through the treatment permutations and producing the exact p-value (Hirji et al. 1987). Besides being easily employed in SAS, exact conditional logistic regression also allows for covariate adjustment and treatment effect estimation, neither of which apply to exact log-rank tests in StatXact or R.

While no built-in procedures provide randomization-based p-values based on Wald or LRT statistics, such p-values can be obtained via user-developed code that calls SAS PHREG or LOGISTIC procedures or R COXPH or GLM functions for all possible treatment assignment permutations. Advanced programing knowledge is required and computations quickly become infeasible as sample size increases as there are $\binom{n}{m}$ possible treatment assignment combinations. In cases where obtaining all permutations is not possible, the exact p-value can approximated using a Monte Carlo sampling approach.

Details on the data construct and syntax required for each of these approaches are detailed in Section 4.9.

4.5.4 Modifying the Likelihood

The likelihood used for the parametric tests previously described assumes a leaky effects model where all macaques within a treatment group have the same probability of being infected at each challenge. This likelihood can be modified to account for heterogeneity in the per-exposure probability of infection or for immunity within a subset of the population (Longini and Halloran 1996, Regoes et al 2005, Hudgens and Gilbert 2009, Hudgens et al 2009). The likelihood modified for heterogeneity is the same as (4.11) except that p_z is replaced by an individual transmission probability for each macaque p_{iz} (often assumed to follow a beta distribution). The modified likelihood for immunity is:

$$L_{full} = \prod_{i=1}^{n} [(1-\theta)(1-p_z)^{X_i^{obs}-1} p_z]^{I[D_i^{obs}=1]} [\theta + (1-\theta)(1-p_z)^{X_i^{obs}}]^{I[D_i^{obs}=0]} \tag{4.15}$$

where θ is the probability that a macaque is immune (i.e., not susceptible to infection).

Likelihoods can also be constructed assuming a leaky effect with both heterogeneity and immunity, an all-or-none vaccine effect such that a macaque susceptible to disease when receiving control is no longer susceptible (i.e., immune) when receiving vaccine as well as a mixture vaccine effect where the vaccine provides both all-or-none and leaky effects. There is currently no default SAS or R procedure that assume these likelihoods; however, user-developed code that manually defines and optimizes the likelihood function can be constructed to obtain the corresponding LRT statistics and large sample p-values.

Since the results obtained from optimization packages available in both R and SAS may be sensitive to the observed data as well as specified optimization methods, parameter boundaries, and starting values, analyses employing this approach should explore various optimization options to assess the stability of the results. Conceptually this code can also be augmented in order to obtain randomization-based p-values by obtaining the LRT statistic based on the modified likelihood for all possible treatment assignment permutations. However, optimization options that consistently provide results for all treatment permutations are not immediately identifiable, and sensitivity of the study results to the specified options cannot easily be assessed for all treatment permutations. Accordingly, randomization-based p-values assuming these modified likelihoods are not further considered; however, the randomization-based p-values assuming the original partial or full likelihoods, (4.9) or (4.11) are still appropriate and valid even when data arise from the modified likelihood in that the type I error is still guaranteed to be less than α.

4.5.5 Power

The LRT, score and Wald test statistics all converge asymptotically to chi-square distributions with one degree of freedom but do not have equivalent power for all sample sizes and treatment effects (Sen 1993). As such, identifying the test with the greatest power is of interest. In the large sample setting, the Neyman-Pearson lemma states that $T = L_0/L_1 \geq \alpha$ where L_0 is the likelihood under the null and L_1 is the likelihood under the alternative is the most powerful test statistic for testing H_0 against a simple alternative assuming the likelihood employed in the test is the correct underlying likelihood (specifically reject H_0 if $LRT \geq k_\alpha$ where $P\{t \geq k_\alpha | H_0\} = P_{L_0}\{t \geq k_\alpha\} = \alpha$). A uniformly most powerful (UMP) test statistic is one that is most powerful against an entire set of possible alternatives.

As discussed the survival model for discrete time under non-informative censoring corresponds to the binomial likelihood. Unfortunately, there is no unconditional UMP test for the logistic distribution (Lehmann and Romano 2005). However, if the assumptions about the likelihood are correct the LRT will have best average power when the number of observations is large and therefore the LRT is the preferred test statistic (Wald 1943). Additionally, the Wald test statistic is not recommended in the binomial likelihood setting. Specifically there is nonmonotonicity in the power function for this test such that as the distance increases between the parameter estimate of treatment effect and its null value (i.e., settings where there appears to be large evidence of a treatment effect), the test statistic counter intuitively decreases to zero in certain settings (Hauck and Donner 1977). This characteristic is due to the test statistic's reliance on the estimated variance of the sufficient statistic under the null hypothesis. Specifically, for data resulting from a setting where there is a large treatment effect the variance estimate is large and subsequently the estimated test statistic is small thus resulting in reduced power for the Wald statistic in comparison to other test statistics in this setting (Ullah, Wan and Chaturvedi 2002). With respect to randomization-based tests (permutation tests based on the LRT, score or Wald statistics), the LRT-based permutation test will asymptotically remain the most powerful as the power of the LRT-based permutation test converges to the power of the parametric LRT as sample sizes goes to infinity (Hoeffding 1952).

4.6 Simulations

Simulations were conducted to compare the operating characteristics of the exact tests as well as to assess type I error rates for the large-sample p-values using varying values of N and C_{max}. RLC trials were simulated with macaques randomized 1:1 to either vaccine or a placebo control challenged repeatedly under three different sets of assumptions about rates of infection in the vaccine and placebo arms. In the first setting (Scenario 1), a leaky effects scenario was simulated where the probability of infection at each challenge for all macaques when receiving placebo was p_0 while the probability of infection when receiving vaccine was $p_1 = \phi p_0$. Simulations were completed by performing Bernoulli trials for each challenge with a p_0 or p_1 event probability depending on treatment assignment. The second setting (Scenario 2) simulates a leaky effect with heterogeneity in the per-exposure probability of infection where mean probability of infection at each challenge for all macaques was p_0 and $p_1 = \phi p_0$ when receiving placebo and vaccine respectively. Individual macaque probabilities of infection used in the Bernoulli trials were obtained using a beta distribution as detailed in Regoes et al. (2005) where $\mu = p_0$ for the placebo group, $\mu = p_1$ for the vaccine group and the coefficient of variation, $CV = \mu/\sigma = 0.5$ for both groups. The last setting (Scenario 3) simulates the leaky effects model from Setting 1 with a proportion of the population immune to infection ($\theta = 0.2$).

