2018

Generalizing Evidence from Randomized Trials using Inverse Probability of Sampling Weights

Ashley L. Buchanan
University of Rhode Island, buchanan@uri.edu

Michael G. Hudgens

Stephen R. Cole

Katie R. Mollan

Paul E. Sax

See next page for additional authors

Citation/Publisher Attribution

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J. and Mugavero, M. J. (2018), Generalizing evidence from randomized trials using inverse probability of sampling weights. J. R. Stat. Soc. A, 181: 1193-1209. doi:10.1111/rssa.12357
Available at: https://doi.org/10.1111/rssa.12357

Generalizing Evidence from Randomized Trials using Inverse Probability of Sampling Weights

Ashley L. Buchanan†

Department of Pharmacy Practice, College of Pharmacy, The University of Rhode Island

Michael G. Hudgens

Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina

Stephen R. Cole

Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina

Katie R. Mollan

School of Medicine, University of North Carolina

Paul E. Sax

Division of Infectious Diseases and Department of Medicine, Brigham and Women's Hospital and Harvard Medical School

Eric S. Daar

Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center

Adaora A. Adimora

School of Medicine, University of North Carolina

Joseph J. Eron

School of Medicine, University of North Carolina

Michael J. Mugavero

School of Medicine, University of Alabama

†*Address for correspondence:* 7 Greenhouse Road, Kingston, RI 02881

E-mail: buchanan@uri.edu

Summary.

Results obtained in randomized trials may not easily generalize to target populations. Whereas in randomized trials the treatment assignment mechanism is known, the sampling mechanism by which individuals are selected to participate in the trial is typically not known and assuming random sampling from the target population is often dubious. We consider an inverse probability of sampling weighted (IPSW) estimator for generalizing trial results to a target population. The IPSW estimator is shown to be consistent and asymptotically normal. A consistent sandwich-type variance estimator is derived and simulation results are presented comparing the IPSW estimator to a previously proposed stratified estimator. The methods are then utilized to generalize results from two randomized trials of HIV treatment to all people living with HIV in the US.

Keywords: Causal inference; External validity/Generalizability; HIV/AIDS; Inverse probability weights; Randomized controlled trial; Target population

1. Introduction

Generalizability is a concern for many scientific studies, including those in public health and medicine (Cole and Stuart, 2010; Hernan and VanderWeele, 2011; Stuart et al., 2011, 2015; Tipton, 2013; Keiding and Louis, 2016). Using information in the study sample, it is often of interest to draw inference about a specified target population. Therefore, it is important to consider the degree to which an effect estimated from a study sample approximates the true effect in the target population. Unfortunately, study participants often do not constitute a random sample from the target population, bringing into question the generalizability of effect estimates based on such studies. For example, in clinical trials of treatment for HIV-infected individuals, there is often concern that trial participants are not representative of the larger population of HIV-positive individuals. Greenblatt (2011) highlighted the over-representation of African American and Hispanic women among HIV cases in the United States (US) and the limited clinical trial participation of members of these groups. The Women's Interagency HIV Study (WIHS) is a prospective, observational, multicenter study considered to be representative of women living with HIV and women at risk for HIV infection in the US (Bacon et al., 2005). However, a review of eligibility criteria of 20 AIDS Clinical Trial Group (ACTG) studies found that 28% to 68% of the HIV-positive women in WIHS cohort would have been excluded from these trials (Gandhi et al., 2005).

There exist several quantitative methods that provide a formal approach to generalize results from a randomized trial to a specified target population. Some of these methods utilize a model of the probability of trial participation conditional on covariates. Herein, we refer to this conditional probability as the sampling score. Generalizabilty methods employing sampling scores are akin to methods that use treatment propensity scores to adjust for (measured) confounding (Rubin, 1980) and include the use of inverse probability of sampling weights and stratification based on sampling scores. For

example, Cole and Stuart (2010) estimated sampling scores using logistic regression and then employed inverse probability of sampling weighted (IPSW) methods to estimate the treatment effect in the target population. Another approach to generalizing trial results entails an estimator based on stratifying individuals according to their estimated sampling scores (Tipton, 2013; O'Muircheartaigh and Hedges, 2013; Tipton et al., 2014). To date, there have been no formal studies or derivations of the large sample statistical properties (e.g., consistency and asymptotic normality) of these generaliziability estimators.

Following Cole and Stuart (2010) and Stuart et al. (2011), we consider an inverse weighting approach based on sampling scores to generalize trial effect estimates to a target population. The inverse weighted estimator is compared to the stratified estimator. In Section 2, assumptions and notation are discussed. The IPSW estimator and the stratified estimator are described in Section 3. Large sample properties of the IPSW estimator are derived, including a closed-form expression for the asymptotic variance and a consistent sandwich-type estimator of the variance. The finite sample performance of the IPSW and stratified estimators are compared in a simulation study presented in Section 4. In Section 5, the IPSW estimator is applied to generalize results from two ACTG trials to all people currently living with HIV in the US. Section 6 concludes with a discussion.

