Published in final edited form as:

Stat Med. 2013 November 20; 32(26): 4526–4539. doi:10.1002/sim.5849.

Rank-Based Principal Stratum Sensitivity Analyses

X. Lu^{a,*}, D. V. Mehrotra^b, and B. E. Shepherd^c

^aDepartment of Biostatistics, University of Florida, Gainesville, FL 32610, USA

^bDepartment of Clinical Biostatistics, Merck Research Laboratories, North Wales, PA 19454, USA

^cDepartment of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

Abstract

We describe rank-based approaches to assess principal stratification treatment effects in studies where the outcome of interest is only well-defined in a subgroup selected after randomization. Our methods are sensitivity analyses, in that estimands are identified by fixing a parameter and then we investigate the sensitivity of results by varying this parameter over a range of plausible values. We present three rank-based test statistics and compare their performance through simulations, and provide recommendations. We also study three different bootstrap approaches for determining levels of significance. Finally, we apply our methods to two studies: an HIV vaccine trial and a prostate cancer prevention trial.

Keywords

Causal inference; Monotonicity; Non-normality; Principal stratification; Ranks

1. Introduction

In many settings, researchers are interested in assessing the effect of treatment on an outcome that is well defined among a subset of participants selected after randomization. For example, in prophylactic HIV vaccine trials there is interest in assessing the effect of vaccination on viral load among patients who become infected (i.e., selected for the viral load analysis) during the course of the trial. Many candidate vaccines have been specifically designed to elicit cell-mediated immune (CMI) responses whereby vaccine-induced immunity may help abort an emerging infection before it becomes fully established or keep the amount of virus (viral load) in the infected subject at low levels. Therefore, what happens to participants post-infection is of primary interest in these trials. A recent trial of a CMI-based HIV vaccine candidate (the Step trial) had dual primary outcomes: HIV infection and post-infection viral load [1].

A second example comes from the Prostate Cancer Prevention Trial (PCPT). In this randomized trial of finasteride versus placebo, men randomized to finasteride had lower rates of prostate cancer, yet their cancers seemed more severe [2]. In this study there is

*Correspondence to: Department of Biostatistics, University of Florida, Gainesville, FL 32610, USA. xlu2@ufl.edu.

interest in assessing the effect of finasteride on the severity of prostate cancer among those men who were diagnosed with prostate cancer (selected) after randomization. Other examples arise in settings where there is truncation or censoring by death [3–5], where researchers are interested in assessing the effect of treatment on an outcome that is only defined among survivors.

As is well documented, an analysis that conditions on post-randomization selection could be subject to selection bias [6]. To avoid this problem in the context of HIV vaccine trials, Gilbert *et al.* [7] (GBH) proposed looking at the effect of vaccination in the subgroup of individuals who would have been infected regardless of their treatment assignment. Subgroups of this nature have been referred to as principal strata [8]. Symbolically, let $S(0)$ be the indicator of being selected (e.g., infected) if randomized to treatment 0 (e.g., placebo), and let $S(1)$ be the indicator of selection if randomized to treatment 1 (e.g., vaccine). Subgroups defined by $\{S(0) = a_0, S(1) = a_1\}$, for $a_0, a_1 \in \{0, 1\}$, are called principal strata. Measures of treatment effect within principal strata enable valid causal comparisons because principal stratum membership is not affected by treatment assignment. Among participants in the “doomed” (or “always infected”) principal stratum, i.e. those with $\{S(0) = S(1) = 1\}$, GBH compared viral load distributions between those assigned to vaccine versus those assigned to placebo. To identify these distributions, GBH first assumed that the vaccine did not cause infections, or that every subject infected in the vaccine arm was in the “doomed” principal stratum. Second, GBH modeled the probability of being infected if randomized to vaccine, given a participant was randomized to placebo and subsequently infected with a known viral load. For identification, this model requires fixing a sensitivity parameter, which is varied over a plausible range of values in a sensitivity analysis. An extension of the approach of GBH has been shown to be more powerful than analyses where uninfected individuals are assigned zero viral load values within a burden of illness analysis [9]. GBH has also been extended to incorporate binary outcomes [10], right-censored outcomes [11], baseline covariates [12–14], and to settings with less stringent assumptions [15, 16]. These and other principal stratification methods have been applied to a wide variety of applications.

The goal of this manuscript is to extend GBH by developing rank-based tests. GBH proposed non-parametric test statistics that did not specify the distribution of outcome variables but used the empirical distribution. However, many trials’ primary analyses incorporate rank-based test statistics. For example, a rank-based analysis of viral loads was planned for the Step trial [1] to prevent the analysis from being dominated by viral load outliers and to accommodate censored values due to assay detection limits. Rank-based comparisons may also be more natural in settings where the outcome is not a continuous variable. For example, cancer severity in the PCPT was measured using Gleason scores, which is an ordered categorical variable ranging from 2–10, with higher scores designating more severe cancer. A secondary, yet important, aim of this manuscript is to evaluate different bootstrap approaches for computing significance levels. GBH proposed using a modification of a ‘controls only’ bootstrap procedure developed by Hudgens *et al.* [17]. In addition to the ‘controls only’ bootstrap, we also consider two additional bootstrap procedures: (1) a modification of another bootstrap procedure proposed by Hudgens *et al.* [17] that incorporates outcome data from both trial arms, and (2) a simple nonparametric bootstrap approach generally used to compute confidence intervals.

In section 2 we introduce notation, our estimands, and inference procedures. In section 3 we contrast different test statistics and bootstrap options using simulation studies. In section 4 we apply our methods to the analysis of data from an HIV vaccine trial and the PCPT. Concluding remarks are provided in section 6.

2. Methods

2.1. Notation and Assumptions

Let Z_i be the treatment assignment for subject i , where $i = 1, \dots, N$, with N being the number of subjects in the trial; $Z_i = 0$ or 1 if subject i is assigned to control or treatment, respectively. Let $S_i(z)$ be the indicator of post-randomization selection for the i th subject if, possibly contrary to fact, s/he was assigned $Z_i = z$. Let $Y_i(z)$ be the outcome of interest given $Z_i = z$. If $S_i(z) = 0$, then $Y_i(z)$ may not be well defined and, following others, will be assigned the value $*$. Our notation follows GBH, and has implicitly assumed Rubin's stable unit treatment value assumption (SUTVA) [18], which implies that potential outcomes for each subject i are unrelated to the assignment Z_j of other subjects $i \neq j$, and the observed outcome under the assignment of treatment z is the same as the corresponding potential outcome $Y_i(z)$. Consequently, under SUTVA, each subject has two potential selection outcomes ($S_i(0)$, $S_i(1)$) and at most two potential response outcomes ($Y_i(0)$, $Y_i(1)$); hereafter the term 'response' will be dropped for the Y variables. The vectors (Z_i , $S_i(0)$, $S_i(1)$, $Y_i(0)$, $Y_i(1)$) for $i = 1, \dots, N$ are assumed to be i.i.d. copies of a random vector (Z , $S(0)$, $S(1)$, $Y(0)$, $Y(1)$). Since subjects are assigned to either treatment or control, but not both, the observed data, denoted (Z_i , S_i , Y_i) for subject i , equals (Z_i , $S_i(0)$, $Y_i(0)$) if $Z_i = 0$ and (Z_i , $S_i(1)$, $Y_i(1)$) if $Z_i = 1$.