Figure 4.1 displays the empirical type I error and power of the randomization-based tests for various values of $\phi = .1, .2, \ldots, 1$ ($\phi = 1$ under the null) based on 2,000 simulations. In all settings, exact p-values based on the LRT (full) assume the leaky effects model with no heterogeneity or immunity. P-values for the exact Wald test and LRT were approximated using Monte Carlo methods with 4,000 samples. All exact tests have appropriate type I error rates even when the underlying likelihood for the simulated data does not correspond to the likelihood used as a basis for the test. The LRT (full) is the most powerful test in all scenarios; however, power is lower when the leaky effects model is not the true underlying likelihood of the simulated data. The Wald tests (both full and partial) consistently have the lowest power. The power for the log-rank/score test which is computationally easiest to obtain compared to the LRT is reduced by up to 0.05 in the first simulation setting with no heterogeneity or immunity and by up to 0.11 in the other settings. Larger values of C_{max} do not substantially reduce the difference in power between these tests; however, for larger N (i.e., $N \geq 20$ in the simulated settings) the power curves are approximately equal for the exact tests (results not shown).

Figure 2 displays the type I error of the large sample tests for various values of C_{max} and N based on 10,000 simulations. The type I error varies by test but tends to be inflated for most values of C_{max} and N. Note in this simulation scenario the inflation amount appears to increase with C_{max}; however, the inflation amount tends to be $< .02$ for N greater than or equal to 30.

4.7 Application

Hessell et al. (2009) presented analyses of a repeated low-dose challenge study with 4 macaques in the control group and 5 macaques in each of two vaccine groups (b12 and LALA). All groups were initially challenged with a very low-dose challenge (3TCID). After only one macaque was infected after 11 challenges (infected at the 6^{th} challenge), the challenge dose was slightly escalated (10TCID). In the published results of the study, the macaque that was infected while being challenged with 3TCID was treated as though they were infected at the first 10TCID challenge; otherwise the 3TCID challenges were ignored in the analyses. Of the 4 control macaques, 3 macaques were infected after 2 10TCID challenges and 1 macaque was infected after 4 10TCID challenges. Of the 5 b12 macaques, 4 macaques were infected after 1, 6, 23, and 38 challenges and 1 macaque remained uninfected after 40 challenges.

Hessell et al. declared a significant difference between these groups based on a Fisher's exact test (p-value=0.0016). As discussed in Section 4.4.2, the calculation for Fisher's exact test considers a set of 2×2 tables inconsistent with permuting treatment assignment and therefore may reject the null hypothesis too often. Based on the score test from exact conditional logistic regression (equivalent to an exact log-rank test), the null (4.2) is not rejected (p-value=0.1095) although there is suggestion of a trend. Exact p-values based on the LRT(full) and Wald test statistics (assuming no heterogeneity or immunity) are all also > 0.05.

4.8 Discussion

In RLC studies with small sample sizes, randomization-based inference should be employed in order to protect type I error. While randomization-based tests constructed using the LRT statistic tend to be the most powerful if the assumed underlying likelihood is correct, such tests are computationally difficult to employ as they require user-defined programming and must rely on Monte-Carlo sampling for even moderate sample sizes. In contrast, randomization-based tests constructed using the log-rank or score statistic which are equivalent for the discrete time to event setting, are easy to obtain in either R or SAS (R Development Core Team 2008, SAS Institute Inc. 2008). Specifically, p-values based on the exact log-rank test can be obtained using the SURVTEST function from the COIN package in R while p-values based on the exact score test can be obtained via exact conditional logistic regression using the LOGISTIC procedure with an EXACT option in SAS. The power of these tests is approximately comparable to the power of the randomization-based test using the LRT when there is no heterogeneity or immune fraction. An added benefit to the exact conditional logistic regression approach in comparison to exact log-rank test approaches is that covariates of potential interest can easily be incorporated into the model. Additionally, estimates of treatment effect can also be obtained from this approach.

4.9 Analysis Code

A. SAS Code

The code below if for version 9.1.3 or higher. Section 1 creates two datasets for the same RLC study (STRUCTURE1 has one observation per subject while STRUCTURE2 has one observation per subject/challenge). Section 2 of A provides code for calculating large sample p-values based on the log-rank test (PROC LIFETEST); time-stratified CMH test which is equivalent to the log-rank test (PROC FREQ); the Wald, score and LRT using the partial likelihood with no heterogeneity or immunity (PROC PHREG); as well as the the Wald and LRT using the full likelihood with no heterogeneity or immunity (PROC LOGISTIC). The LRT using the full likelihood is not included in the SAS output. Instead it is calculated as -2[LogL(full)-LogL(reduced)] where the reduced model is obtained via an additional PROC LOGISTIC call that only includes time. Section 3 of A provides code for calculating randomization-based p-values using built-in SAS procedures including the exact p-value based on the log-rank test statistic using Savage scores (PROC TWO SAMPLE, requires STATXACT PROCs SAS license) and the exact p-value based on the score test statistic obtained via exact conditional logistic regression (PROC LOGISTIC with EXACT statement). Section 4 of A provides a MACRO (with inputs of the number of MC samples and number of individuals randomized to active treatment) for calculating randomization-based p-values using user-defined code and Monte-Carlo approximation. The p-values calculated are based on the Wald, Score and LRT statistics using the partial likelihood while assuming no heterogeneity or immunity. A similar process can be used for obtaining the randomization-based p-values for the Wald and LRT statistics using the full likelihood while assuming no heterogeneity or immunity.

1. Data Structure

```
DATA STRUCTURE1;
INPUT ID TRT TIME INFECTION @@;
CARDS;
1 1 3 1   2 1 4 0   3 1 4 1   4 1 5 1   5 1 5 1
6 0 1 1   7 0 2 0   8 0 3 1   9 0 3 1   10 0 5 1;

DATA STRUCTURE2;
INPUT ID TRT TIME CHALLENGE_INFECTION @@;
CARDS;
1 1 1 0   1 1 2 0   1 1 3 1
2 1 1 0   2 1 2 0   2 1 3 0   2 1 4 0
3 1 1 0   3 1 2 0   3 1 3 0   3 1 4 1
4 1 1 0   4 1 2 0   4 1 3 0   4 1 4 0   4 1 5 1
5 1 1 0   5 1 2 0   5 1 3 0   5 1 4 0   5 1 5 1
6 0 1 1
7 0 1 0   7 0 2 0
8 0 1 0   8 0 2 0   8 0 3 1
9 0 1 0   9 0 2 0   9 0 3 1
10 0 1 0  10 0 2 0  10 0 3 0  10 0 4 0  10 0 5 1;
```

