University of North Carolina at Chapel Hill

The University of North Carolina at Chapel Hill Department of
Biostatistics Technical Report Series

Year 2010

Paper 17

Randomization-Based Inference within Principal Strata

Tracy L. Nolen*

Michael Hudgens[†]

*RTI International, Research Triangle Park, NC, tnolen@rti.org

[†]University of North Carolina at Chapel Hill, mhudgens@bios.unc.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/uncbiostat/art17>

Copyright ©2010 by the authors.

Randomization-Based Inference within Principal Strata

Tracy L. Nolen and Michael Hudgens

Abstract

In randomized studies, treatment comparisons conditional on intermediate post-randomization outcomes using standard analytic methods do not have a causal interpretation. An alternate approach entails treatment comparisons within principal strata by the potential outcomes for the intermediate outcome that would be observed under each treatment assignment. In this paper, we develop methods for randomization-based inference within principal strata. The proposed methods are compared with existing large-sample methods as well as traditional intent-to-treat approaches. This research is motivated by HIV prevention studies where few infections are expected and inference is desired within the always-infected principal stratum, i.e., all individuals who would become infected regardless of randomization assignment.

Randomization-Based Inference within Principal Strata

Tracy L. Nolen^{1,2} and Michael G. Hudgens^{2*}

¹ RTI International, Research Triangle Park, NC 27709

² Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599

* email: mhudgens@bios.unc.edu phone: 919.966.7253 fax: 919.966.3804

May 31, 2010

Tracy L. Nolen is Research Statistician, RTI International, Research Triangle Park, NC 27709 (email: tnolen@rti.org). Michael G. Hudgens is Research Associate Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599 (email: mhudgens@bios.unc.edu). TLN and MGH were partially support by NIH grant 1 R01 AI085073-01. The authors thank the BAN investigators for access to data from their study.

Abstract

In randomized studies, treatment comparisons conditional on intermediate post-randomization outcomes using standard analytic methods do not have a causal interpretation. An alternate approach entails treatment comparisons within principal strata defined by the potential outcomes for the intermediate outcome that would be observed under each treatment assignment. In this paper, we develop methods for randomization-based inference within principal strata. The proposed methods are compared with existing large-sample methods as well as traditional intent-to-treat approaches. This research is motivated by HIV prevention studies where few infections are expected and inference is desired within the always-infected principal stratum, i.e., all individuals who would become infected regardless of randomization assignment.

Keywords: causal inference, covariate adjustment, exact test, randomization

1 Introduction

1.1 Principal Stratification

Often 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 pre-treatment 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 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 standard comparisons between infected treated individuals and infected controls are subject to selection bias (Halloran and Struchiner 1995; Hernán, Hernández-Daz and Robins 2004). Moreover, because the set of individuals who would become infected if treated is likely not identical to the set of those who would become infected if given control, 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, Jemai and Rotnitzky 2006; Jemai, Rotnitzky, Shepherd and Gilbert 2007; 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; 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).

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, e.g., a person who is infected when treated would also become infected if not treated. Additional assumptions are needed in order to completely identify principal strata membership in the control group. 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.

1.2 Randomization-Based Inference

Randomized studies are the gold standard in clinical trials for evaluating the effects of biomedical interventions because randomization (*i*) allows for comparable groups with respect to measured and unmeasured covariates and (*ii*) provides a basis for statistical inference. 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). 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. In theory, principal stratification can be used in this setting, but to date no randomization-based inference methods have been developed for inference within principal strata.

An additional 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 not be appropriate. In the HIV vaccine setting, small to intermediate-sized 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) described 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 based inference may not be ideal if investigators are hesitant to make assumptions regarding prior distributions. Additionally, the use of large sample based frequentist inference in such settings may lead to incorrect inferences. For example, simulation studies of such methods have demonstrated inflated type I error of bootstrap based tests and under-coverage of bootstrap and Wald based confidence intervals (CIs) when the number of intermediate post-randomization events is small (Hudgens et al. 2003; Gilbert et al. 2003; Shepherd et al. 2007).

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. Section 2 describes the construction of the principal stratum exact test (PSET). Section 3 presents simulation results comparing the PSET to a large sample frequentist approach for testing for a treatment effect within principal strata. In Section 4 the PSET is applied to two studies on mother to child transmission of HIV. Section 5 compares the PSET to a traditional intent-to-treat (ITT) approach. Section 6 describes an extension of the PSET to allow for adjustments for covariates. Lastly, in Section 7 exact CIs for treatment effect are derived by inverting the PSET.

2 Principal Stratum Exact Test

2.1 Preliminaries

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 no hidden versions of treatment.