A causal estimand of treatment effect on Y can be any function that measures a contrast between the distributions of the two potential outcomes $Y_i(0)$ and $Y_i(1)$, within a subgroup of subjects with a common pair of potential selection outcomes, $\{S_i(0) = a_1, S_i(1) = a_2\}$, where $a_1, a_2 \in \{0, 1\}$ are fixed values. Such a subgroup was referred to as a *principal stratum* by Frangakis and Rubin [8]. In GBH, the authors only considered the principal stratum defined by $\{S_i(0) = S_i(1) = 1\}$. This stratum consists of those who would be selected regardless of treatment assignment, and following GBH, we will refer to it as the "always selected" principal stratum and consider a causal estimand in this stratum. The distributions of the two potential response outcomes are given by

$$F_0^{11}(y) \equiv \Pr\{Y_i(0) \leq y | S_i(0) = S_i(1) = 1\}$$

$$F_1^{11}(y) \equiv \Pr\{Y_i(1) \leq y | S_i(0) = S_i(1) = 1\}. \quad (1)$$

Based on (1), a null hypothesis for no causal effect of treatment on outcome in the always selected principal stratum can be expressed as

$$H_0: F_0^{11}(y) = F_1^{11}(y) \text{ for all } y. \quad (2)$$

Since $S_A(0)$ and $S_A(1)$ are not both observable, neither of the distributions in (1) is identifiable, and it is not possible to test the null hypothesis in (2) without additional assumptions. Therefore, in addition to SUTVA, GBH added the following two assumptions.

A1: The assignment Z_i of each subject is independent of his/her potential outcomes.

A2: For each subject i , $\Pr\{S_A(1) = 1, S_A(0) = 0\} = 0$.

Assumption A1, often referred to as “ignorable treatment assignment,” holds by design in a randomized experiment. Assumption A2 is often referred to as “monotonicity” and states that everyone who would get selected if randomized to treatment would also be selected if randomized to control. This assumption is strong, but may be reasonable in certain settings. For example, in an HIV vaccine trial, it may be reasonable to assume that every person infected in the vaccine arm would also have been infected if randomized to placebo. Similarly, in the Prostate Cancer Prevention Trial, one might be willing to assume that everyone who got cancer in the finasteride arm would have gotten cancer if randomized to placebo. Based on SUTVA, A1, and A2, it is easy to derive that

$$F_1^{11}(y) = F_1(y) \equiv \Pr\{Y_i \leq y | S_i = 1, Z_i = 1\}, \quad (3)$$

where $F_1(\cdot)$ is the distribution of Y among those selected in the treatment arm, and thus $F_1^{11}(\cdot)$ is identified from the observed data. Therefore, the null hypothesis can be rewritten as $H_0: F_1(y) = F_0^{11}(y)$ for all y . Notice that these assumptions also identify $\Pr\{S(1) = 1 | S(0) = 1\} = \Pr\{S(1) = 1\} / \Pr\{S(0) = 1\}$, which is the relative risk (RR) of selection if assigned to treatment over control. In the HIV vaccine context, GBH referred to this quantity as one minus the vaccine efficacy (VE) to prevent infection.

$F_0^{11}(y)$ can be expressed as a biased sampling model

$$F_0^{11}(y) = RR^{-1} \int_{-\infty}^y w(t) dF_0(t), \quad (4)$$

where $dF_0(y)$ is the density of Y for selected subjects randomized to control, which is a mixture of the always infected ($S(0) = S(1) = 1$) and the protected ($S(0) = 1, S(1) = 0$), and $w(y) = \Pr\{S(1) = 1 | Y(0) = y, S(0) = 1\}$ is the conditional relative risk of selection if assigned to treatment given selection and a known potential response y when assigned to placebo. For example, in the context of the vaccine trial, $w(y)$ is the probability that a person infected in the placebo arm would have been infected had they been randomized to the vaccine arm based on their placebo viral load. Clearly, $F_0^{11}(y)$ will be identified if RR , $w(y)$ and $dF_0(y)$ are identified. The density $dF_0(y)$ and RR are identified; however, $w(y)$ is not.

Therefore, to identify the distribution $F_0^{11}(\cdot)$, we need an additional assumption specifying $w(y)$. GBH used the logistic function indexed by the parameter β :

$$w(y; \alpha, \beta) = \frac{\exp(\alpha + \beta y)}{1 + \exp(\alpha + \beta y)}, \quad (5)$$

where α is determined by satisfying $\int w(y; \alpha, \beta) dF_0(y) = RR$, and therefore is identified for a given β . β is a sensitivity parameter which is treated as fixed and known in a single analysis, but then varied over a range of plausible values in a sensitivity analysis. β can be interpreted as a log-odds ratio: given $S(0) = 1$, for a 1-unit increase in $Y(0)$ the odds that $S(1) = 1$ multiplicatively increase e^β (A figure of $w(y; \alpha, \beta)$ as a function of β for some fixed values of α and y is provided in the Supplementary Material.). Note that if $\beta > 0$ and the null hypothesis is true, $F_1(\cdot)$ is stochastically larger than $F_0(\cdot)$; that is, if $\beta > 0$, the distribution of outcomes among selected persons in the treatment arm will be right-shifted relative to those among selected persons in the control arm. Similarly, if $\beta < 0$ and the null hypothesis is true, $F_1(\cdot)$ is stochastically smaller than $F_0(\cdot)$.