2. Assumptions and Notation

Suppose we are interested in drawing inference about the effect of a treatment (e.g., drug) on an outcome (e.g., disease) in some target population. Assume each individual in the target population has two potential outcomes Y^0 and Y^1, where Y^0 is the outcome that would have been seen if (possibly contrary to fact) the individual received control, and Y^1 is the outcome that would have been seen if (possibly contrary to fact) the individual received treatment. It is assumed throughout that the stable unit treatment value assumption (SUTVA) (Rubin, 1978) holds, i.e., there are no variations of treatment and there is no interference between individuals (the outcome of one individual is assumed to be unaffected by treatment of other individuals). Let $\mu_1 = E\left(Y^1\right)$ and $\mu_0 = E\left(Y^0\right)$ denote the mean potential outcomes in the target population. The parameter of interest is the population average treatment effect (PATE) $\Delta = \mu_1 - \mu_0$.

Consider a setting where two data sets are available. A random sample (e.g., cohort study) of m individuals is drawn from the target population. A second sample of n individuals participate in a randomized trial. Unlike the cohort study, the trial participants are not necessarily assumed to be a random sample from the target population but rather may be a biased sample. The following random variables are observed for the cohort and trial participants. Let Z be a $1 \times p$ vector of covariates and assume that information on Z is available for those in the trial and those in the cohort. Let $S = 1$ denote trial participation and $S = 0$ otherwise. For those individuals who participate in the trial, define X as the treatment indicator, where $X = 1$ if assigned to treatment and $X = 0$ otherwise.

Let $Y = Y^1X + Y^0(1-X)$ denote the observed outcome. Assume (S, Z) is observed for cohort participants and (S, Z, X, Y) is observed for trial participants.

Assume the trial participants are randomly assigned to receive treatment or not such that the treatment assignment mechanism is ignorable, i.e., $P(X = x|S = 1, Z, Y^0, Y^1) = P(X = x|S = 1)$. Assume an ignorable trial participation mechanism conditional on Z, i.e., $P(S = s|Z, Y^0, Y^1) = P(S = s|Z)$. In other words, participants in the trial are no different from nonparticipants regarding the treatment-outcome relationship conditional on Z. Trial participation and treatment positivity (Westreich and Cole, 2010) are also assumed, i.e., $P(S = 1|Z = z) > 0$ for all z such that $P(Z = z) > 0$ and $P(X = x|S = 1) > 0$ for $x = 0, 1$. Assume participants in the trial are adherent to their treatment assignment (i.e., there is no noncompliance).

3. Inference about the Population Average Treatment Effects

3.1. Estimators

A traditional approach to estimating treatment effects is a difference in outcome means between the two randomized arms of the trial. Let $i = 1, \ldots, n + m$ index the trial and cohort participants. The within-trial estimator is defined as

$$\hat{\Delta}_T = \frac{\sum_i S_i Y_i X_i}{\sum_i S_i X_i} - \frac{\sum_i S_i Y_i (1 - X_i)}{\sum_i S_i (1 - X_i)},$$

where here and in the sequel $\sum_i = \sum_{i=1}^{n+m}$. If trial participants are assumed to constitute a random sample from the target population, it is straightforward to show $\hat{\Delta}_T$ is a consistent and asymptotically normal estimator of Δ. On the other hand, if we are not willing to assume trial participants are a random sample from the target population, then $\hat{\Delta}_T$ is no longer guaranteed to be consistent.

Below we consider two estimators of Δ that do not assume trial participants are a random sample from the target population. Both estimators utilize sampling scores. Following Cole and Stuart (2010), assume a logistic regression model for the sampling scores such that $P(S = 1|Z = z) = \{1 + \exp(-z\beta)\}^{-1}$ where β is a $p \times 1$ vector of coefficient parameters. Note here and throughout we assume the $p \times 1$ vector Z includes 1 as the first component in order to accommodate an intercept term in the sampling score model. Let $\hat{\beta}$ denote the weighted maximum likelihood estimator of β where each trial participant has weight $\Pi_{S_i}^{-1} = 1$ and each individual in the cohort has weight $\Pi_{S_i}^{-1} = m/(N - n)$, where N is the size of the target population (Scott and Wild, 1986). Let $P(S = 1|Z = z) = w(z, \beta)$, $w_i = w(Z_i, \beta)$, and $\hat{w}_i = w(Z_i, \hat{\beta})$. The IPSW estimator (Cole and Stuart, 2010) of the PATE is

$$\hat{\Delta}_{IPW} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{\sum_i S_i Y_i X_i / \hat{w}_i}{\sum_i S_i X_i / \hat{w}_i} - \frac{\sum_i S_i Y_i (1 - X_i)/\hat{w}_i}{\sum_i S_i (1 - X_i)/\hat{w}_i}. \quad (1)$$