Under SUTVA, let $s_i(z)$ denote the value for 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 post-randomization outcome is binary. For ease of presentation, assume for the remaining discussion of notation that 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 as those individuals with $s_i(0) = 1, s_i(1) = 0$; and the immune (never-infected) stratum as those individuals with $s_i(0) = s_i(1) = 0$.

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 $\{y_i(1) : s_i(0) = s_i(1) = 1\}$ and $\{y_i(0) : s_i(0) = s_i(1) = 1\}$. While motivated by infectious disease settings where the AI stratum is of interest, the PSET can be applied to alternative principal stratum defined by infection such as the immune stratum as well as a variety of other scenarios. 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 (Rubin 2006).

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

$$H_0 : y_i(1) = y_i(0) \text{ for all } i \in \mathcal{AJ}, \quad (1)$$

where $\mathcal{AJ} \equiv \{i : s_i(1) = s_i(0) = 1\}$ is the set of individuals in the AI stratum, against

$$H_A : y_i(1) \geq y_i(0) \text{ for all } i \in \mathcal{AJ}, \quad (2)$$

where the inequality in (2) is strict for at least one $i \in \mathcal{AJ}$. Using the terminology of VanderWeele (2008), the null (1) corresponds to no principal stratum direct effect. An exact test requires that the resulting p-value, p , must be exact in the sense that $\Pr[p \leq \alpha] \leq \alpha$ for each $\alpha \in [0, 1]$ under the null hypothesis (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 for each individual dependent on treatment assignment, either $(s_i(1), y_i(1))$ or $(s_i(0), y_i(0))$. Let Z_i represent the treatment assignment for the i^{th} individual where $Z_i = 0$ for control and $Z_i = 1$ for treatment, and $\mathbf{Z} = (Z_1, \dots, Z_n)$ represent the vector of treatment assignments for all individuals. To make inference about the treatment effect on the outcome, 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 $(s_i(1), s_i(0), y_i(1), y_i(0))$ 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. Hence, S_i^{obs} and Y_i^{obs} are also random variables since they depend on Z_i . To develop a test of (1), assume independent treatment assignment:

A.2 Independent treatment assignment: $\Pr[\mathbf{Z} = \mathbf{z}] = \Pr[\mathbf{Z} = \mathbf{z}']$ for any $\mathbf{z}, \mathbf{z}'$ such that $\sum_{i=1}^n z_i = \sum_{i=1}^n z'_i$ where $\mathbf{z} = (z_1, \dots, z_n)$ and $\mathbf{z}' = (z'_1, \dots, z'_n)$ are treatment assignment

vectors.

If principal stratum membership was known, for the AI stratum in particular, the development of an exact test of (1) against (2) would be straight-forward. However, 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 $i = 1, \dots, n$.

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., $s_i(0) = 1$). As such, all infected treated individuals are members of the AI stratum i.e., $\{i : S_i^{obs} = 1, Z_i = 1\} \subseteq \mathcal{AJ}$. Unfortunately, A.1-A.3 do not identify AI stratum membership for individuals assigned control because infected individuals in the control group are a mixture of individuals from both the AI and protected strata. Therefore, AI stratum membership for control individuals remains unknown making the development of an exact test challenging, but not impossible.

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 (1), first imagine we know exactly which individuals are in $\mathcal{AJ}$. Then the following 2×2 table can be constructed

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

(3)

where $m \equiv \sum_{i \in \mathcal{AJ}} 1$ is the number of individuals in $\mathcal{AJ}$. Under the sharp null, $Y_i^{obs} = y_i(0)$.

Thus (3) can equivalently be written as

	Event	No Event	
Treatment	$\sum_{i \in \mathcal{AJ}} Z_i y_i(0)$	$\sum_{i \in \mathcal{AJ}} Z_i (1 - y_i(0))$	$\sum_{i \in \mathcal{AJ}} Z_i$
Control	$\sum_{i \in \mathcal{AJ}} (1 - Z_i) y_i(0)$	$\sum_{i \in \mathcal{AJ}} (1 - Z_i) (1 - y_i(0))$	$\sum_{i \in \mathcal{AJ}} (1 - Z_i)$
	$\sum_{i \in \mathcal{AJ}} y_i(0)$	$\sum_{i \in \mathcal{AJ}} (1 - y_i(0))$	m

(4)

For randomization inference about the finite population of n individuals, the column totals of (4) are considered fixed. Conditional on the row totals, (1) versus (2) can be tested using Fisher's exact test where the p-value is obtained by calculating the probability of each potential table using the hypergeometric distribution.