2.2. Test Statistics

To test the null hypothesis (2), GBH introduced the test statistic $T_M(\beta)$, which compares means in the always selected principal stratum. Specifically, let N_z be the number of subjects randomized to treatment z , n_z be the number selected in treatment z , and $n = n_0 + n_1$. Define

$\hat{F}_\lambda(y)$ to be the nonparametric estimate of $F_\lambda(y)$ and $\widehat{RR} = \min\left(\frac{n_1 N_0}{n_0 N_1}, 1\right)$. Let $\hat{\alpha}$ solve the equation $\widehat{RR}^{-1} \int w(y; \hat{\alpha}, \beta) d\hat{F}_0(y) = 1$; note that if $\widehat{RR} = 1$ then $\hat{\alpha} = \infty$, and $w(y; \hat{\alpha}, \beta) = 1$ for all y . Let $\hat{F}_0^{11}(y; \beta) = \widehat{RR}^{-1} \int_{-\infty}^y w(t; \hat{\alpha}, \beta) d\hat{F}_0(t)$ the non-parametric maximum likelihood estimate for (4). Therefore, the average causal effect in the always selected principal stratum can be estimated as

$$T_M(\beta) = \int y \{d\hat{F}_0^{11}(y; \beta) - d\hat{F}_1(y)\} = \widehat{RR}^{-1} \frac{1}{n_0} \sum (1 - Z_i) S_i w(Y_i; \hat{\alpha}, \beta) Y_i - \frac{1}{n_1} \sum Z_i S_i Y_i.$$

The test statistic, $T_M(\beta)$ can also be written in an alternative manner, which motivates our rank-based test statistics. Define $Y_i^*(\beta)$ as

$$Y_i^*(\beta) = Y_i - (1 - Z_i) \left[\frac{\sum (1 - Z_i) S_i Y_i}{\sum (1 - Z_i) S_i} - \frac{\sum (1 - Z_i) S_i w(Y_i; \hat{\alpha}, \beta) Y_i}{\sum (1 - Z_i) S_i w(Y_i; \hat{\alpha}, \beta)} \right],$$

where $Y_i^*(\beta) = *$ if $S_i = 0$ (and therefore $Y_i = *$). $Y_i^*(\beta)$ was introduced by Mehrotra *et al.* [9] and adjusts the outcome among those selected in the control arm by subtracting the difference between the mean outcome under control and the mean outcome under control in the always selected stratum. Thus the average of $Y_i^*(\beta)$ among those selected in the control arm equals the average of Y_i among those randomized to control in the always selected

stratum. The appendix contains a brief proof demonstrating that $T_M(\beta)$ can therefore be written as

$$T_M(\beta) = \frac{1}{n_0} \sum (1 - Z_i) S_i Y_i^*(\beta) - \frac{1}{n_1} \sum Z_i S_i Y_i^*(\beta),$$

the mean difference between the adjusted outcomes.

Rather than use outcomes or adjusted outcomes, one might want to compare corresponding ranks of these quantities. Let R_i be the pooled rank for Y_i for $i = 1, \dots, N$; let $R_i^*(\beta)$ be the pooled rank of $Y_i^*(\beta)$; and define R_i^{**} as follows:

$$R_i^{**}(\beta) = (1 - Z_i) \frac{w(Y_i; \hat{\alpha}, \beta) R_i}{\widehat{RR}} + Z_i R_i.$$

If $Y_i = *$ then $R_i = R_i^*(\beta) = R_i^{**}(\beta) = *$ (does not receive a rank). $R_i^{**}(\beta)$ can be thought of as a weighted rank of the pooled sample, where the ranks for those with $Z = 0$ are up- or down-weighted according to their probability of being in the always selected stratum. In contrast, $R_i^*(\beta)$ corresponds to the ranks of the outcome after its distribution in the control arm has been shifted so that it is centered at the estimated mean in the always selected stratum.

Our rank-based test statistics are then defined in a manner analogous to $T_M(\beta)$ except replacing outcomes and adjusted outcomes with ranks and adjusted ranks, respectively:

$$T_1(\beta) = \frac{1}{n_0} \sum (1 - Z_i) S_i R_i^{**}(\beta) - \frac{1}{n_1} \sum Z_i S_i R_i^{**}(\beta),$$

$$T_2(\beta) = \frac{1}{n_0} \sum (1 - Z_i) S_i R_i^*(\beta) - \frac{1}{n_1} \sum Z_i S_i R_i^*(\beta).$$

These test statistics therefore correspond to differences between mean ranks in the always selected principal stratum.

As a final test statistic we consider a variation of the Fligner-Policello test [19] to allow for unequal variances of the distributions being compared. Let

$$P_i^*(\beta) = S_i \sum_{j=1}^N S_j I\{Z_j \neq Z_i\} \left[I\{Y_j^*(\beta) < Y_i^*(\beta)\} + \frac{1}{2} I\{Y_j^*(\beta) = Y_i^*(\beta)\} \right],$$

where $I\{\cdot\}$ designates the indicator function. $P_i^*(\beta)$ is simply the number of outcomes in the other trial arm less than $Y_i^*(\beta)$ plus half the number that are equal. Define

$\overline{P_0^*}(\beta) = \frac{1}{n_0} \sum (1 - Z_i) S_i P_i^*(\beta)$, $\overline{P_1^*}(\beta) = \frac{1}{n_1} \sum Z_i S_i P_i^*(\beta)$, $V_{P_0} = \sum (1 - Z_i) S_i (P_i^*(\beta) - \overline{P_0^*}(\beta))^2$, and $V_{P_1} = \sum Z_i S_i (P_i^*(\beta) - \overline{P_1^*}(\beta))^2$. Then the Fligner-Policello test statistic is

$$T_{FP}(\beta) = \frac{n_0 \overline{P_0^*}(\beta) - n_1 \overline{P_1^*}(\beta)}{2(V_{P_0} + V_{P_1} + \overline{P_0^*}(\beta) \overline{P_1^*}(\beta))^{1/2}}.$$

All three of these new test statistics compare ranks of the outcome after accounting for post-randomization selection. $T_2(\beta)$ and $T_{FP}(\beta)$ apply the selection weights, $w(\cdot)$, on the original scale and then rank; in contrast, $T_1(\beta)$ first ranks the data and then applies the selection weights to the ranks. Notice that the selection probability, $w(\cdot)$, is a function of Y , not the rank of Y .

2.3. Bootstrap Approaches

We use the nonparametric bootstrap to compute p -values for our test statistics, and we will examine three different approaches: the first is essentially a standard bootstrap and the latter two re-sample data under the null. For ease of presentation, we define $\mathbf{Y}_Z = \{Y_{Z1}, Y_{Z2}, \dots, Y_{Zn_Z}\}$ as the set of all Y such that $Z = z$ and $S = 1$. Our application of each of the three bootstraps involves the following steps for each of N_{boot} bootstrap replications:

1. Generate $n_1^* \sim \text{Bin}(n, \hat{p})$ and let $n_0^* = n - n_1^*$, where $\hat{p} = n_1/n$.
2. Sample with replacement n_0^* values $\mathbf{Y}^*$ from $(\mathbf{Y}_1, \mathbf{Y}_2)$ with probability $(\mathbf{P}_1, \mathbf{P}_2)$ and assign $Z^* = 0$, $S^* = 1$.
3. Sample with replacement n_1^* values $\mathbf{Y}^*$ from $(\mathbf{Y}_1, \mathbf{Y}_2)$ with probability $(\mathbf{Q}_1, \mathbf{Q}_2)$ and assign $Z^* = 1$, $S^* = 1$.
4. Assign $N_0 - n_0^*$ records $Z^* = 0$, $S^* = 0$ and $N_1 - n_1^*$ records $Z^* = 1$, $S^* = 0$. That is, bootstrap only resamples data from selected patients. Non-selected patients are not considered in the bootstrap calculations.
5. Compute bootstrap test statistics $T_{\{\cdot\}}^*(\beta)$.