2. Large Sample P-values

```
PROC LIFETEST data=STRUCTURE1;
  time TIME*INFECTION(0); strata TRT;
```

```
PROC FREQ data=STRUCTURE2;
  tables TIME*TRT*CHALLENGE_INFECTION /cmh;


PROC PHREG data=STRUCTURE1;
  model TIME*INFECTION(0)=TRT/ ties=discrete;


PROC LOGISTIC data=STRUCTURE2 descending;
  class TRT TIME /param=ref; model CHALLENGE_INFECTION=TRT TIME;
```

3. Randomization-Based P-values (built-in procedures)

```
PROC TWOSAMPL data=STRUCTURE1;
  lo/ex; po TRT; re TIME; ce INFECTION;


PROC LOGISTIC data=STRUCTURE2 descending exactonly;
  class TRT TIME /param=ref; model CHALLENGE_INFECTION=TRT TIME;
  exact TRT /estimate=both;
```

4. Randomization-Based P-values (manual; monte-carlo approximation)

```
%MACRO RAND(SAMPLES=,NTRT=);
  *create dataset with set of observations for each permutation
  DATA STRUCTURE1_RANDT1 (drop=TRT); set STRUCTURE1;
    do PERM=1 to &SAMPLES; RANDORDER=rand("Uniform"); output; end;
  PROC SORT data=STRUCTURE1_RANDT1; by PERM RANDORDER;
  DATA STRUCTURE1_RAND; set STRUCTURE1_RANDT1;
    retain PERMID; by PERM;
    if first.PERM then PERMID=1; else PERMID=PERMID+1;
    if PERMID<=&ntrt then TRT=1; else TRT=0; run;
  *calculate test statistics for all permutations;
  PROC PHREG data=STRUCTURE1_RAND;
```

```
    ODS OUTPUT GLOBALTESTS=PHTESTSPERM;
    by PERM; model TIME*INFECTION(0)=TRT/ ties=discrete;
  *calculate test statistics for observed data;
  PROC PHREG data=STRUCTURE1;
    ODS OUTPUT GLOBALTESTS=PHTESTSMAIN (rename=(CHISQ=MAIN));
    model TIME*INFECTION(0)=TRT/ ties=discrete; run;
  *shell dataset to ensure 1 observation for each permutation;
  DATA SHELL;
    do PERM=1 to &SAMPLES;
      test="Likelihood Ratio"; output;
      test="Score"; output;
      test="Wald"; output;
    end;
  PROC SORT data=PHTESTSPERM; by TEST PERM;
  PROC SORT data=SHELL; by TEST PERM;
  DATA PHTESTSPERM; merge PHTESTSPERM SHELL; by TEST PERM;
  PROC SORT data=PHTESTSMAIN; by TEST;
  *calculate the probability of treatment assignment as or more
   extreme than that observed based on each test statistic;
  DATA PHTESTS; merge PHTESTSPERM PHTESTSMAIN; by TEST;
    if CHISQ>=MAIN then PVALUE=1; else PVALUE=0;
  PROC MEANS data=PHTESTS; by TEST; var PVALUE; run;
%MEND;
```

B. R Code

The code below is for version 2.12.0. Section 1 creates two vector sets for the same RLC study (TRT, TIME, INFECTION have one observation per subject while TRTCHALL, CHALLENGE and INFECHALL have one observation per subject/challenge). Section 2 of B provides code for calculating large sample p-values based on the log-rank test (SURV); time-stratified CMH test (CMH_TEST); the Wald, score and LRT using the partial likelihood with no heterogeneity or immunity (COXPH); as well as the the Wald and LRT using the full likelihood with no heterogeneity or immunity (GLM). As with the SAS code, the LRT using the full likelihood is not included in the output and must be calculated by constructing a reduced model. Section 3 of B provides code for calculating randomization-based p-values using built-in procedures including the exact p-value based on the log-rank test statistic using conditional variance (SURV with exact statement). Section 4 of B provides code (with input of the number of MC samples) for calculating randomization-based p-values using user-defined code and Monte-Carlo approximation. The p-values calculated are based on the Wald, Score and LRT statistics using the partial likelihood while assuming no heterogeneity or immunity. A similar process can be used for obtaining the randomization-based p-values for the Wald and LRT statistics using the full likelihood while assuming no heterogeneity or immunity.

1. Data Structure

```
TRT<-c(rep(1,5),rep(0,5))
TIME<-c(3,4,4,5,5, 1,2,3,3,5)
INFECTION<-c(1,0,1,1,1,1,0,1,1,1)

TRTCHALL<-c(rep(1,21),rep(0,14))
CHALLENGE<-c(1,2,3, 1,2,3,4, 1,2,3,4, 1,2,3,4,5, 1,2,3,4,5,
             1, 1,2, 1,2,3, 1,2,3, 1,2,3,4,5)
INFECHALL<-c(0,0,1, 0,0,0,0, 0,0,0,1, 0,0,0,0,1, 0,0,0,0,1,
             1, 0,0, 0,0,1, 0,0,1, 0,0,0,0,1)
```

2. Large Sample P-values

```
survdiff(Surv(TIME,INFECTION)~TRT)
```

```
CMHTEST<-array(c(0,5,1,4, 0,5,0,4, 1,4,2,1, 1,3,0,1, 2,0,1,0),
  dim=c(2,2,5), dimnames=list(Treatment=c("Vacc","Cont"),
  Response=c("Infected","Not"),Time=c("1","2","3","4","5")))
CMHTABLE <- as.table(CMHTEST)
cmh_test(CMHTABLE)


PHTEST<-coxph(formula=Surv(TIME, INFECTION)˜TRT, method="exact")
summary(PHTEST)


LOGFULL<-glm(INFECHALL˜TRTCHALL+strata(CHALLENGE),family=binomial)
summary(LOGFULL)
```

3. Randomization-Based P-values (built-in procedures)

```
TRTC <-c(rep(’A’,5),rep(’B’,5))
SURVDATA<-data.frame(TIME,TRTC,INFECTION)
surv_test(Surv(TIME, INFECTION)˜TRTC,data=SURVDATA,
          distribution="exact")
```

4. Randomization-Based P-values (manually computed)

```
*calculate test statistic for observed data
PHTEST<-coxph(formula=Surv(TIME,INFECTION)˜TRT,method="exact")
MAINLRT<--2*(PHTEST$loglik[1]-PHTEST$loglik[2])
MAINWALD<-PHTEST$wald.test
MAINSCORE<-PHTEST$score
MCrejects<-matrix(0,SAMPLES,3)
for (KK in 1:SAMPLES){
  *permute treatment assignment
  TRT<-sample(TRT)
  *calculate test statistics for permutation
  PHTEST<-coxph(formula=Surv(TIME,INFECTION)~TRT,method="exact")
  PERMLRT<--2*(PHTEST$loglik[1]-PHTEST$loglik[2])
  PERMWALD<-PHTEST$wald.test
  PERMSCORE<-PHTEST$score
  if (PERMLRT>MAINLRT) MCrejects[KK,1]<-1
  if (PERMWALD>MAINWALD) MCrejects[KK,2]<-1
  if (PERMSCORE>MAINSCORE) MCrejects[KK,3]<-1
 }
LRTPVALUE<-mean(MCrejects[,1])
WALDPVALUE<-mean(MCrejects[,2])
SCOREPVALUE<-mean(MCrejects[,3])
```