Unfortunately principal strata membership is not completely known, so we cannot construct (4). Instead the following 2×2 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	$\sum (1 - Z_i) S_i^{obs} Y_i^{obs}$	$\sum (1 - Z_i) S_i^{obs} (1 - Y_i^{obs})$	$\sum (1 - Z_i) S_i^{obs}$
	$\sum S_i^{obs} Y_i^{obs}$	$\sum S_i^{obs} (1 - Y_i^{obs})$	$\sum S_i^{obs}$

(5)

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

To develop an exact test of (1) versus (2), information from (5) can be used to make inference about the unobservable 2×2 table (4). Under A.3, $Z_i S_i^{obs} = 1$ implies $i \in \mathcal{AJ}$; that is, all treated individuals who become infected must be members of the AI stratum. Thus assuming A.3, under (1) the observable table (5) can be written as

	Event	No Event	
Treatment	$\sum_{i \in \mathcal{AJ}} Z_i y_i(0)$	$\sum_{i \in \mathcal{AJ}} Z_i (1 - y_i(0))$	$\sum_{i \in \mathcal{AJ}} Z_i$
Control	$\sum (1 - Z_i) S_i^{obs} y_i(0)$	$\sum (1 - Z_i) S_i^{obs} (1 - y_i(0))$	$\sum (1 - Z_i) S_i^{obs}$
	$\sum S_i^{obs} Y_i^{obs}$	$\sum S_i^{obs} (1 - Y_i^{obs})$	$\sum S_i^{obs}$

(6)

Table (6) differs from (4) 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 a nuisance parameter. One solution for obtaining a p-value is to conduct tests over the entire range of the nuisance parameter and report the largest p-value (Barnard 1947; Silvapulle 1996). Accordingly, an exact p-value for testing (1) can be obtained by taking the largest p-value from a set of exact tests (e.g. Fisher's test) performed for all possible AI tables (4) consistent with the observed data table (6).

However, by considering all possible AI tables (4) consistent with (6) and selecting the worst case p-value, the test could be overly conservative; such criticism has been levied against Barnard's test. Berger and Boos (1994) suggest considering only values of the nuisance parameter restricted to some $100(1 - \gamma)\%$ confidence interval and adding γ to the largest p-value over the restricted set of tests. The observed data table (6) provides information about the likelihood of various underlying AI tables (4) that can be used to develop an exact test of (1) using this approach.

2.3 PSET Development

Now assume that $y_i(1)$ and $y_i(0)$ are any type of event and not necessarily binary. Let $\mathbf{Y}_1^{ai} \equiv \{Y_i^{obs} : Z_i = 1, i \in \mathcal{AI}\}$ and $\mathbf{Y}_0^{ai} \equiv \{Y_i^{obs} : Z_i = 0, i \in \mathcal{AI}\}$ and define $p(\mathbf{Y}_1^{ai}, \mathbf{Y}_0^{ai})$ as the p-value for an exact randomization-based test of (1) versus (2) assuming AI membership were known. For example, if no ties exist in $(\mathbf{Y}_1^{ai}, \mathbf{Y}_0^{ai})$ then the usual exact Wilcoxon rank sum test could be employed. In this scenario, the test statistic is equal to the sum of the ranks of the elements of $\mathbf{Y}_1^{ai}$ when all elements of $\mathbf{Y}_1^{ai}$ and $\mathbf{Y}_0^{ai}$ are ordered and large values of the test statistic provide evidence in favor of (2).

However, as illustrated in Section 2.2, it is not possible to identify which infected control individuals belong to $\mathcal{AI}$. Accordingly, the set of AI stratum membership indica-

tors 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. Unfortunately, this approach is overly conservative because it almost always fails to reject (1). Specifically, the set of possible always-infected 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 less than the $\{y_i(1) : s_i(1), Z_i = 1\}$, (1) will not be rejected. Furthermore, this approach ignores information available about the AI stratum. While the observed data do not provide information on individual AI stratum membership for control individuals, the data do provide some information on bounds of possible values for the number of control individuals in the AI stratum. This suggests 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 $M_0 \equiv \sum_{i \in \mathcal{AI}} (1 - Z_i)$ and $M_1 \equiv \sum_{i \in \mathcal{AI}} Z_i$, i.e., M_0 and M_1 are the number of individuals in $\mathcal{AI}$ assigned control and treatment such that $M_0 + M_1 = m$. Since the number of individuals in $\mathcal{AI}$ does not depend on $\mathbf{Z}$, m is fixed, whereas M_0 and M_1 are random variables since they depend on $\mathbf{Z}$. Under A.3, $M_1 = \sum_{i=1}^n I[Z_i = S_i^{obs} = 1]$ 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 $\{i : Z_i = 0, S_i^{obs} = 1\}$ and taking the maximum of the resulting p-values. Define this p-value as

$$p^{ai}(M_0) \equiv \max\{p(\mathbf{Y}_1^{ai}, \mathbf{Y}_0) : \mathbf{Y}_0 \in \Omega(M_0)\} \quad (7)$$