Another approach for estimating the PATE uses stratification based on the sampling scores (Tipton, 2013; O'Muircheartaigh and Hedges, 2013; Tipton et al., 2014) and is computed in the following steps. First, β is estimated using a logistic regression model as described above and the estimated

sampling scores $\hat{w}_i$ are computed. These estimated sampling scores are used to form L strata. The difference of sample means within each stratum is computed among those in the trial. The PATE is then estimated as a weighted sum of the differences of sample means across strata. The stratum specific weights used in computing this weighted average equal estimates of the proportion of individuals in the target population within the stratum. Specifically, let n_l be the number of individuals in the trial in stratum l and m_l be the number of individuals in the cohort in stratum l. Let $S_{il} = 1$ denote trial participation for individual i in stratum l for $i = 1, \ldots, (n_l + m_l)$ and $l = 1, \ldots, L$ (and $S_{il} = 0$ otherwise). If $S_{il} = 1$, then let X_{il} and Y_{il} denote the treatment assignment and outcome for individual i in stratum l; otherwise, if $S_{il} = 0$, then let $X_{il} = Y_{il} = 0$. The sampling score stratified estimator is defined as

$$\hat{\Delta}_S = \sum_{l=1}^{L} \omega_l \left(\frac{\sum_{i=1}^{n_l+m_l} S_{il} X_{il} Y_{il}}{\sum_{i=1}^{n_l+m_l} S_{il} X_{il}} - \frac{\sum_{i=1}^{n_l+m_l} S_{il}(1 - X_{il}) Y_{il}}{\sum_{i=1}^{n_l+m_l} S_{il}(1 - X_{il})} \right),$$

where $\omega_l = N_l/N$, $N_l = \sum_{i=1}^{n_l+m_l} \Pi_{S_{il}}^{-1}$, and $\Pi_{S_{il}}$ is the weight for individual i in stratum l.

3.2. Large Sample Properties of the IPSW Estimator

Because the trial participants are not assumed to be a random sample from the target population, the observed random variables $(S_i, Z_i, S_i X_i, S_i Y_i)$ for $i = 1, \ldots, n+m$ are assumed to be independent but not necessarily identically distributed. Below, the IPSW estimator is expressed as the solution to an unbiased estimating equation to establish asymptotic normality and provide a consistent sandwich-type estimator of the variance.

First, consider the case when β is known. Let $\hat{\theta}^* = (\hat{\mu}_1, \hat{\mu}_0)$, $\theta^* = (\mu_1, \mu_0)$ and note that $\hat{\theta}^*$ is the solution for θ^* of the estimating equation

$$\sum_i \Psi_\Delta^*(Y_i, Z_i, X_i, S_i, \theta^*) = \begin{pmatrix} \sum_i \{S_i X_i (Y_i - \mu_1)\}/w_i \\ \sum_i \{S_i (1 - X_i)(Y_i - \mu_0)\}/w_i \end{pmatrix} = 0.$$

Define the following matrices:

$$A_{m,n}(\theta^*) = (n+m)^{-1} \sum_i E\{ \frac{\partial}{\partial \theta^*} \Psi_\Delta^*(Y_i, Z_i, X_i, S_i, \theta^*) \}$$

$$B_{m,n}(\theta^*) = (n+m)^{-1} \sum_i \mathrm{cov}\{ \Psi_\Delta^*(Y_i, Z_i, X_i, S_i, \theta^*) \}$$

Define $A(\theta^*) = \lim_{m,n\to\infty} A_{m,n}(\theta^*)$ and $B(\theta^*) = \lim_{m,n\to\infty} B_{m,n}(\theta^*)$. Note $E\{\Psi_\Delta^*(Y_i, Z_i, X_i, S_i, \theta^*)\} = 0$ for $i = 1, ..., n+m$, implying under suitable regularity conditions that as $n, m \to \infty$, $\hat{\theta}^*$ converges in probability to θ^* and $(n+m)^{1/2}(\hat{\theta}^* - \theta^*)$ converges in distribution to $N(0, \Sigma_\theta^*)$ where

$$\Sigma_\theta^* = A(\theta^*)^{-1} B(\theta^*) A(\theta^*)^{-T} \tag{2}$$

(Carroll et al. 2010, Appendix A.6). By Slutsky's theorem and the delta method, $\hat{\Delta}_{IPW}$ is a consistent estimator of Δ and $(n+m)^{1/2}(\hat{\Delta}_{IPW} - \Delta)$ converges in distribution to $N(0, \Sigma_{IPW}^*)$ where

$$\Sigma_{IPW}^* = \Sigma_\theta^{*(11)} + \Sigma_\theta^{*(22)} - 2\Sigma_\theta^{*(12)} \tag{3}$$

and in general $\Sigma^{(ij)}$ refers to the entry in the i^{th} row and the j^{th} column of the matrix Σ. A consistent estimator of (3) is given in Appendix A.