Figure 4.1: Empirical type 1 error and power for $\alpha = 0.05$ significance level, $p_0 = 0.5$ probability of infection, $N = 10$ macaques, and $C_{max} = 12$ maximum challenges per macaque, where Scenario 1 includes no heterogeneity or immunity, Scenario 2 includes heterogeneity, and Scenario 3 includes immunity (LRT=solid line [full=black, partial=gray], Score/Log-rank=dot/dashed line, Wald=dashed line [full=black, partial=gray])

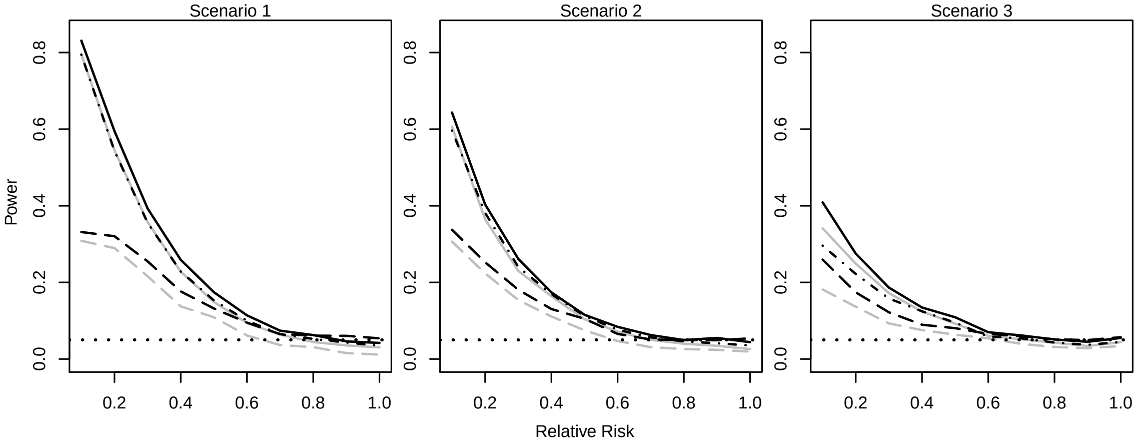
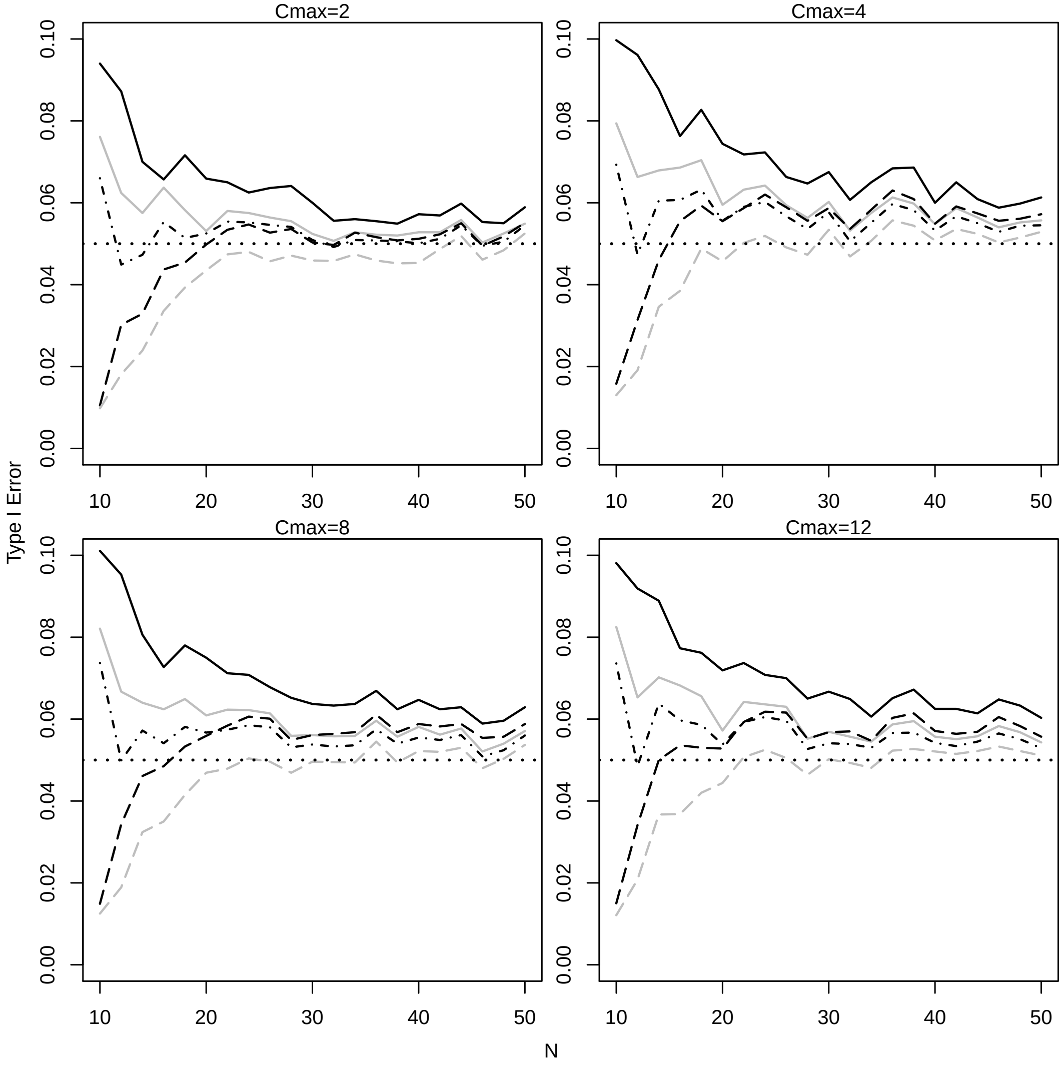


Figure 4.2: Empirical type 1 error for $\alpha = 0.05$ significance level, $p_0 = p_1 = 0.5$ probability of infection under Scenario 1 (no heterogeneity or immunity) for various values of N and C_{max} (LRT=solid line [full=black, partial=gray], Score/Log-rank=dot/dashed line, Wald=dashed line [full=black, partial=gray])