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

Although M_0 is not observed, it is bounded above by $\sum_{i=1}^n I[Z_i = 0, S_i^{obs} = 1]$ un-

der A.3. Moreover, the observed M_1 provides some information about m and thus M_0 . Specifically, conditional on the total number assigned treatment $\sum_{i=1}^n Z_i$, 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 $\sum_i Z_i S_i^{obs} = \sum_{i \in \mathcal{A}^J} Z_i$ using standard results under simple random sampling (e.g., Thompson 2002). Then, following Berger and Boos (1994), define

$$p_\gamma^{ai} \equiv \max\{p^{ai}(\tilde{m} - M_1) : \tilde{m} \in C_\gamma\} + \gamma. \quad (8)$$

The following proposition indicates that p_γ^{ai} is an exact p-value for testing (1) versus (2) .

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 (1).

The proof of the proposition is in the Appendix. 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 3 presents simulation studies which provide empirical evidence suggesting $\gamma = \alpha/2$ may be recommended in certain settings. For tests where the conditional p-values $p^{ai}(\tilde{m} - M_1)$ tend to decrease as $\tilde{m}$ increases, letting $U_m = \sum_i S_i^{obs}$ and computing a one-sided $(1 - \gamma)\%$ CI for m to obtain L_m should result in a test with higher power compared to a test using a two-sided CI.

2.4 Computations

Calculating $p^{ai}(M_0)$ can be computationally intensive as it requires performing $\left(\sum_{M_0}^{(1-Z_i)S_i^{obs}}\right)$ exact tests corresponding to $\Omega(M_0)$. In certain settings, the computation requirements can be reduced by identifying the maximum $p(\mathbf{Y}_1^{ai}, \mathbf{Y}_0)$ for $\mathbf{Y}_0 \in \Omega(M_0)$ without having to calculate $p(\mathbf{Y}_1^{ai}, \mathbf{Y}_0)$ for all $\mathbf{Y}_0$. For example, if the p-value increases monotonically as the sum of the ranks of $\mathbf{Y}_0$ increases, the maximum p-value for tests of (1) against (2) is obtained when the ranks of $\mathbf{Y}_0$ are as large as possible. Specifically, the maximum p-value is obtained for $\mathbf{Y}_0 = \mathbf{Y}_0^{ai}[1 : M_0]$ where $\mathbf{Y}_0^{ai}[1 : M_0]$ is the M_0 largest values of

$\{Y_i^{obs} : Z_i = 0, S_i^{obs} = 1\}$. In this case $p^{ai}(M_0) = p(\mathbf{Y}_1^{ai}, \mathbf{Y}_0^{ai}[1 : M_0])$.

2.5 Plug-in P-value Alternative

An alternative, simple approach for dealing with nuisance parameters entails conditioning on estimates of the unknown parameters; the resulting p-value is referred to as the “plug-in p-value” (Bayarri and Berger 2000). A plug-in p-value testing (1) versus (2) can be obtained by conditioning on an unbiased estimate of m , $\hat{M} = nM_1 / \sum_{i=1}^n Z_i$ (Thompson 2002) such that

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

Although intuitive, this approach does not take into account uncertainty about m and therefore p_{plug}^{ai} is not guaranteed to be exact. For example, consider the following population of 6 individuals.

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	6	6	4	1	1	3	3
2	1	1	5	5	5	1	0	2	*
3	1	1	4	4	6	1	0	1	*

Suppose by design $\sum_{i=1}^n Z_i = 3$ is fixed. Then, for $\alpha = 0.20$, $\Pr[p_{plug}^{ai} \leq \alpha] = 0.25 > \alpha$ because $p_{plug}^{ai} \leq \alpha$ for 5 of the $\binom{6}{3} = 20$ possible treatment assignment permutations.

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 large sample frequentist 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 in the placebo arm, 31.5 in the vaccine arm), the proposed tests demonstrated inflated type I error. Therefore we conducted a simulation study under identical assumptions to assess how the PSET from Section 2.3 performs in comparison.

In this setting $z = 0$ for placebo individuals and $z = 1$ for vaccinated individuals. For assignment z , $s_i(z) = 1$ if an individual is infected and $y_i(z)$ represents log-transformed viral load when $s_i(z) = 1$. Thus, the null (1) corresponds to the vaccine having no effect on viral load in individuals in the AI principal stratum. It is of interest to test the null against the one-sided alternative (2) 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). Therefore simulations under (2) were conducted assuming $y_i(1) = y_i(0) + \delta$ for $\delta > 0$.