Steps 2–3 differ across our three bootstrap approaches and are detailed below. Note that this general bootstrap procedure fixes the number of selection events n , as well as N_0 and N_1 across bootstrap replications. Such a bootstrap procedure mimics an event-driven trial, and has been seen to produce similar results to bootstraps where only N is fixed [20].

Simple Bootstrap—In the first bootstrap approach, $\mathbf{Y}^*$ is generated in a standard manner:

- Sample with replacement n_0^* values $\mathbf{Y}^*$ with equal probability from $\mathbf{Y}_0$.
- Sample with replacement n_1^* values $\mathbf{Y}^*$ with equal probability from $\mathbf{Y}_1$.

A two-sided p -value is then computed as $2 \times$ the minimum of the proportion of $T_{\{\cdot\}}^*(\beta) > 0$ or the proportion of $T_{\{\cdot\}}^*(\beta) < 0$. A one-sided p -value can be computed as the proportion of

$T_{\{.\}}^*(\beta) > 0$ (or < 0 , depending on the direction of the null hypothesis). Note that this bootstrap procedure essentially computes a confidence interval for the test statistic and then determines whether this confidence interval excludes zero. This approach works because under H_0 , the distributions of the outcome in the always selected stratum are the same for $Z = 0$ and $Z = 1$, so each test statistic is approximately 0 under the null.

Controls-Only Bootstrap—The second bootstrap approach, which we refer to as the *controls-only bootstrap* was introduced by Hudgens *et al.* [17] and employed by GBH. The basic idea is to generate data under the null by sampling outcomes for the treatment arm from the control arm in a weighted manner such that they are drawn from the distribution of the always selected under control. Specifically,

- Sample with replacement n_0^* values Y^* with equal probability from $\mathbf{Y}_0$.
- Sample with replacement n_1^* values Y^* from $\mathbf{Y}_0$ with probability $w(\mathbf{Y}_0; \hat{\alpha}, \beta) / (n_0 \widehat{RR})$.

A two-sided p-value is then computed as the proportion with $|T_{\{.\}}^*(\beta)| \geq |T_{\{.\}}(\beta)|$. One-sided p-values are computed as the proportion with $T_{\{.\}}^*(\beta)$ greater (less) than or equal to $T_{\{.\}}(\beta)$, depending on the direction of the null. We see two potential disadvantages to this bootstrap approach. First, this approach ignores any outcomes in the treatment arm. This could be particularly undesirable in settings where there are relatively few selected subjects in the control arm (e.g., trials where N_1 is substantially greater than N_0 and $\widehat{RR}$ close to one). Second, for those selected in the treatment arm in the bootstrap sample, Y^* is sampled based on the estimated weights $w(\mathbf{Y}_0; \hat{\alpha}, \beta)$ which are fixed across bootstrap replications. Therefore, even though α is re-estimated within each bootstrap sample in the process of computing $T_{\{.\}}^*(\beta)$, the sampling of Y^* ignores variation in the estimation of $\hat{\alpha}$.

This latter problem can be overcome with an additional bootstrap procedure that first obtains the distribution of $\hat{\alpha}$, and then uses a sample from this distribution in computations. Specifically, perform steps 1–4 of the *simple bootstrap*, estimate $\hat{\alpha}$ from the bootstrap data, and repeat many (say 500) times to get a distribution of $\hat{\alpha}$. Then perform steps 1–5 of the *controls-only bootstrap*, in each iteration drawing a value of $\hat{\alpha}^*$ from the bootstrapped distribution of $\hat{\alpha}$ and then plugging $\hat{\alpha}^*$ into step 3 instead of $\hat{\alpha}$. We refer to this as the *augmented controls-only-bootstrap*.

Treatment-and-Control Bootstrap—A third bootstrap procedure, which we refer to as the *treatment-and-control bootstrap*, is a modification of another bootstrap approach proposed by Hudgens *et al.* [17]. The *treatment-and-controls bootstrap* similar to the *controls-only bootstrap* except that data are generated under the null by sampling with replacement in a weighted manner from both the treatment and control arms. Specifically,

- Sample with replacement n_0^* values Y^* from $(\mathbf{Y}_0, \mathbf{Y}_1)$ with probability $(\mathbf{P}_0, \mathbf{P}_1)$, where $\mathbf{P}_0 = 1/n$ and $\mathbf{P}_1 = (n_1/w(\mathbf{Y}_1; \hat{\alpha}, \beta)) / (n \sum (1/w(\mathbf{Y}_1; \hat{\alpha}, \beta)))$.

- Sample with replacement n_1^* values Y^* from (Y_0, Y_1) with probability (Q_0, Q_1) , where $Q_0 = n_0 w(Y_0; \hat{\alpha}, \beta) / (n \sum w(Y_0; \hat{\alpha}, \beta))$ and $Q_1 = 1/n$.

Once test statistics have been computed on the bootstrap data, p-values are computed in an identical manner to that described above for the controls-only bootstrap. This bootstrap procedure has the advantage over the controls-only bootstrap in that it incorporates outcomes from both arms. However, if $w(Y_1; \hat{\alpha}, \beta)$ is close to zero, this approach could be problematic. And similar to the controls-only bootstrap, this approach ignores variation in $\hat{\alpha}$. Again, the latter problem can be avoided by performing an *augmented treatment-and-control bootstrap*, in a manner analogous to that described above for the *controls-only bootstrap*. Note that if $\beta = 0$ this bootstrap approach is equivalent to a permutation test among selected subjects that accounts for uncertainty in the number of selected events in each arm. However, if $\beta \neq 0$, a simple permutation test among the selected subjects would be incorrect.

3. Simulations

3.1. Simulation Procedure

We investigated the performance of the test statistics and bootstrap procedures through a series of simulations. Each simulation was comprised of 1000 simulated data sets. Each data set was simulated as if it came from a trial with an equal number of subjects assigned to treatment and control, and the number of selection events, n , equal to 100. Data were generated in the following manner:

1. Generate $n_1 \sim \text{Bin}(n, p)$ and let $n_0 = n - n_1$, where $p = RR/(1 + RR)$.
2. Generate n_1 values Y from $F_1^{11}(y + \Delta)$, as specified below, and assign $Z = 1, S = 1$.
3. Generate n_0 values Y from $F_0(y, \alpha, \beta)$, as specified below, and assign $Z = 0, S = 1$.