Chapter 5

Conclusion

Approximately 2.6 million new cases of human immunodeficiency virus (HIV) infection occur each year (Dieffenbach and Fauci 2011). Accordingly a primary goal of future research is to identify interventions that will successfully prevent future infections. In research for preventing infectious diseases including HIV, randomized studies are often employed as a means for assessing the effectiveness of an intervention in preventing infection as well as in assessing the effect of the intervention on secondary outcomes such as a post-infection outcomes of death or disease severity. Unfortunately randomized studies are sometimes complicated by small sample sizes or the presence of intermediate post-randomization events. In such settings, researchers may employ inappropriate analytic methods for the data available and therefore the results reported may be invalid. Examples of invalid analytic approaches used in infectious disease prevention research include making treatment comparisons conditional on post-randomization events using standard analytic methods as well as using analytic methods for repeated low-dose challenge studies requiring assumptions inconsistent with the study data.

With respect to presence of intermediate post-randomization events, which is the focus of Chapters 2 and 3, the biggest concern with researchers employing standard analytic approaches is that findings of a significant treatment effect on an outcome of interest may be unfounded. Specifically, the observed data may not support such a finding when the potential selection bias is taken into account. The publication of significant findings based on invalid approaches incorrectly informs decisions regarding continued research or future use of potential preventive therapies. Principal stratification approaches have been developed and proposed for addressing potential selection bias due to intermediate post-randomization events. Unfortunately, approaches that are valid in small sample settings were previously not available and in general, use of principal stratification approaches is extremely limited in published infectious disease prevention research.

Accordingly, in this dissertation, we first addressed these issues by developing methods for exact randomization-based causal inference within principal strata in the presence of selection bias due to the presence of post-randomization events. This development work also included augmenting the proposed test via adjusting for baseline covariates to increase power as well as the development of exact CIs for the principal stratum direct effect. Secondly, we presented a broader discussion of the use of principal stratification analytic approaches for handling selection bias in randomized studies. This discussion is targeted at subject matter experts and includes an overview of the problem with using standard analytic approaches in the presence of potential selection bias, describes how principal stratification approaches are generally employed and also compares the results obtained from principal stratification approaches to published results based on standard analytic approaches that ignore the potential selection bias.

The major goal of the research associated with these chapters is to better inform analyses conducted on outcomes subject to selection bias due to intermediate post-randomization events. The introduction of the principal stratum exact test (PSET) provides researchers with a valid analytic approach using the principal stratification framework for small sample settings while the general discussion of principal stratification introduces clinical researchers to a potential analytic option for addressing this type of selection bias. One potential limitation of the PSET is that the test may be overly conservative as the type I error rate is exceptionally low. Areas for future research regarding randomization-based inference within principal strata include pursuing other avenues for deriving an exact test that results in a more powerful test, relaxing the assumptions of monotonicity or independent treatment assignment for the PSET, and relaxing the additivity assumption with respect to constructing a CI for the principal stratum direct effect.

Repeated low-dose challenge (RLC) studies, the focus of Chapter 4, were developed such that preclinical randomized studies of potential vaccines more accurately mirrored 'real-life' human transmission of infection and therefore, vaccines identified for future research would be more likely to be effective when applied to the human population. However, the application of invalid analytic approaches such as methods that rely on assumptions about asymptotic distributions of test statistic that may be inappropriate in the small sample setting or assumptions that are not met by the data collected in RLC studies may lead to inaccurate study results and therefore remove all benefits of conducting a RLC study. Again, significant findings based on invalid approaches in this setting will also incorrectly inform decisions regarding continued research of potential preventive therapies.

Therefore, we presented a discussion of appropriate analytic approaches for RLC preclinical vaccine studies that included a comparison of the operating characteristics of these approaches as well as detail on how to employ these approaches using standard statistical software. A possible avenue of future research is determining a computationally feasible route for constructing an exact, randomization-based test of vaccine effect for the more complex models such as a leaky effects model assuming heterogeneity or an all-or-none or mixture vaccine effects model. Additionally, the current research presented in Chapter 4 is primarily intended for an audience with a statistical background. To increase general knowledge of appropriate analytic approaches for RLC studies, it would be helpful to present an abbreviated version of this paper to a journal geared towards clinical investigators.

Bibliography

P. D. Allison. Discrete-time methods for the analysis of event histories. Sociological Methodology, 13:61–98, 1982.

P. D. Allison. Survival Analysis Using SAS: A Practical Guide, Second Edition. SAS Press, Cary, NC, 2010.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables (disc: P456-472). J Am Stat Assoc, 91:444–455, 1996.

S. G. Baker, Y. Wax, and P. H. Patterson. Regression analysis of grouped survival data: informative censoring and double sampling. Biometrics, 49:379–389, 1993.

S. G. Baker, C. Frangakis, and K. S. Lindeman. Estimating efficacy in a proposed randomized trial with initial and later non-compliance. J R Stat Soc Ser C Appl Stat, 56:211–221, 2007.

G. A. Barnard. Significance tests for 2 × 2 tables. Biometrika, 34:123–138, 1947. ISSN 0006-3444.

M. J. Bayarri and J. O. Berger. *P* values for composite null models. J Am Stat Assoc, 95 (452):1127–1142, 2000.

R. L. Berger and D. D. Boos. *P* values maximized over a confidence set for the nuisance parameter. J Am Stat Assoc, 89:1012–1016, 1994.

M. Boily, R. F. Baggaley, L. Wang, B. Masse, R. G. White, R. J. Hayes, and M. Alary. Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. Lancet Infectious Diseases, 9:118–129, 2009.

C. C. Brown. On the use of indicator variables for studying the time-dependence of parameters in a response-time model. Biometrics, 31:863–872, 1975.

S. P. Buchbinder, D. V. Mehrotra, A. Duerr, D. W. Fitzgerald, R. Mogg, D. Li, P. B. Gilbert, J. R. Lama, M. Marmor, C. Del Rio, M. J. McElrath, D. R. Casimiro, K. M. Gottesdiener, J. A. Chodakewitz, L. Corey, M. N. Robertson, and Step Study Protocol Team. Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. Lancet, 372:1881–1893, Nov 2008.

H. Callaert. Comparing statistical software packages: the case of the logrank test in StatXact. American Statistician, 57:214–217, 2003.

G. Casella and R. L. Berger. Statistical Inference. Duxbury Press, Belmont, CA, 2002.

M. N. Chang, H. A. Guess, and J. F. Heyse. Reduction in burden of illness: A new efficacy measure for prevention trials. Stat Med, 13:1807–1814, 1994.