The following steps were performed for each trial simulation. First, $s_i(0)$ was set equal to 1 for $i = 1, \dots, 90$ and $s_i(0) = 0$ for $i = 91, \dots, 2000$. For $i = 1, \dots, 90$, $y_i(0)$ was randomly generated from a normal distribution with mean 4.5 and standard deviation 0.6. Next, selection bias was simulated by setting $s_i(1) = 1$ and $y_i(1) = y_i(0) + \delta$ for the 63 individuals with the largest values of $y_i(0)$ such that individuals that would become mildly infected under control (i.e. low viral load) are protected under vaccine. For the remaining individuals $s_i(1) = 0$ and $y_i(1)$ was set to missing. Finally, exactly 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 and the nonparametric mean bootstrap test of Hudgens et al. (2003) 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 (7). Simulations using a two-sided $100(1 - \gamma)\%$ CI to obtain L_m and U_m for the PSET resulted in reduced power compared to the one-sided approach (results not shown). For each simulation, the plug-in p-value from Section 2.5 was also computed. Table 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 nonparametric mean bootstrap test and plug-in p-value were 0.11 and 0.19 respectively, i.e. over twice the nominal level. Similar results were obtained for $\alpha = 0.10$. 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.

4 Applications

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 at birth to the intervention and 477 randomized at birth to standard practice of continued breastfeeding. 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 the authors to conclude that there is evidence of a harmful effect of weaning on survival.

Because the reported analysis conditions on infection and survival status at 4 months, the results do not have a causal interpretation and could be due to selection bias. The principal stratum of interest for this analysis 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 against the one-sided alternative that the intervention caused at least one death in the AI stratum. To compute $p(\mathbf{Y}_1^{ai}, \mathbf{Y}_0)$ in equation (7), a Fisher's exact test was used where each individual was classified as having died or not. Ideally an exact log-rank test would be preferable for calculating the conditional p-values, however the individual death and censoring times were not reported by Kuhn et al. For $\alpha = 0.05$ and $\gamma = 0.025$, 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 includes a plot of the p-values from Fisher's exact test conditional on $m = \tilde{m}$ for each $\tilde{m} \in C_\gamma$. While a Fisher's exact test using all available data and ignoring the potential for selection bias is significant (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 a non-significant $p_{plug}^{ai} = 0.1611$. In order to obtain a significant p-value 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 out of 70

in the standard practice arm ($p_{\gamma}^{ai} = 0.0375$).

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 stopped enrollment in this arm early). The effect of NVP on infection status through 28 weeks was assessed using a stratified log-rank test that compared infection between treatment groups for infants who were not infected at 2 weeks. Of the 632 infants not infected at 2 weeks in the control arm, 32 (5.1%) were infected by 28 weeks. Likewise, of the 815 infants not infected at 2 weeks in the NVP arm, 12 (1.5%) were infected by 28 weeks. The log-rank p-value was < 0.001 and the authors concluded NVP prevents breastmilk transmission of HIV.

The principal stratum of interest for this analysis is the never-infected (NI) stratum, defined as all individuals who would not be HIV-infected at 2 weeks regardless of randomization assignment. The PSET is still applicable in this setting; however, the test is performed reversing the roles of treated and control individuals. Membership in the NI stratum is known for control individuals since individuals not infected at 2 weeks when receiving control would also not be infected at 2 weeks when receiving NVP by the monotonicity assumption. However, membership in the NI stratum is unknown for treated infants not infected at 2 weeks. The PSET can be used to test the null hypothesis of no treatment effect on infection by 28 weeks versus the one-sided alternative that at least

one infection was averted due to treatment in the NI stratum. To compute (7), an exact stratified log-rank tests was employed using StatXact (Mehta and Patel 2007). Due to the large sample size of the BAN study, Monte Carlo sampling was used to approximate the exact conditional p-values. For $\alpha = 0.025$ and $\gamma = 0.0125$, the PSET resulted in $p_{\gamma}^{ai} = 0.0125$ indicating a benefit of NVP among infants who were immune to infection at two weeks. Using a Fisher's exact test to compute (7) as in Section 4.1 yielded a similar result with $p_{\gamma}^{ai} = 0.0126$.