Next consider the more likely case that β is unknown. Using weighted maximum likelihood, the estimator $\hat{\beta}$ is the solution for β of the $p \times 1$ vector estimating equation

$$\sum_i \psi_\beta(S_i, Z_i, \beta) = \sum_i \Pi_{S_i}^{-1} \frac{S_i - w_i}{w_i(1-w_i)} \frac{\partial}{\partial \beta} w_i = 0$$

(Scott and Wild, 1986). Let $\hat{\theta} = (\hat{\mu}_1, \hat{\mu}_0, \hat{\beta})$, $\theta = (\mu_1, \mu_0, \beta)$ and note that $\hat{\theta}$ is the solution for θ of the $(p+2) \times 1$ vector estimating equation

$$\sum_i \Psi_\Delta(Y_i, Z_i, X_i, S_i, \theta) = \begin{pmatrix} \sum_i \{S_i X_i (Y_i - \mu_1)\}/w_i \\ \sum_i \{S_i (1 - X_i)(Y_i - \mu_0)\}/w_i \\ \sum_i \psi_\beta(S_i, Z_i, \beta) \end{pmatrix} = 0.$$

Define the following matrices:

$$A_{m,n}(\theta) = (n+m)^{-1} \sum_i E\{\frac{\partial}{\partial \theta} \Psi_\Delta(Y_i, Z_i, X_i, S_i, \theta)\}$$

$$B_{m,n}(\theta) = (n+m)^{-1} \sum_i \text{cov}\{\Psi_\Delta(Y_i, Z_i, X_i, S_i, \theta)\}.$$

Define $A(\theta) = \lim_{m,n \to \infty} A_{m,n}(\theta)$ and $B(\theta) = \lim_{m,n \to \infty} B_{m,n}(\theta)$. Note $E\{\Psi_\Delta(Y_i, Z_i, X_i, S_i, \theta)\} = 0$ for $i = 1, ..., n+m$, implying under suitable regularity conditions that as $n, m \to \infty$, $\hat{\theta}$ converges in probability to θ and $(n+m)^{1/2}(\hat{\theta} - \theta)$ converges in distribution to $N(0, \Sigma_\theta)$ where

$$\Sigma_\theta = A(\theta)^{-1} B(\theta) A(\theta)^{-T} \tag{4}$$

(Carroll et al., 2010). By Slutsky's theorem and the delta method, $\hat{\Delta}_{IPW}$ is a consistent estimator of Δ and $(n+m)^{1/2}(\hat{\Delta}_{IPW} - \Delta)$ converges in distribution to $N(0, \Sigma_{IPW})$ where

$$\Sigma_{IPW} = \Sigma_\theta^{(11)} + \Sigma_\theta^{(22)} - 2\Sigma_\theta^{(12)}. \tag{5}$$

A consistent estimator of (5) is given in Appendix A. This variance estimator can be used to construct Wald-type confidence intervals (CIs) for Δ.

Comparison of (3) and (5) shows that the variance is smaller when the sampling scores are estimated (see Appendix B). Therefore, even if the correct sampling scores are known, estimation of the sampling scores is preferable due to improved efficiency. This is analogous to a well-known result for inverse probability of treatment weighted estimators (Hirano et al., 2003; Robins et al., 1992; Wooldridge, 2007). In general, it is common practice to compute the variance of the inverse probability weighted estimators using standard software assuming the weights are known. This leads to valid but conservative CIs. In the Supplementary Material, an R function is provided which computes the IPSW estimator and the corresponding (consistent) sandwich-type estimator of the variance described in Appendix A which does not assume β is known.

3.3. Estimator of the Variance of the Stratified Estimator

One approach to obtain an estimator of the variance of the stratified estimator is to express $\hat{\Delta}_S$ as the solution to an unbiased vector of estimating equations, which include an estimating equation for the potential outcome means, the L quantiles, and each element of β. This approach can be used to show $\hat{\Delta}_S$ is asymptotically normal (Lunceford and Davidian, 2004). In practice, it is routine to approximate the sampling variance of $\hat{\Delta}_S$ by treating the estimator as the average of L independent, within-stratum, treatment effect estimators (Tipton, 2013; Lunceford and Davidian, 2004). Specifically, the approximate variance of $\hat{\Delta}_S$ is

$$\sum_{l=1}^{L} \omega_l^2 \hat{\sigma}_l^2, \tag{6}$$

where $\hat{\sigma}_l^2 = \sum_{x=0}^{1} n_{xl}^{-1} s_{xl}^2$, $n_{xl} = \sum_{i=1}^{n_l+m_l} S_{il} I(X_{il} = x)$, $s_{xl} = n_{xl}^{-1} \sum_{i=1}^{n_l+m_l} S_{il} I(X_{il} = x)(Y_{il} - \bar{Y}_{xl})^2$ and $\bar{Y}_{xl} = n_{xl}^{-1} \sum_{i=1}^{n_l+m_l} S_{il} I(X_{il} = x) Y_{il}$ for $x = 0, 1$.