C. Chasela, M. G. Hudgens, D. J. Jamieson, D. Kayira, M. Hosseinipour, A. P. Kourtis, R. Knight, Y. Ahmed, D. Kamwendo, I. Hoffman, S. Ellington, J. Wiener, S. A. Fiscus, I. Mofolo, D. Sichali, and C. van der Horst for the BAN Study team. Maternal antiretrovirals or infant nevirapine to reduce HIV-1 transmission. New Engl J Med, 362:2271–2281, 2010.

D. R. Cox. Analysis of Binary Data. Chapman and Hall, London, 1970.

D. R. Cox. Regression models and life-tables. Journal of the Royal Statistical Society. Series B, 34:187–220, 1972.

Cytel Software Corporation. StatXact 8.0 For Windows User Manual. Cytel Software Corporation, Cambridge, MA, 2007.

C. W. Dieffenbach and A. S. Fauci. Thirty years of HIV and AIDS: Future challenges and opportunities. Ann Intern Med., 154:766–771, 2011.

D. W. Fitzgerald, H. Janes, M. Robertson, R. Coombs, L. Frank, P. Gilbert, M. Loufty, D. Mehrotra, and A. Duerr for the Step Study Protocol Team. An Ad5-vectored HIV-1 vaccine elicits cell-mediated immunity but does not affect disease progression in HIV-1-infected male subjects: results from a randomized placebo-controlled trial (the Step study). JID, 203:765–772, 2011.

D. Follmann, M. P. Fay, and M. Proschan. Chop-lump tests for vaccine trials. Biometrics, 65:885–893, 2009.

E. M. Foster. Causal inference and developmental psychology. Developmental Psychology, 46:1454–1480, 2010.

C. E. Frangakis and D. B. Rubin. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika, 86:365–379, 1999.

C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. Biometrics, 58: 21–29, 2002.

J.G. García-Lerma, R.A. Otten, S.H. Qari, E. Jackson, M.E. Cong, S. Masciotra, W. Luo, C. Kim, D.R. Adams, M. Monsour, J. Lipscomb, J.A. Johnson, D. Delinsky, R.F. Schinazi, R. Janssen, T.M. Folks, and W. Heneine. Prevention of rectal SHIV transmission in macaques by daily or intermittent prophylaxis with emtricitabine and tenofovir. PLoS Med., 5:e28, 2008.

P. B. Gilbert and M. G. Hudgens. Evaluating candidate principal surrogate endpoints. Biometrics, 64:1146–1154, 2008.

P. B. Gilbert, R. Bosch, and M. G. Hudgens. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. Biometrics, 59:531–541, 2003.

P. B. Gilbert, J. O. Berger, D. Stablein, S. Becker, M. Essex, S. M. Hammer, J. H. Kim, and V. G. DeGruttola. Statistical interpretation of the RV144 HIV vaccine efficacy trial in Thailand: A case study for statistical issues in efficacy trials. JID, 203:969–975, 2011.

R. M. Grant, J. R. Lama, P. L. Anderson, V. McMahan, A. Y. Liu, L. Vargas, P. Goicochea, M. Casapia, J. V. Guanira-Carranza, M. E. Ramirez-Cardich, O. Montoya-Herrera, T. Fernandex, V. G. Veloso, S. P. Buchbinder, S. Chariyalertsak, M. Schechter, L. G Bekker, K. H. Mayer, E. G. Kallas, K. R. Amico, K. Mulligan, L. R. Bushman, R. J. Hance, C. Ganoza, P. Defechereux, B. Postle, F. Wang, J. J. McConnell, J. H. Zheng, J. Lee, J. F. Rooney, A. I. Jaffe, H. S. anad Martinez, D. N. Burns, and D. V. Glidden for the iPrEx study team. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. N Engl J Med, 363:2587–2599, 2010.

R. H. Gray, M. J. Wawer, and R. Brookmeyer et al. Probability of HIV-1 transmission per coital act in monogamous, heterosexual HIV-1-discordant couples in Rakai, Uganda. Lancet, 357:1149–1153, 2001.

R. H. Gray, D. Serwadda, X. Kong, F. Makumbi, G. Kigozi, P. E. Gravitt, S. Watya, F. Nalugoda, V. Ssempijja, A. A. R. Tobian, N. Kiwanuka, L. H. Moulton, N. K. Sewankambo, S. J. Reynolds, T. C. Quinn, B. Iga, O. Laeyendecker, A. E. Oliver, and M. J. Wawer. Male circumcision decreases acquisition and increases clearanace of high-risk human papillomavirus in HIV-negative men: a randomized trial in Rakai, Uganda. JID, 201:1455–1462, 2010.

S. Greenland. Causal inference in the health sciences. J Am Statist Assoc, 95:286–289, 2000.

B. M. Greenwood. What can be expected from malaria vaccines? Annals of Tropical Medicine and Parasitology, 91:S9–S13, 1997.

M. E. Halloran and C. J. Struchiner. Causal inference in infectious diseases. Epidemiology, 6:142–151, 1995.

B. B. Hansen and J. Bowers. Attributing effects to a cluster-randomized Get-Out-The-Vote campaign. J Am Stat Assoc, 104:873–885, 2009.

W. W. Hauck and A. Donner. Wald's test as applied to hypotheses in logit analysis. JASA, 72:851–853, 1977.

J. J. Heckman and E. J. Vytlacil. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proc Natl Acad Sci, 96:4730–4734, 1999.

M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A structural approach to selection bias. Epidemiology, 15:615–625, Sep 2004.

A. J. Hessell, P. Poignard, M. Hunter, L. Hangartner, D. M. Tehrani1, W. K. Bleeker, P. W. H. I. Parren, P. A. Marx, and D. R. Burton. Effective, low-titer antibody protection against low-dose repeated mucosal SHIV challenge in macaques. Nature Medicine, 15: 951–954, 2009.

K. F. Hirji, C. R. Mehta, and N. R. Patel. Computing distributions for exact logistic regression. JASA, 82:1110–1117, 1987.

W. Hoeffding. The large-sample power of tests based on permutations of observations. The Annals of Mathematical Statistics, 23:169–192, 1952.

P. W. Holland. Statistics and causal inference (C/R: p961-970). J Am Stat Assoc, 81:945–960, 1986.

T. Hothorn and B. Lausen. On the exact distribution of maximally selected rank statistics. Comp Statist Data Analysis, 43:121–137, 2003.

M. G. Hudgens and P. B. Gilbert. Assessing vaccine effects in repeated low-dose challenge experiments. Biometrics, 65:1223–1232, 2009.