5 Comparisons with ITT

Principal stratification provides a method for dealing with potential selection bias induced by conditioning on an intermediate post-randomization outcome. As an alternative, an intent-to-treat (ITT) based approach can be employed. The ITT principle generally refers to analyzing all randomized 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 example, in infectious disease prevention studies, it is conceptually challenging to define post-infection outcomes for non-infected individuals. Also, in contrast to principal stratification, the ITT approach does not clearly differentiate treatment effects on infection and post-infection outcomes.

Chang, Guess and Heyse (1994) proposed an ITT-based approach for assessing treatment effect on disease severity by assigning disease severity scores to each incident infection, with individuals who escape infection receiving a score of zero. In a similar manner, a randomization-based ITT test can be conducted using the observed disease severity

scores 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 $\mathbf{W}_1^{ITT} \equiv \{W_i^{obs} : Z_i = 1\}$, $\mathbf{W}_0^{ITT} \equiv \{W_i^{obs} : Z_i = 0\}$ and $p(\mathbf{W}_1^{ITT}, \mathbf{W}_0^{ITT})$ as the p-value for an exact randomization-based test comparing $\mathbf{W}_1^{ITT}$ and $\mathbf{W}_0^{ITT}$. As opposed to testing (1) against (2), this approach tests $H_0 : w_i(1) = w_i(0)$ for all i against the one-sided alternative $H_A : w_i(1) \geq w_i(0)$ for all i where the inequality is strict for at least one i . Assuming $Y_i^{obs} > 0$ whenever $S_i^{obs} = 1$, it follows that this approach is equivalent to testing the composite hypothesis:

$$H_0 : s_i(1) = s_i(0) \text{ and } y_i(1) = y_i(0) \text{ for all } i, \quad (9)$$

against the one-sided alternative

$$H_A : s_i(1) \geq s_i(0) \text{ and } y_i(1) \geq y_i(0) \text{ for all } i, \quad (10)$$

where one of the inequalities of (10) is strict for at least one i . When $s_i(1) = s_i(0)$ for all i , testing (9) against (10) is a test of the post-infection outcome of interest in the finite population of n individuals.

The ITT approach was assessed using a simulation scenario similar to that described in Section 3 except that $s_i(1)$ was set equal to $s_i(0)$ for all i such that both the ITT and PSET tested only the post-infection outcome of interest. For the ITT test, W_i^{obs} was set to zero for individuals who escaped infection and set to the observed viral load (Y_i^{obs}) for infected individuals ($S_i^{obs} = 1$). The Wilcoxon rank sum test was used to compare the observed disease severity scores W_i^{obs} between the 1000 vaccine and 1000 control individuals. For $\alpha = 0.05$, the empirical type I error and power for $\delta = 1/3$ and $2/3$ were all < 0.05 for the ITT approach and were 0.001, 0.09 and 0.65 respectively for the PSET (with $\gamma = 0.025$). That is, the PSET is markedly more powerful test than the ITT approach in this scenario. Mehrotra et al. (2006) presented similar findings comparing large-sample frequentist based principal stratification tests with an ITT approach.

6 Adjusting for Covariates

Covariate adjustment is often used in analysis of randomized experiments to account for chance imbalances that may exist in the observed data, thereby allowing for more precise inference. Following Rosenbaum (2002b), in this section we consider extending the PSET described above 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.

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 (1), $y_i(1) = y_i(0)$ and therefore $e_i(1) = e_i(0)$ for all $i \in \mathcal{AJ}$. Therefore, a test of (1) can be constructed using the residuals.

Testing $H_0 : e_i(1) = e_i(0)$ for all $i \in \mathcal{AJ}$ against $H_A : e_i(1) \geq e_i(0)$ for all $i \in \mathcal{AJ}$ where the inequality is strict for at least one i can be used to test (1) against (2) provided $g(y, x)$ is monotonically increasing in y . Note under (2) that $y_i(1) = y_i(0) + \delta_i$ where $\delta_i \geq 0$ and the inequality is strict for at least one i . If $g(y, x)$ is monotonically increasing in y such that $g(y_1, x) \geq g(y_0, x)$ if $y_1 \geq y_0$, then under (2) $e_i(1) = g(y_i(1), x_i) = g(y_i(0) + \delta_i, x_i) \geq$

$g(y_i(0), x_i) = e_i(0)$. An example of a residual function monotone in y is $g(y, x) = y - (\hat{\beta}_0 + \hat{\beta}_1 x)$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are solutions to the normal equations for the linear regression model of y on x .

The covariate-adjusted PSET is constructed in a similar fashion to the PSET from Section 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 $p(\mathbf{Y}_1^{ai}, \mathbf{Y}_0^{ai})$ 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 $g(\cdot, \cdot)$. 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 $p(\mathbf{Y}_1^{ai}, \mathbf{Y}_0^{ai})$ while a Wilcoxon rank sum test could be used on the residuals.