4. Simulations

A simulation study was conducted to compare the performance of the IPSW and stratified estimators in scenarios with a continuous or discrete covariate and a continuous outcome. The following quantities were computed for each scenario: the bias for each estimator, the average of the estimated standard errors, empirical standard error, and empirical coverage probability of the 95% CIs.

A total of 5,000 data sets were simulated per scenario as follows. There were $N = 10^6$ observations in the target population with sample score $w_i = \{1 + \exp(-\beta_0 - \beta_1 Z_{1i})\}^{-1}$. In the first two scenarios, one binary covariate $Z_{1i} \sim$ Bernoulli(0.2) was considered and, for scenarios 3 to 6, one continuous covariate $Z_{1i} \sim N(0, 1)$ was considered. The covariate Z_{1i} was associated with trial participation and a treatment effect modifier. A Bernoulli trial participation indicator, S_i, was simulated according to the true sampling score w_i in the target population and those with $S_i = 1$ were included in the trial. The parameters β_0 and β_1 were set were set such that the sample size in the trial was approximately $n \approx 1000$. The cohort was a random sample of size $m =$ 4,000 from the target population (less those selected into the trial) and S_i was set to zero for those in the cohort. The trial was small compared to the size of the target, so the cohort was essentially a random sample from the target.

For those included in the randomized trial ($S_i = 1$), X_i was generated as Bernoulli(0.5) and the outcome Y was generated according to $Y_i = \nu_0 + \nu_1 Z_{1i} + \xi X_i + \alpha Z_{1i} X_i + \epsilon_i, \epsilon_i \sim N(0, 1)$. For scenarios 1 to 4, $(\nu_0, \nu_1, \xi, \alpha) = (0, 1, 2, 1)$. For scenarios 5 to 6, $(\nu_0, \nu_1, \xi, \alpha) = (0, 1, 2, 2)$. Two sampling score models were considered: Scenario 1, 3, and 5 set $\beta = (-7, 0.4)$; Scenario 2, 4, and 6 set $\beta = (-7, 0.6)$. The truth was calculated for each scenario based on the distribution of Z_{1i} in the target population. The truth was $\Delta = 2.2$ for scenarios 1 and 2 and $\Delta = 2$ for scenarios 3 through 6. To estimate the sampling scores, the combined trial ($S_i = 1$) and cohort ($S_i = 0$) data was used to fit a (weighted) logistic regression model with S_i as the outcome and the covariate Z_{1i} as described in Section 3.1.

Comparisons between the IPSW and stratified estimator when the sampling score model was correctly specified are summarized in Table 1. The within-trial estimator $\hat{\Delta}_T$ was biased for all scenarios and had low coverage (results not shown). For all scenarios, $\hat{\Delta}_{IPW}$ was unbiased. For scenarios 1 to 2, $\hat{\Delta}_S$ was unbiased and standard errors were comparable to $\hat{\Delta}_{IPW}$. For scenarios 3 to 6, $\hat{\Delta}_S$ was biased, possibly due to residual confounding from a continuous covariate in the sampling score model. For the IPSW estimator, the average of the estimated standard error was approximately equal to the empirical standard error, supporting the derivations of the sandwich-type estimator of the variance. For all scenarios, coverage was approximately 95% for the Wald CI of $\hat{\Delta}_{IPW}$. With a continuous covariate, the Wald CI of the stratified estimator had poor coverage, particularly in the presence of stronger effect modification (e.g., scenarios 5 and 6). Histograms of the three estimators for scenario 4 are given in Figure 1; the IPSW was approximately unbiased and normally distributed.

Simulations were also performed with the sampling score model misspecified. A second covariate was generated for each member of the target population and the true sampling score was $w_i = \{1+\exp(-\beta_0-\beta_1 Z_{1i}-\beta_2 Z_{2i})\}^{-1}$. For the first two scenarios, $Z_{2i} \sim$ Bernoulli(0.6), and for scenarios 3 to 6, $Z_{2i} \sim N(0,1)$. For those included in the randomized trial ($S_i = 1$), X_i was generated as Bernoulli(0.5) and the outcome Y was generated according to $Y_i = \nu_0 + \nu_1 Z_{1i} + \nu_2 Z_{2i} + \xi X_i + \alpha_1 Z_{1i} X_i + \alpha_2 Z_{2i} X_i + \epsilon_i, \epsilon_i \sim N(0,1)$. For scenarios 1 to 4, $(\nu_0, \nu_1, \nu_2, \xi, \alpha_1, \alpha_2) = (0,1,1,2,1,1)$. For scenarios 5 to 6, $(\nu_0, \nu_1, \nu_2, \xi, \alpha_1, \alpha_2) = (0,1,1,2,2,2)$. The estimated sampling scores were computed using logistic regression with Z_{1i} as the only covariate. Two sampling score models were considered: Scenario 1, 3, and 5 set $\beta = (-7, 0.4, 0.4)$; Scenario 2, 4, and 6 set $\beta = (-7, 0.6, 0.6)$. Based on the distribution of $Z_i = (Z_{1i}, Z_{2i})$ in the target population, the truth was $\Delta = 2.8$ for scenarios 1 and 2 and $\Delta = 2$ for scenarios 3 through 6.