Data were generated such that in the always-selected principal stratum, $Y(1)$ and $Y(0)$ were identically distributed as one of the following:

- normal with mean=4.5 and variance=0.74², $N(4.5, 0.74^2)$
- stochastic mixture of two normal distributions: $0.8N(4.3, 0.6^2) + 0.2N(5.3, 1)$
- 2 + gamma distribution with shape=3, scale= 2.5/3
- $(4.5 - 0.98) + \text{exponential distribution with rate}=1/0.98$

These distributions represent varying levels of skewness and kurtosis, and parameter values were chosen to approximately resemble the distribution of log-transformed viral load measurements in HIV infected subjects. (A figure showing these densities is in the Supplementary Material.) Furthermore, if $Y < \log_{10}(50)$, then Y was set to equal $\log_{10}(50)$ as the viral load assay has a lower quantification limit of 50. To generate n_0 values of Y from $F_0(y, \alpha, \beta)$, we incorporated the biased sampling models of Section 2.1, drawing Y from

$$dF_0(y; \alpha, \beta) = \frac{w^{-1}(y; \alpha, \beta) dF_0^{11}(y)}{\int w^{-1}(y; \alpha, \beta) dF_0^{11}(y)},$$

where $w(\cdot)$ was as defined in (5). Data were generated with $RR \in \{0.4, 0.7\}$, $\beta \in \{-1, 0, 1\}$, and α such that $\int w(y; \alpha, \beta) dF_0(y) = RR$. The mean shift parameter, μ , was set to 0 under the null and 0.5 under the alternative except for the gamma distribution where $\mu = 0.8$ under the alternative. In summary, we generated data under $4 \times 2 \times 3 \times 2 = 48$ simulation scenarios, corresponding to all combinations of the different choices for $F_0^{11}(y)$, RR , β , and μ , respectively. Finally, in an additional set of simulations to investigate performance with large amounts of selection bias, we also considered $\beta \in \{-3, 3\}$ for $\mu = 0$ and $Y(0)$ normally distributed.

In all simulated data sets, we computed each of the test statistics given in Section 2.2 with β properly specified. One- and two-sided p-values were computed using each of the 3 bootstrap approaches of Section 2.3, each consisting of 3000 bootstrap replications. All model fitting steps for all tests (including the adaptive test discussed later) were performed within each bootstrap replication. The null was rejected for one- and two-sided tests if the p-value was less than 0.025 and 0.05, respectively. Simulation code is given at <http://biostat.mc.vanderbilt.edu/RankBasedGBH>.

3.2. Simulation Results

We first considered our simulations under the null hypothesis. Each bootstrap procedure resulted in approximately the nominal type I error rate of 0.05 when $\beta = 0$, i.e., when data were generated with no selection bias. However, the controls-only and the treatment-and-control bootstraps that ignored the estimation of α had inflated type I error rates as β deviated farther from 0. For example, when data were generated using a normal distribution with $RR = 0.7$ (Figure 1), the type I error rate of 2-sided tests was greater than 10% for all test statistics using the controls-only bootstrap and the treatment-and-control bootstrap when $\beta = -3$. However, type I error rates were more conservative – perhaps too conservative – when accounting for variation in $\hat{\alpha}$ by performing the augmented controls-only and augmented treatment-and-control bootstraps.

Because the controls-only and treatment-and-control bootstraps could not preserve the type I error rate, we focus on simulation results using the simple bootstrap and the augmented versions of the controls-only and treatment-and-control bootstraps. Figure 2 shows the type I error rate for the different test statistics and distributions considered with $RR = 0.7$ and $\beta = 1$. The type I error rate was controlled using the simple bootstrap for all test statistics and all distributions. The other bootstraps tended to be somewhat conservative under the null. Other simulation results under the null, including results for one-sided tests, are reported in the Supplementary Material. These results were fairly similar, except type I error rates for $T_2(\beta)$ and $T_{FP}(\beta)$ were inflated when $\beta = -1$ under the gamma and exponential distributions.

Figure 3 shows the power for the different test statistics and bootstrap procedures when data were simulated with $RR = 0.7$ and $\beta = 1$. Test statistics using the simple bootstrap were more

powerful than their counterparts using the augmented controls-only and treatment-and-control bootstraps. Similar patterns were also observed with other values of RR and β (see Supplementary Material). Given the conservative nature of the augmented controls-only and treatment-and-control bootstraps, we will hereafter focus our attention on the simple bootstrap.

Figure 4 shows the power of the different test statistics under different distributions using the simple bootstrap and $RR = 0.7$. Not surprisingly, under normality, $T_M(\beta)$ was more powerful than the rank-based test statistics. However, this improvement in power was not substantial (0.829 vs. 0.801 for $T_1(\beta)$). As expected, the rank-based statistics tended to be more powerful than $T_M(\beta)$ when the data were not normal, with the improvement being greater as the data became more skewed (going from mixture normal, to gamma, to exponential). For example, with the exponential distribution and $\beta = 1$, the power of $T_1(\beta)$ was 0.908 compared to 0.650 for $T_M(\beta)$. For $\beta = 0$ and $\beta = 1$, it is difficult to distinguish between any of the rank-based statistics. However, with $\beta = -1$, $T_1(\beta)$ clearly outperformed $T_2(\beta)$ and $T_{FP}(\beta)$. Given this fact, similar results when $RR = 0.4$, and the inability of $T_2(\beta)$ and $T_{FP}(\beta)$ to control the type I error rate under certain simulation settings, our simulations favor using $T_1(\beta)$ as the rank-based test statistic of choice.

Figure 4 also shows results using an additional, adaptive test, $T_{adapt}(\beta)$. Specifically, in each simulation we tested the distribution of $Y|Z = 1$, $S = 1$ for normality using a Shapiro-Wilk test. If normality was rejected ($p < 0.01$) then our adaptive test statistic was set as $T_1(\beta)$, otherwise $T_M(\beta)$. This process was repeated in each bootstrap replication. When data were non-normal, this $T_{adapt}(\beta)$ tended to perform similarly to $T_1(\beta)$; when data were normal, it performed similarly to $T_M(\beta)$. By bootstrapping the entire model selection procedure, the type I error rate for this adaptive test statistic was preserved (see Supplementary Material).

In summary, our simulations demonstrate that the type I error rate and power can be highly dependent on the type of bootstrap and the choice of test statistic. The simple bootstrap generally appears to be a safe choice, although in some settings the augmented treatment-and-control bootstrap may be better. The test statistic based on the expectation of untransformed data, $T_M(\beta)$, tended to be most powerful under normality. With greater departures from normality, the rank-based test statistics performed better, and among rank-based test statistics, $T_1(\beta)$ generally outperformed $T_2(\beta)$ and $T_{FP}(\beta)$.