6.2 Simulations

To assess the power of the covariate adjusted PSET, the simulation scenario described in Section 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 847 and standard deviation 307 (Oladepo et al. 2009). 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 ($\rho = 0.0, 0.1, \dots, 0.9$). 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 Wilcoxon rank sum test of the residuals was used to obtain the conditional p-values $p^{ai}(M_0)$ in (7).

Results of the simulations for $\alpha = 0.05$ and $\gamma = 0.025$ are displayed in Table 2.

Comparing the results to Table 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. Although, even for $\rho=0$ there was no loss in power when adjusting for the covariate.

7 Confidence Interval for Principal Stratum Direct Effect

In this section we show how the PSET can be used to form an exact confidence interval (CI) for a principal stratum direct effect. Suppose the treatment effect in AI is additive such $y_i(1) - y_i(0) = \delta_0$ for all $i \in \mathcal{AJ}$. 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 2.3. 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 2002).

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:

$$H_0 : y_i(1) - y_i(0) = \delta \text{ for all } i \in \mathcal{AJ}, \quad (11)$$

against

$$H_A : y_i(1) - y_i(0) \geq \delta \text{ for all } i \in \mathcal{AJ}. \quad (12)$$

where the inequality in (12) is strict for at least one $i \in \mathcal{AJ}$. Note under (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 (11) versus (12) can be constructed just as in Section 2.3 except Y_i^{obs} is replaced with $Y_i^{obs} - \delta$ for individuals where $Z_i = 1$ (Rosenbaum 2002b). Then a one-

sided $100(1 - \alpha)\%$ CI for the true (unknown) δ_0 is formed by the set of all δ where (11) is not rejected.

More specifically, let $p_\delta(\mathbf{Y}_1^{ai}, \mathbf{Y}_0^{ai}) = p(\mathbf{Y}_1^{ai} - \delta, \mathbf{Y}_0^{ai})$ denote the p-value from an exact randomization-based test of (11) using $Y_i^{obs} - \delta$ for individuals where $Z_i = 1$, and let $p_\delta^{ai}(M_0) = \max\{p_\delta(\mathbf{Y}_1^{ai}, \mathbf{Y}_0) : \mathbf{Y}_0 \in \Omega(M_0)\}$ and $p_{\gamma, \delta}^{ai} = \max\{p_\delta^{ai}(\tilde{m} - M_1) : \tilde{m} \in C_\gamma\} + \gamma$. Let δ_{min} and δ_{max} denote the lower and upper limits of the range of possible values for δ . Following Mehta and Patel (2007), define the lower bound of the $100(1 - \alpha)\%$ one-sided CI of δ_0 by $\Delta_\gamma^\alpha = \sup\{\delta : p_{\gamma, \tilde{\delta}}^{ai} \leq \alpha \text{ for all } \tilde{\delta} < \delta\}$. If there exists no 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 $100(1 - \alpha)$. The proof of Proposition 2 is in the Appendix.

Proposition 2: For $\gamma \in [0, 1]$ and $\alpha \in [0, 1]$, $Pr[\delta_0 \in [\Delta_\gamma^\alpha, \delta_{max}]] \geq 1 - \alpha$.

8 Conclusion

The PSET provides a means for randomization-based inference about the direct effect of treatment within a principal stratum. Assumptions of random sampling from an infinite population or parametric distributions are not required for the PSET. Simulation studies indicate the PSET can be more powerful than an ITT approach when treatment has no impact on the intermediate post-randomization outcome. The power of the PSET can be increased by adjusting for baseline covariates. Finally, exact confidence intervals for the principal stratum direct effect can be obtained by inverting the PSET.

Appendix

Proof of Proposition 1

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

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

where the 2nd inequality holds because $\max\{p^{ai}(\tilde{m} - M_1) : \tilde{m} \in C_\gamma\} \geq p^{ai}(m - M_1)$ when $m \in C_\gamma$ and the 4th 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 (1).

Proof of Lemma: Suppose H_0 (1) is true.

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

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

Collection of Biostatistics
Research Archive

Proof of Proposition 2

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 (11) for any choice of δ . Therefore, $Pr[\delta_0 \notin [\Delta_\gamma^\alpha, \delta_{max}]] = Pr[\delta_0 \notin [\delta_{min}, \delta_{max}]] = 0 \leq \alpha$. If $\gamma \leq \alpha$ then

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

where the 3rd 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 4th inequality follows for reasons analogous to the proof of Proposition 1.