Comparisons between the IPSW and stratified estimator are summarized in Table 2 when the sampling score model was misspecified. The bias was reduced by approximately half when either the IPSW or the stratified estimator was employed as compared to the within-trial estimator. The sandwich-type estimator of the variance of the IPSW estimator performed reasonably well when the sampling score model was misspecified; however, CI coverage was below the nominal level.

5. Applications

5.1. Trials and Cohorts

In this section, the methods described in Section 3 are applied to generalize results from two different ACTG randomized clinical trials, ACTG 320 and ACTG A5202. Two different target populations are considered, namely all women currently living with HIV in the US and all people currently living with HIV in the US.

The ACTG 320 trial examined the safety and efficacy of adding a protease inhibitor (PI) to an HIV treatment regimen with two nucleoside analogues. A total of 1,156 participants were enrolled in

ACTG 320 between January 1996 and January 1997 and were recruited from 33 AIDS clinical trial units and 7 National Hemophilia Foundation sites in the US and Puerto Rico. These participants were HIV-positive, highly active antiretroviral therapy (HAART) naive, and had CD4 T cell counts ≤ 200 cells/mm^3 at screening. Of the 1,156 participants, 200 were women (Hammer et al., 1997). Among ACTG 320 participants, 116 (10%) were missing the primary outcome of CD4 count at week 4, so they are excluded from the analysis below. The baseline characteristics of the ACTG 320 participants are shown in Supplemental Table 1.

The ACTG A5202 trial assessed equivalence of abacavir-lamivudine (ABC-3TC) or tenofovir disoproxil fumarate-emtricitabine (TDF-FTC) plus efavirenz or ritonavir-boosted atazanavir. A total of 1,857 participants were enrolled in ACTG A5202 between September 2005 and November 2007 and were recruited from 59 ACTG sites in the US and Puerto Rico. These participants were HIV-positive, antiretroviral (ART) naive, and had viral load $>$ 1,000 copies/ml at screening. Of the 1,857 participants, 322 were women (Sax et al., 2009, 2011). Among ACTG A5202 participants, 417 (22%) were missing the primary outcome of CD4 count at week 48, so they are excluded from the analysis below. The baseline characteristics of the ACTG A5202 participants are shown in Supplemental Table 2.

Data from two cohort studies, WIHS and Center for AIDS Research Network of Integrated Clinical Systems (CNICS), are used in the analysis below to generalize the ACTG 320 and A5202 trial results. Participants in WIHS and CNICS were considered to be representative samples of the target populations, i.e., all women living with HIV in the US and all people living with HIV in the US, respectively. A total of 4,129 women (1,065 HIV-uninfected) were enrolled in WIHS between October 1994 and December 2012 at six US sites (Bacon et al., 2005). The CNICS captures comprehensive and standardized clinical data from point-of-care electronic medical record systems for population-based HIV research (Kitahata et al., 2008). The CNICS cohort includes over 27,000 HIV-infected adults (at least 18 years of age) engaged in clinical care since January 1995 at eight CFAR sites in the US.

For generalizing results from ACTG 320, the analysis included cohort participants who were HIV-positive, HAART naive, and had CD4 cell counts ≤ 200 cells/mm^3 at the previous visit ($m = 493$ women and $m = 6{,}158$ men and women combined). For generalizing results from A5202, the analysis included cohort participants who were HIV-positive, ART naive, and had viral load $>$ 1,000 copies/ml at the previous visit ($m = 1{,}012$ women and $m = 12{,}302$ men and women combined). Supplemental Table 1 displays the characteristics of the women in the WIHS sample and the participants in the CNICS sample used to generalize results from ACTG 320. Likewise, the characteristics of the women in the WIHS sample and participants in the CNICS sample used to generalize results from ACTG A5202 are displayed in Supplemental Table 2.

5.2. Analysis

The IPSW and stratified estimators were employed to generalize the difference in the average change in CD4 from baseline between treatment groups observed among women in the trials to all women currently living with HIV in the US and among all participants in the trials to all people currently living with HIV in the US. Based on Centers for Disease Control and Prevention (2012) estimates, the size of the first target population was assumed to be 280,000 women and the size of the second target population was assumed to be 1.1 million people.