4. Data Examples

4.1. HIV Vaccine Trial

We first illustrate the different rank-based test statistics by applying them to data from an HIV vaccine trial, VaxGen's trial of AIDSVAX B/B. Detailed results of this trial have been reported elsewhere [21, 22]. In short, this was a double blinded trial that randomized approximately 5,400 HIV-negative volunteers (engaging in activities that increased their risk of HIV acquisition) to vaccine or placebo in a 2:1 ratio. The vaccine was found not to be effective at preventing HIV infection, although subgroup analyses suggested that the vaccine may have been partially protective against infection for non-whites. Of 914 non-whites in the trial, 30/604 (5.0%) randomized to vaccine became infected whereas 29/310 (9.4%)

became infected in placebo. Our goal is to determine whether the vaccine had an effect on the viral load among non-white volunteers who would have been infected regardless of treatment assignment. This question has been addressed with these data in earlier work using the approach of GBH [23]; here we re-examine the question using our rank-based test statistics. In the notation given above, $Z = 0, 1$ implies randomization to placebo, vaccine; S is the indicator of infection; and Y is the viral load if infected. SUTVA, A1, and A2 are all thought to be reasonable in this trial [12].

The distribution of $\log_{10}$ -transformed viral loads for those infected in the vaccine and placebo arms are shown in the left-panel of Figure 5. The right panel shows two-sided p-values testing the null hypothesis that the distribution of viral loads is the same between the vaccine and placebo arms among those who would have been infected regardless of treatment assignment. P-values, based on 10,000 bootstrap replications, are shown as a function of the sensitivity parameter, β , which is interpreted as follows: given someone was infected in the placebo arm, for a $1 - \log_{10}$ increase in viral load, the odds that they would have been infected had they been randomized to vaccine multiplicatively increases by e^β . Therefore, $\beta = 3$ corresponds to a very large odds ratio ≈ 20 . In our past experience eliciting this sensitivity parameter, most subject-matter experts have believed that β was non-negative, with an odds ratio (e^β) ranging between 1 and 5 [23]. An odds ratio greater than 1 suggests that infected subjects with high viral loads in the placebo arm are those who would be more likely infected had they been randomized to the vaccine arm.

Results were nearly identical for $T_1(\beta)$ and $T_M(\beta)$ using the simple bootstrap. For $\beta = 1, 1.25$, and 1.5 , the p-values for $T_1(\beta)$ were 0.060, 0.042, and 0.037 whereas for $T_M(\beta)$ they were 0.060, 0.044, and 0.038, respectively. Consistent with our simulations, p-values for $T_2(\beta)$ and $T_{FP}(\beta)$ were similar to each other and slightly higher than the other two test statistics: for $\beta = 1, 1.25$, and 1.5 , the p-values for $T_2(\beta)$ were 0.080, 0.059, and 0.050 and for $T_{FP}(\beta)$ they were 0.079, 0.059, and 0.050, respectively. These results using $T_M(\beta)$ or $T_1(\beta)$ suggest that the sensitivity parameter odds ratio would have to be at least $\exp(1.25) \approx 3.5$ for there to be evidence that the vaccine lowers viral loads for non-whites who would have been infected regardless of treatment assignment.

4.2. Prostate Cancer Prevention Trial

Our second illustration comes from the Prostate Cancer Prevention Trial (PCPT), a double blinded trial that randomized 18,882 men aged 55 years or older to daily doses of finasteride or placebo for 7 years of follow-up. Trial details are given elsewhere [2]. Our goal is to determine whether finasteride had an effect on the severity of prostate cancer among participants who would have gotten prostate cancer regardless of their treatment assignment. Among those with prostate cancer, its severity was measured using the Gleason score, an ordered categorical variable assigned the integer values 2–10, with 10 being the most severe. A previous analysis of this data used principal stratification techniques but dichotomized severe cancer as a Gleason score ≥ 7 [16]; in this re-analysis we use raw, non-dichotomized Gleason scores. Z is the indicator of being randomized to finasteride, S is the indicator of getting prostate cancer, and Y is the Gleason score. We will limit our analysis to the 10,031

men with non-missing S and Y . (A large portion of patients did not receive an end-of-study biopsy necessary to determine their cancer status.)

Prostate cancer was detected in 769/4899 (15.7%) in the finasteride arm compared with 1109/5132 (21.6%) in the placebo arm, suggesting that finasteride lowered the risk of prostate cancer. The distribution of Gleason scores among those who got cancer in both arms is shown in the left panels of Figure 6.

The right panel of Figure 6 shows the p-value using $T_M(\beta)$ and $T_1(\beta)$ with the simple bootstrap (based on 10,000 bootstrap replications) as a function of the sensitivity parameter β , which in this example is interpreted as follows: given someone got prostate cancer in the placebo arm, for a 1-unit increase in Gleason score, the odds that they would have gotten prostate cancer had they been randomized to finasteride arm multiplicatively increases by e^β . For $\beta = 1.25$, p-values are very similar and less than 0.05; as β increases, p-values become more disparate but lead to similar conclusions. These results suggest that unless the odds of getting cancer if randomized to finasteride are $> \exp(1.25) \approx 3.5$ per 1-category increase in Gleason score given someone got cancer in the placebo arm, there is evidence that finasteride causes higher Gleason scores among those who were doomed to get cancer regardless of their treatment assignment. These results are consistent with prior analyses of these data suggesting finasteride causes more severe cancers among those doomed to get cancer [16], although subsequent investigation has suggested that this finding may be an artifact of the technique used to compute Gleason scores [24, 25]. Notice that $T_M(\beta)$ compares mean Gleason scores in the group doomed to get cancer regardless of treatment assignment, whereas $T_1(\beta)$ compares the mean rank of Gleason scores, which may be more appropriate given the ordinal nature of Gleason scores.

It is important to note that Figure 6 does not contain the p-values for $T_2(\beta)$ and $T_{FP}(\beta)$ because they are not useful in this setting, and almost always zero. As a reminder, both $T_2(\beta)$ and $T_{FP}(\beta)$ are based on ranks computed for the adjusted outcomes, in which outcomes among those selected in the control arm are reduced by the difference between the mean outcome under control and the mean outcome under control in the always selected stratum. The reason for their lack of utility is because of the large number of ties in Gleason scores. Consider the definition of $Y_i^*(\beta)$ given in Section 2.2: $Y_i^*(\beta) = Y_i$ for those in the finasteride arm, whereas $Y_i^*(\beta) = Y_i - K(\beta)$ for those in the placebo arm, where for a given β , $K(\beta)$ is a constant. From Figure 6 we can see that there were lots of men who got cancer with a Gleason score of 6, for example. $Y_i^*(\beta)$ for each of these men in the finasteride arm remains the same; however $Y_i^*(\beta)$ is shifted by a constant $K(\beta)$ for all placebo participants who had a Gleason score of 6, regardless of the size of $K(\beta)$. Therefore, whenever $K(\beta)$ is non-zero (which always is the case unless $\beta = 0$), when ranks are applied to $Y_i^*(\beta)$, there is a huge shift in the distribution of the ranks between the finasteride and placebo arms, which leads to strong differences between groups. This is not a problem when using $T_1(\beta)$ because small values of $K(\beta)$ only result in small mean differences of $Y_i^*(\beta)$ between the finasteride and placebo arms.