M. G. Hudgens and M. E. Halloran. Causal vaccine effects on binary post-infection outcomes. J Am Stat Assoc, 101:51–64, 2006.

M. G. Hudgens, A. Hoering, and S. G. Self. On the analysis of viral load endpoints in HIV vaccine trials. Stat Med, 22:2281–2298, 2003.

M. G. Hudgens, P. B. Gilbert, J. R. Mascola, C. Wu, D. H. Barouch, and S. G. Self. Power to detect the effects of HIV vaccination in repeated low-dose challenge experiments. JID, 200:609–613, 2009.

K. Imai. Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". Stat Probab Lett, 78:144–149, 2008.

G. W. Imbens and P. R. Rosenbaum. Robust, accurate confidence intervals with a weak instrument, quarter of birth and education. J R Stat Soc Ser A General, 25:305–327, 2005.

Y. Jemiai, A. Rotnitzky, B. E. Shepherd, and P. B. Gilbert. Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs. J R Stat Soc Series B Stat Methodol, 69:879 – 901, 2007.

H. Jin and D. B. Rubin. Public schools versus private schools: causal inference with partial compliance. J Educ Behav Stat, 34:24–45, 2009.

B. Jo. Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance. Stat Med, 21:3161–3181, 2002.

M. M. Joffe. Principal stratification and attribution prohibition: good ideas taken too far. Int J Biostat, 7, 2011.

M. M. Joffe and T. Greene. Related causal frameworks for surrogate outcomes. Biometrics, 65:530–538, 2009.

J. D. Kalbfleisch and R. L. Prentice. The Statistical Analysis of Failure Time Data. Wiley, New York, 1980.

J. D. Kalbfleisch and R. L. Prentice. The Statistical Analysis of Failure Time Data, Second Edition. Wiley, New York, 2002.

G. G. Koch, D. B. Gillings, and M. E. Stokes. Biostatistical implications of design, sampling, and measurement to health science data analysis. Ann Rev Public Health, 1:163–225, 1980.

W. C. Koff, P. R. Johnson, D. I. Watkins, D. R. Burton, J. D. Lifson, K. J. Hasenkrug, A. B. McDermott, A. Schultz, T. J. Zamb, R. Boyle, and R. C. Desrosiers. HIV vaccine design: insights from live attenuated SIV vaccines. *Nature Immunology*, 7:19–23, 2006.

L. Kuhn, G. M. Aldrovandi, M. Sinkala, C. Kankasa, K. Semrau, M. Mwiya, P. Kasonde, N. Scott, C. Vwalika, J. Walter, M. Bulterys, W. Y. Tsai, D. M. Thea, E. Abrams, T. Colton, W. Fawzi, S. Kapiga, E. Chomba, S. Allen, C. Luo, L. Mofenson, E. Piwoz, K. Ryan, J. Simon, Z. Stein, J. Stringer, and S. Vermund. Effects of early, abrupt weaning on HIV-free survival of children in Zambia. *N Engl J Med*, 359:130–141, Jul 2008.

J. M. Lachin. Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials*, 21:167–189, 2000.

E. L. Lehmann. *Testing Statistical Hypotheses*. New York: Wiley, 1959.

E. L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Prentice Hall, New Jersey, 1998.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. New York: Springer, 2005.

R J Little and D B Rubin. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health*, 21:121–145, 2000.

R. J. Little, Q. Long, and X. Lin. A comparison of methods for estimating the causal effect of treatment in randomized clinical trials subject to noncompliance. *Biometrics*, 65:640–649, 2009.

I. M. Longini and M. E. Halloran. A frailty mixture model for estimating vaccine efficacy. *Applied Statistics*, 45:165–173, 1996.

N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50:163–170, 1966.

A. B. McDermott, J. Mitchen, S. Piaskowski, I. De Souza, L. J. Yant, J. Stephany, J. Furlott, and D. I. Watkins. Repeated low-dose mucosal simian immunodeficiency virus SIVmac239 challenge results in the same viral and immunological kinetics as high-dose challenge: a model for the evaluation of vaccine efficacy in nonhuman primates. *Journal of Virology*, 78:3140–3144, 2004.

D. V. Mehrotra, X. Li, and P. B. Gilbert. A comparison of eight methods for the dual-endpoint evaluation of efficacy in a proof-of-concept HIV vaccine trial. *Biometrics*, 62:893–900, 2006.

C. Mehta and N. Patel. *StatXact 8 User Manual*. Cytel, Cambridge, MA, 2007.

G. Mick. Vaccination: A new option to reduce the burden of herpes zoster. *Expert Rev Vaccines*, 9(3 Suppl):31–35, 2010.

P. A. Mock, N. Shaffer, and C. Bhadrakom et al. Maternal viral load and timing of mother-to-child HIV transmission, Bangkok, Thailand. *AIDS*, 13:407–414, 1999.

L. M. Mofenson. Prevention of breast milk transmission of HIV: The time is now. J. Acquir. Immune Defic. Syndr., 52:305–308, 2009.

L. M. Mofenson. Prevention in neglected subpopulations: Prevention of mother-to-child transmission of HIV infection. CID, 50(S3):130–148, 2010.

J. Neyman. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. (In Polish) *Roczniki Nauk Roiniczych, Tom X,* pp. 1-51. Reprinted in *Statist Sci* 1990, 5, 463-480, with discussion by T. Speed and D. Rubin. 1923.

U. M. Parikh, C. Dobard, S. Sharma, M. Cong, H. Jia, A. Martin, C. Pau, D. L. Hanson, P. Guenthner, J. Smith, E. Kersh, J. G. García-Lerma, F. J. Novembre, R. Otten, T. Folks, and W. Heneine1. Complete protection from repeated vaginal simian-human immunodeficiency virus exposures in macaques by a topical gel containing tenofovir alone or with emtricitabine. Journal of Virology, 83:1035810365, 2009.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

R. R. Regoes, I. M. Longini, M. B. Feinberg, and S. I. Staprans. Preclinical assessment of HIV vaccines and microbicides by repeated low-dose virus challenges. PLoS Medicine, 2(8): e249, 2005.

M. R. Reynolds, A. M. Weiler, S. M. Piaskowski, H. L. Kolar, A. J. Hessell, M. Weiker, K. L. Weisgrau, E. J. León, W. E. Rogers, R. Makowsky, A. B. McDermott, R. Boyle, N. A. Wilson, D. B. Allison, D. R. Burton, W. C. Koff, and D. I. Watkins. Macaques vaccinated with simian immunodeficiency virus SIVmac239δnef delay acquisition and control replication after repeated low-dose heterologous SIV challenge. Journal of Virology, 84:91909199, 2010.