The population average treatment effect was estimated using the IPSW estimator in equation (1). To estimate the sampling scores, the data from the ACTG trial (i.e., 320 and A5202) and cohort (i.e., WIHS or CNICS) were analyzed together, with $S = 1$ for those in the ACTG trial and $S = 0$ for those in the cohort. In the model to estimate the sampling scores, the outcome was trial participation and the possible covariates for ACTG 320 included sex, race/ethnicity, age, history of injection drug use (IDU), and baseline CD4 and for ACTG A5202 included sex, race/ethnicity, age, history of IDU, hepatitis B/C, AIDS diagnosis, baseline CD4 and baseline $\log_{10}$ viral load. Variables associated with trial participation, the outcome, or effect modifiers, as well as all pairwise interactions, were included in the sampling score model. Sex was not included as a covariate in analyses generalizing the trial results among women.

5.3. Results

Estimates of the mean differences based on the within-trial estimator among women and all participants are given in Table 3. Among all participants and among just women in ACTG 320, there was a significant difference in the change in CD4 from baseline to 4 weeks between the PI and non-PI groups. Among women in A5202 at week 48, those randomized to ABC-3TC had an average change in CD4 cell count comparable to those randomized to a regimen with TDF-FTC. Among all participants in A5202, those randomized to ABC-3TC had an average change in CD4 cell count slightly higher than those randomized to a regimen with TDF-FTC, but this did not achieve statistical significance.

Table 3 also displays the results for the two ACTG trials generalized to both target populations. In the target population of all women living with HIV in the US, the IPSW estimate was approximately double the within-trial estimate ($\hat{\Delta}_{IPW} = 46$ compared to $\hat{\Delta}_T = 24$), suggesting that the within-trial result may underestimate the effects of PIs in all HIV-infected women in the US. The IPSW estimator also indicated a much stronger protective effect of ABC-3TC (vs. TDF-FTC) in the target population of all HIV-infected women in the US ($\hat{\Delta}_{IPW} = 35$ compared to $\hat{\Delta}_T = 1$), providing evidence that this particular ART combination may increase CD4 cell counts more on average than what was observed in the trial. In the target population of all people living with HIV in the US, the IPSW estimates were comparable to the within-trial effect estimates, suggesting that both the effect of PIs and the effect of the ART combination ABC-3TC (vs. TDF-FTC) from the trials may be generalizable to all

people living with HIV in the US. In summary, these results suggest the ACTG trial results are more generalizable for US men with HIV than US women with HIV.

6. Discussion

In this paper, we considered generalizing results from a randomized trial to a specific target population using inverse probability of sampling weights. The IPSW estimator was shown to be consistent and asymptotically normal and a consistent sandwich-type estimator of the variance was provided. In a simulation study, the IPSW outperformed the stratified estimator when the sampling score was correctly specified. The IPSW was unbiased for all scenarios and the CIs exhibited coverage approximately at the nominal level. With a continuous covariate, the stratified estimator exhibited bias and the corresponding CI had poor coverage, particularly in the presence of stronger effect modification.

In the illustrative example, the ACTG 320 results appear to be generalizable to all people living with HIV in the US. On the other hand, the within-trial effect estimates of ACTG 320 and ACTG A5202 among women were not comparable to the effect estimated in the target population of women. For the A5202 results among women, the difference in the effect estimates was primarily due to hepatitis, which was negatively associated with trial participation. Results from both ACTG A5202 and ACTG 320 were not sensitive to the specification of the size of the target population, although some results were sensitive to the specification of the sampling score model (results not shown). In the data example, a complete case analysis was performed; however, in practice, one would want to address the possibility that the missinginess was not completely at random.

When applying these methods, the analysis is subject to the following considerations. The ignorable trial participation mechanism assumption, i.e., that participants' decisions to participate in a trial is independent of their outcomes conditional on covariates, is untestable. In trials with non-neglible rates of non-compliance, effect estimates based on IPSW or stratification should be interpreted as estimates of treatment assignment rather than treatment receipt. Future research could entail extending these estimators to account for non-compliance. The sampling score model was assumed to be correctly specified (e.g., correct covariate functional forms). Because some degree of model mis-specification is inevitable, sensitivity analysis of inferences about the treatment effect in the population to the sampling score model specification is recommended. The stratified estimator (Tipton et al., 2014; O'Muircheartaigh and Hedges, 2013) requires that individuals sharing the same stratum of the sampling score distribution can be identified. This estimator may be biased when there is residual confounding within strata and, in general, is not a consistent estimator of the PATE (Lunceford and Davidian, 2004).

In the application, the cohort study was assumed to be a random sample (i.e., representative) of the target population. If the cohort is not representative, one possibility is weighting the cohort data

to the distribution of covariates in a census (e.g., Centers for Disease Control and Prevention (CDC) estimates). A limitation of this approach is that the census may not have covariate information as rich as the cohort data. The CDC estimates used to quantify the size of the target population in the example were for all people living with HIV. Use of surveillance studies that report on the number of ART and HAART naive HIV patients in the US could further sharpen the information about the target population.