Because those participants with missing data were excluded from the analysis, our results may be biased unless data are missing completely at random [26], which is unlikely. These rank-based approaches could be extended using covariates and inverse-probability weighting techniques in a manner similar to that described elsewhere [16], to account for differences between those with and without end-of-study biopsies. Other approaches to account for data that are not missing completely at random could also be applied. Although scientifically important, such analyses are beyond the scope of this manuscript.

5. Discussion

We have considered several rank-based statistics and bootstrap approaches for performing sensitivity analyses of treatment effects in the always selected principal stratum. Our simulations and data analysis experience suggest that $T_1(\beta)$ is the preferred rank-based statistic; it was about as powerful as $T_{MM}(\beta)$ under normality, but notably more powerful under non-normality. In many settings it may be preferable to use an adaptive test statistic, where an initial test for normality is performed and then $T_1(\beta)$ is applied if normality is rejected, otherwise $T_{MM}(\beta)$ (bootstrapping the entire model selection procedure). Our results also suggest that a controls-only bootstrap can have inflated type I error; we suggest using the simple bootstrap procedure to compute significance levels.

All of our rank-based test statistics incorporate the original data into the selection probability $w(\cdot)$. One could consider making the probability of selection a function of the ranks, not Y . However, such an approach does not seem as scientifically reasonable, and it makes it more difficult to select the range β over which to perform the sensitivity analysis. More generally, results using rank-based approaches – not just for $w(\cdot)$ but for the final test statistic – tend to be more difficult to interpret.

The rank-based methods described here are applicable to principal stratification approaches for dealing with ‘truncation by death.’ $T_1(\beta)$ (and $T_{MM}(\beta)$) with the simple bootstrap can be easily implemented using the R package *sensitivityPStrat*.

Principal stratification approaches have recently come under criticism [27–30]. Broadly stated, in the context of ‘truncation by death,’ the primary criticisms are that principal stratum treatment effects 1) do not improve decision making because principal stratum membership is not known, 2) focus on an effect of secondary importance given the primary importance of treatment effect on selection (i.e., infection, cancer, or death), and 3) have limitations when applied to time-dependent selection events. These concerns are valid and should be considered when applying principal stratification methods such as those described herein. However, studies of this nature are inherently difficult, and we are unaware of a ‘silver bullet’ approach that fully addresses the ‘truncation by death’ question. We recommend the use of principal stratification methods in conjunction with other approaches (e.g., [31]) to better understand treatment effects. To this end, the rank-based approaches described in this manuscript are important methods to add to an analyst’s tool kit.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the associate editor and two referees for several suggestions that helped improve the manuscript.

Appendix

Notice that

$$Y_i^*(\beta) = Y_i - (1 - Z_i) \left\{ \frac{\sum (1 - Z_i) S_i Y_i}{\sum (1 - Z_i) S_i} - \frac{\sum (1 - Z_i) S_i w(Y_i; \hat{\alpha}, \beta) Y_i}{\sum (1 - Z_i) S_i w(Y_i; \hat{\alpha}, \beta)} \right\}$$

The equality of

$$T_M(\beta) = \frac{1}{n_0} \sum (1 - Z_i) S_i Y_i^*(\beta) - \frac{1}{n_1} \sum Z_i S_i Y_i^*(\beta) = \widehat{RR}^{-1} \frac{1}{n_0} \sum (1 - Z_i) S_i w(Y_i; \hat{\alpha}, \beta) Y_i - \frac{1}{n_1} \sum Z_i S_i Y_i.$$

immediately follows from the following results:

1. $Z_i = 0$ or 1. So $Z_i S_i Y_i^*(\beta) = Z_i S_i Y_i$ and

$$\sum (1 - Z_i) S_i \left\{ Y_i - (1 - Z_i) \frac{\sum (1 - Z_i) S_i Y_i}{\sum (1 - Z_i) S_i} \right\} = 0.$$

2. $n_0 = \sum (1 - Z_i) S_i$ and

$$\frac{1}{\widehat{RR}} \int w(y; \hat{\alpha}, \beta) d\hat{F}_0(y) = \frac{1}{(n_0 \widehat{RR})} \sum (1 - Z_i) S_i w(Y_i; \hat{\alpha}, \beta). \text{ So}$$

$$\sum (1 - Z_i) S_i \frac{\sum (1 - Z_i) S_i w(Y_i; \hat{\alpha}, \beta) Y_i}{\sum (1 - Z_i) S_i w(Y_i; \hat{\alpha}, \beta)} = \widehat{RR}^{-1} \sum (1 - Z_i) S_i w(Y_i; \hat{\alpha}, \beta) Y_i.$$

References

1. Buchbinder S, Mehrotra D, Duerr A, Fitzgerald DW, Mogg R, Li D, Gilbert PB, Lama JR, Marmor M, Del Rio C, McElrath MJ, Casimiro DR, Gottesdiener KM, Chodakewitz JA, Corey L, Robertson MN. Team. SSP. Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the step study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *Lancet*. 2008; 372:1881–93. [PubMed: 19012954]
2. Thompson I, Goodman P, Tangen C, Lucia M, Miller G, Ford L, Lieber M, Cespedes R, Atkins J, Lippman S, Carlin S, Ryan A, CM S, Crowley J, Coltman C. The influence of finasteride on the development of prostate cancer. *The New England Journal of Medicine*. 2003; 349:213–222. [PubMed: 12867605]
3. Rubin D. Comment on ‘causal inference without counterfactuals’ by AP Dawid. *Journal of the American Statistical Association*. 2000; 95:435–437.
4. Zhang J, Rubin D. Estimation of causal effects via principal stratification when some outcomes are truncated by death. *Journal of Educational and Behavioral Statistics*. 2003; 28:353–368.
5. Hayden D, Pauler D, Schoenfeld D. An estimator for treatment comparisons among survivors in randomized trials. *Biometrics*. 2005; 61:305–310. [PubMed: 15737107]
6. Rosenbaum P. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A*. 1984; 147:656–666.