W. N. Rida, P. Fast, R. Hoff, and T. R. Fleming. Intermediate-sized trials for the evaluation of HIV vaccine candidates: a workshop summary. J Acquir Immune Defic Syndr Hum Retrovirol, 16:195–203, 1997.

J. M. Robins. An analytic method for randomized trials with informative censoring: Part I. Lifetime Data Anal, 1:241–254, 1995.

J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. Epidemiology, 3:143–155, 1992.

J. M. Robins, A. van der Vaart, and V. Ventura. The asymptotic distribution of p-values in composite null models. J Am Stat Assoc, 95:1143–1156, 2000.

P. R. Rosenbaum. The consequences of adjustment for a concomitant variable that has been affected by the treatment. J R Stat Soc Ser A General, 147:656–666, 1984.

P. R. Rosenbaum. Conditional permutation tests and the propensity score in observational studies. J Am Stat Assoc, 79:565–574, 1984b.

P. R. Rosenbaum. Identification of causal effects using instrumental variables: comment. JASA, 91:465–468, 1996.

P. R. Rosenbaum. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. Biometrika, 88:219–231, 2001.

P. R. Rosenbaum. Observational Studies. New York: Springer-Verlag, 2002a.

P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. Statist Sci, 17:286–327, 2002b.

P. R. Rosenbaum. Design of observational studies. New York: Springer, 2010.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70:41–55, 1983.

J. Roy, J. W. Hogan, and B. H. Marcus. Principal stratification with predictors of compliance for randomized trials with 2 active treatments. Biostatistics, 9:277–289, 2008.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66:688–701, 1974.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. The Annals of Statistics, 6:34–58, 1978.

D. B. Rubin. Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. J Am Stat Assoc, 75:591–593, 1980.

D. B. Rubin. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics, 47:1213–1234, 1991.

D. B. Rubin. Comment on "Causal inference without counterfactuals". J Am Stat Assoc, 95: 435–437, 2000.

D. B. Rubin. Causal inference using potential outcomes: design, modeling, decisions. J Am Stat Assoc, 100:322–331, 2005.

D. B. Rubin. Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death. Statist Sci, 3:299–309, 2006.

D. B. Rubin. For objective causal inference, design trumps analysis. Ann Appl Stat, 2: 808–840, 2008.

SAS Institute Inc. SAS/STAT 9.2 User Guide. Cary, NC, 2008.

D. O. Scharfstein, M. E. Halloran, H. Chu, and M. J. Daniels. On estimation of vaccine efficacy using validation samples with selection bias. Biostatistics, 7:615–629, Oct 2006.

P. K. Sen. Large sample methods in statistics: an introduction with applications. Chapman & Hall, 1993.

R. L. Shapiro, M. D. Hughes, A. Ogwu, and et al. Antiretroviral regimens in pregnancy and breast-feeding in Botswana. N Engl J Med, 362:2282–2294, 2010.

D. J. Shedlock, G. Silvestri, and D. B. Weiner. Monkeying around with HIV vaccines: using rhesus macaques to define gatekeepers for clinical trials. Nature, 9:717–728, 2009.

B. Shepherd, P. B. Gilbert, Y. Jemiai, and A. Rotnitzky. Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. Biometrics, 62(2):332–342, 2006.

B. E. Shepherd, P. B. Gilbert, and T. Lumley. Sensitivity analyses comparing time-to-event outcomes existing only in a subset selected postrandomization. J Am Stat Assoc, 102(478): 573–582, 2007a.

B. E. Shepherd, P. B. Gilbert, and D. V. Mehrotra. Eliciting a counterfactual sensitivity parameter. Am Stat, 61(1):56–63, 2007b.

J. D. Singer and J. B. Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. Journal of Educational Statistics, 18:155–195, 1993.

S. M. Smith. HIV vaccine development in the nonhuman primate model of AIDS. J Biomed Sci, 9:100–111, 2009.

M. Sobel. Causal inference in the social sciences. J Am Stat Assoc, 95:647–651, 2000.

B. E. Storer and C. Kim. Exact properties of some exact test statistics for comparing two binomial proportions. J Am Stat Assoc, 85:146–155, 1990.

S. I. Straprans, M. B. Feinberg, J. W. Shiver, and D. R. Casimiro. Role of nonhuman primates in the evaluation of candidates AIDS vaccines: an industry perspective. Current Opinion in HIV and AIDS, 5:377–385, 2010.

The Petra Study Team. Efficacy of three short-course regimens of zidovudine and lamivudine in preventing early and late transmission of hiv-1 from mother to child in Tanzania, South Africa, and Uganda (Petra study): a randomized, double-blind, placebo-controlled trial. Lancet, 353:1178–1186, 1999.

T. R. Ten Have, M. Joffe, and M. Cary. Causal logistic models for non-compliance under randomized treatment with univariate binary response. Stat Med, 22:1255–1283, 2003.

I. Thior, S. Lockman, L. M. Smeaton, R. L. Shapiro, C. Wester, S. J. Heymann, P. B. Gilbert, L. Stevens, T. Peter, S. Kim, E. van Widenfelt, C. Moffat, P. Ndase, P. Arimi, P. Kebaabetswe, P. Maxonde, J. Makhema, K. McIntosh, V. Novitsky, T. H. Lee, R. Marlink, S. Lagakos, and M. Essex for the Mashi study team. Breastfeeding plus infant zidovudine prophylaxis for 6 months vs formula feeding plus infant zidovudine for 1 month to reduced mother-to-child HIV transmission in Botswana. JAMA, 296:794–805, 2006.

S. K. Thompson. Sampling. New York: John Wiley & Sons, 2002.

S.M. Tirado and K.J. Yoon. Antibody-dependent enhancement of virus infection and disease. Viral Immunol., 16:69–86, 2003.

A. Ullah, A. Wan, and A. Chaturvedi. Handbook of Applied Econometrics And Statistical Inference. CRC Press, New York, 2002.

T. J. VanderWeele. Simple relations between principal stratification and direct and indirect effects. Stat Probab Lett, 78:2957–2962, 2008.

A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society, 54:426–482, 1943.

WHO/UNAIDS/IAVI International Expert Group. Executive summary and recommendations from the WHO/UNAIDS/IAVI expert group consultation on 'Phase IIB-TOC trials as a novel strategy for evaluation of preventive HIV vaccines', 31 January-2 February 2006, IAVI, New York, USA. AIDS, 21:539–546, 2007.

C. Winship and S. L. Morgan. Estimation of causal effects from observational data. Ann Rev Sociol, 25:659–706, 1999.

J. L. Zhang and D. B. Rubin. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". J Educ Behav Stat, 28:353–368, 2003.