References

- [1] 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.
- [2] 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.
- [3] G. A. Barnard. Significance tests for 2×2 tables. Biometrika, 34:123–138, 1947.
- [4] M. J. Bayarri and J. O. Berger. P values for composite null models. J Am Stat Assoc, 95(452):1127–1142, 2000.
- [5] 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.

- [6] G. Casella and R. L. Berger. Statistical Inference. Duxbury Press, 2002.
- [7] 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.
- [8] 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, In press.
- [9] C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. Biometrics, 58:21–29, 2002.
- [10] 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.
- [11] P. B. Gilbert and M. G. Hudgens. Evaluating candidate principal surrogate endpoints. Biometrics, 64:1146–1154, 2008. NIHMSID 92233.
- [12] M. E. Halloran and C. J. Struchiner. Causal inference in infectious diseases. Epidemiology, 6:142–151, 1995.
- [13] M. A. Hernán, S. Hernández-Daz, and J. M. Robins. A structural approach to selection bias. Epidemiology, 15:615–625, Sep 2004.
- [14] M. G. Hudgens and M. E. Halloran. Causal vaccine effects on binary post-infection outcomes. J Am Stat Assoc, 101:51–64, 2006.
- [15] 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.

- [16] K. Imai. Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". Stat Probab Lett, 78:144–149, 2008.
- [17] Y. Jemai, 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.
- [18] H. Jin and D. B. Rubin. Public schools versus private schools: causal inference with partial compliance. J Educ Behav Stat, 34:24–45, 2009.
- [19] M. M. Joffe and T. Greene. Related causal frameworks for surrogate outcomes. Biometrics, 65:530–538, 2009.
- [20] 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.
- [21] 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.
- [22] E. L. Lehmann. Testing Statistical Hypotheses. New York: Wiley, 1959.
- [23] 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.
- [24] C. Mehta and N. Patel. StatXact 8 User Manual. Cytel, Cambridge, MA, 2007.

- [25] D. K. Oladepo, E. O. Idigbe, R. A. Audu, G. E. Inyang, A. O. Philip, G. O. Okafor, D. Olaleye, S. B. Mohammed, N. N. Odunukwe, T. O. Harry, M. Edyong-Ekpa, J. Idoko, A. Z. Musa, A. Adedeji, A. Nasidi, Y. Ya'aba, and K. Ibrahim. Establishment of reference values of CD4 and CD8 lymphocyte subsets in healthy Nigerian adults. Clin Vaccine Immunol, 16:1374–1377, 2009.
- [26] 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.
- [27] J. M. Robins. An analytic method for randomized trials with informative censoring: Part I. Lifetime Data Anal, 1:241–254, 1995.
- [28] 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.
- [29] P. R. Rosenbaum. Observational Studies. New York: Springer-Verlag, 2002a.
- [30] P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. Statist Sci, 17:286–327, 2002b.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.

- [35] B. Shepherd, P. B. Gilbert, Y. Jemai, 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.
- [36] 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, 2007.
- [37] B. E. Shepherd, P. B. Gilbert, and D. V. Mehrotra. Eliciting a counterfactual sensitivity parameter. Am Stat, 61(1):56–63, 2007.
- [38] M. J. Silvapulle. A test in the presence of nuisance parameters. J Am Stat Assoc, 91(436):1690–1693, 1996.
- [39] S. K. Thompson. Sampling. New York: John Wiley & Sons, 2002.
- [40] T. J. VanderWeele. Simple relations between principal stratification and direct and indirect effects. Stat Probab Lett, 2008.
- [41] 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.

Table 1: Empirical type 1 error and power for various values of γ , significance level (α) and location shifts (δ) in distribution of $\log_{10}$ viral load where $\delta = 0$ under the null hypothesis (1)

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

Table 2: Empirical type 1 error and power for $\alpha = 0.05$ and $\gamma = 0.025$ and various location shifts (δ) in distribution of $\log_{10}$ viral load where $\delta = 0$ under the null hypothesis (1), when adjusting for baseline CD4 count at various levels of ρ between viral load and CD4 count

ρ	$\delta = 0$	$\delta = 1/3$	$\delta = 2/3$
0.0	0.004	0.15	0.75
0.1	0.004	0.15	0.75
0.2	0.004	0.15	0.76
0.3	0.003	0.15	0.77
0.4	0.003	0.16	0.79
0.5	0.002	0.17	0.81
0.6	0.001	0.18	0.84
0.7	0.001	0.20	0.90
0.8	0.001	0.26	0.96
0.9	< 0.001	0.49	> 0.99

Figure 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_{plug}^{ai} where $\tilde{m} = \hat{M} = nM_1 / \sum_{i=1}^n Z_i = 123$