7. Gilbert P, Bosch R, Hudgens M. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*. 2003; 59:531–541. [PubMed: 14601754]
8. Frangakis C, Rubin D. Principal stratification in causal inference. *Biometrics*. 2002; 58:21–29. [PubMed: 11890317]
9. Mehrotra D, Li X, Gilbert P. A comparison of eight methods for the dual-endpoint evaluation of efficacy in a proof-of-concept HIV vaccine trial. *Biometrics*. 2006; 62:893–900. [PubMed: 16984333]
10. Hudgens M, Halloran M. Causal effects on binary postinfection outcomes. *Journal of the American Statistical Association*. 2006; 101:51–64. [PubMed: 19096723]
11. Shepherd B, Gilbert P, Lumley T. Sensitivity analyses comparing time-to-event outcomes only existing in a subset selected post-randomization. *Journal of the American Statistical Association*. 2007; 102:573–582. [PubMed: 19122791]
12. Shepherd B, Gilbert P, Jemai Y, Rotnitzky A. Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics*. 2006; 62:332–342. [PubMed: 16918897]
13. Jemai Y, Rotnitzky A, Shepherd B, Gilbert P. Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs. *Journal of the Royal Statistical Society, Series B*. 2007; 69:879–901.
14. Sjolander A, Humphreys K, Vansteelandt S, Bellocco R, Palmgren J. Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. *Biometrics*. 2009; 65:514–520. [PubMed: 18759834]
15. Jemai Y. Unpublished Ph.D. dissertation under the supervision of A. Rotnitzky. Harvard School of Public Health, Department of Biostatistics; 2005. Semiparametric methods for inferring treatment effects on outcomes defined only if a post-randomization event occurs.
16. Shepherd B, Redman M, Ankerst D. Does finasteride affect the severity of prostate cancer? A causal sensitivity analysis. *Journal of the American Statistical Association*. 2008; 103:1392–1404.
17. Hudgens M, Hoering A, Self S. On the analysis of viral load endpoints in HIV vaccine trials. *Statistics in Medicine*. 2003; 22:2281–98. [PubMed: 12854093]
18. Rubin D. Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*. 1978; 6:34–58.
19. Fligner M, Policello G. Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association*. 1981; 76:162–168.
20. Shepherd B, Gilbert P, Dupont C. Sensitivity analyses comparing time-to-event outcomes only existing in a subset selected postrandomization and relaxing monotonicity. *Biometrics*. 2011; 67:1100–10. [PubMed: 21114663]
21. Flynn N, Forthal D, Harro C, Judson F, Mayer K, Para M. rgp120 HIV Vaccine Study Group. Placebo-controlled trial of a phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *Journal of Infectious Diseases*. 2005; 191:647–649. [PubMed: 15688276]
22. Gilbert P, Ackers M, Berman P, Francis D, Popovic V, Hu D, Heyward W, Sinangil F, Shepherd B, Gurwith M. Hiv-1 virologic and immunologic progression and antiretroviral therapy initiation among HIV-1-infected participants in an efficacy trial of recombinant glycoprotein 120 vaccine. *Journal of Infectious Diseases*. 2005; 192:974–983. [PubMed: 16107949]
23. Shepherd B, Gilbert P, Mehrotra D. Eliciting a counterfactual sensitivity parameter. *The American Statistician*. 2007; 61:56–63.
24. Cohen YC, Liu KS, Heyden NL, Carides AD, Anderson KM, Daifotis AG, Gann PH. Detection bias due to the effect of finasteride on prostate volume: a modeling approach for analysis of the Prostate Cancer Prevention Trial. *Journal of the National Cancer Institute*. 2007; 99:1366–74. [PubMed: 17848668]
25. Redman MW, Tangen CM, Goodman PJ, Lucia MS, Coltman CAJ, Thompson IM. Finasteride does not increase the risk of high-grade prostate cancer: a bias-adjusted modeling approach. *Cancer Prevention Research*. 2008; 1:174–81. [PubMed: 19138953]
26. Little, RJA., Rubin, DB. *Statistical Analysis with Missing Data*, Second Edition. Wiley; New York, NY, USA: 2002.

27. Robins J, Rotnitzky A, Vansteelandt S. Discussion of ‘Principal stratification designs to estimate input data missing due to death’ by Frangakis CE, Rubin DB, An M, MacKenzie E. *Biometrics*. 2007; 63:650–654. [PubMed: 17824996]
28. Pearl J. Principal stratification – a goal or a tool? *International Journal of Biostatistics*. 2011; 7:20. [PubMed: 21556288]
29. Sjolander A. Reaction to Pearl’s critique of principal stratification. *International Journal of Biostatistics*. 2011; 7:22.
30. Joffe M. Principal stratification and attribution prohibition: good ideas taken too far. *International Journal of Biostatistics*. 2011; 7:35.
31. Follmann D, Fay M, Proschan M. Chop-lump tests for vaccine trials. *Biometrics*. 2009; 65:885–893. [PubMed: 19210727]

Figure 1.

Type I error rates based on 1000 datasets simulated under normality for the different test statistics and bootstrap procedures. All tests were two-sided. Dotted lines are at the nominal 0.05-level ± 0.0135 (the Monte Carlo margin of error, $1.96 \sqrt{0.05(1 - 0.05)/1000}$). Bootstraps 1, 2, 3, 2a, and 3a refer to the simple, controls-only, treatment-and-control, augmented controls-only, and augmented treatment-and-control bootstraps, respectively.

Figure 2.

Type I error rates based on 1000 datasets simulated under different distributions with $RR = 0.7$ and $\beta = 1$. All tests were two-sided. Dotted lines are at the nominal 0.05-level ± 0.0135

(the Monte Carlo margin of error, $1.96 \sqrt{0.05(1 - 0.05)/1000}$). Bootstraps 1, 2a, and 3a refer to the simple, augmented controls-only, and augmented treatment-and-control bootstraps, respectively.

Figure 3.

Power based on 1000 datasets simulated under different distributions with $RR = 0.7$ and $\beta = 1$. All tests were two-sided. Bootstraps 1, 2a, and 3a refer to the simple, augmented controls-only, and augmented treatment-and-control bootstraps, respectively.

Figure 4.

Power based on 1000 datasets simulated under different distributions with $RR = 0.7$ and using the simple bootstrap. All tests were two-sided.

Figure 5.

Left plots: Histograms of log₁₀-transformed viral load among non-whites infected in the vaccine and placebo arms of the VaxGen trial. The vertical line represents the mean. Right plot: Sensitivity analysis of the test of association between vaccination and viral load among those who would have been infected regardless of treatment assignment. β represents the log-odds of infection in vaccine given infection in placebo and placebo viral load. All tests were two-sided.

Figure 6.

Left plots: Histograms of Gleason scores among men diagnosed with prostate cancer in the finasteride and placebo arms of the PCPT. The vertical line represents the mean. Right plot: Sensitivity analysis of the test of association between finasteride and Gleason score among those who would have gotten cancer regardless of treatment assignment. β represents the log-odds of getting cancer if randomized to finasteride given cancer when randomized to placebo and based on placebo Gleason score. All tests were two-sided.