Weighted logistic regression was used to estimate the sampling scores. Future research could entail instead using machine learning methods (e.g., as in Westreich et al. (2010)) to estimate the sampling scores. Additional research to develop an augmented estimator could improve efficiency (Zhang et al., 2008). This method could be extended to accommodate interference or right-censored outcomes.

Acknowledgments

These findings are presented on behalf of the Women's Interagency HIV Study (WIHS), the Center for AIDS Research (CFAR) Network of Integrated Clinical Trials (CNICS), and the AIDS Clinical Trials Group (ACTG). We would like to thank all of the WIHS, CNICS, and ACTG investigators, data management teams, and participants who contributed to this project. Funding for this study was provided by National Institutes of Health (NIH) grants R01AI100654, R01AI085073, U01AI042590, U01AI069918, R56AI102622, 5 K24HD059358-04, 5 U01AI103390-02 (WIHS), R24AI067039 (CNICS), and P30AI50410 (UNC CFAR). The views and opinions of authors expressed in this manuscript do not necessarily state or reflect those of the NIH.

Data in this manuscript were collected by the WIHS. WIHS (Principal Investigators): UAB-MS WIHS (Michael Saag, Mirjam-Colette Kempf, and Deborah Konkle-Parker), U01-AI-103401; Atlanta WIHS (Ighovwerha Ofotokun and Gina Wingood), U01-AI-103408; Bronx WIHS (Kathryn Anastos), U01-AI-035004; Brooklyn WIHS (Howard Minkoff and Deborah Gustafson), U01-AI-031834; Chicago WIHS (Mardge Cohen and Audrey French), U01-AI-034993; Metropolitan Washington WIHS (Mary Young), U01-AI-034994; Miami WIHS (Margaret Fischl and Lisa Metsch), U01-AI-103397; UNC WIHS (Adaora Adimora), U01-AI-103390; Connie Wofsy Womens HIV Study, Northern California (Ruth Greenblatt, Bradley Aouizerat, and Phyllis Tien), U01-AI-034989; WIHS Data Management and Analysis Center (Stephen Gange and Elizabeth Golub), U01-AI-042590; Southern California WIHS (Joel Milam), U01-HD-032632 (WIHS I - WIHS IV). The WIHS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), the National Cancer Institute (NCI), the National Institute on Drug Abuse (NIDA), and the National Institute on Mental Health (NIMH). Targeted supplemental funding for specific projects is also provided by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute on Alcohol Abuse and Alcoholism (NIAAA), the National Institute on Deafness and other Communication Disorders (NIDCD), and the NIH Office of Research on Womens Health. WIHS data collection is also supported by UL1-TR000004 (UCSF CTSA) and UL1-TR000454 (Atlanta CTSA).

The data and computer code used to generate the results in the manuscript can be provided upon request and are subject to approval from the WIHS, CNICS, and ACTG study principal investigators and executive committees. Approvals for data sharing requests are subject to any rules and regulations specific to the studies analyzed in this manuscript or otherwise at the NIH. A request for data and computer code can be initiated by contacting the corresponding author.

Table 1: Summary of Monte Carlo results for estimators of the population average treatment effect when the sampling score model was correctly specified with a continuous outcome for $5,000$ simulated data sets with m = 4,000 and $n \approx$ 1,000 per data set. Scenarios are described in Section 4. For scenarios 1 and 2 $\Delta = 2.2$ and for scenarios 3 to 6 $\Delta = 2.0$ (T = within-trial; S = stratified; IPW = inverse probability of sampling weighted; ESE = empirical standard error ($\times$100); ASE = average estimated standard error ($\times$100); ECP = empirical coverage probability; Cov = covariate; Bin = binary; Cont= continuous)

			Bias			**ESE**		**ASE**		**ECP**	
Scenario	Cov.	(β_1,α)	$\hat{\Delta}_T$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPW}$
1	Bin.	(0.4,1)	0.07	1e-3	2e-3	6.2	7.1	7.1	7.3	0.98	0.95
2	Bin.	(0.6,1)	0.11	-3e-5	-6e-4	6.3	7.1	6.6	7.1	0.96	0.95
3	Cont.	(0.4,1)	0.20	0.04	1e-3	8.1	13.4	7.9	13.4	0.91	0.95
4	Cont.	(0.6,1)	0.60	0.07	-1e-4	8.6	15.0	8.6	14.9	0.88	0.95
5	Cont.	(0.4,2)	0.80	0.09	3e-3	9.4	17.2	8.9	17.2	0.81	0.95
6	Cont.	(0.6,2)	1.20	0.14	-1e-3	10.1	19.9	9.8	19.6	0.70	0.95

Table 2: Summary of Monte Carlo results for estimators of the population average treatment effect when the sampling score model was misspecified with a continuous outcome for $5,000$ simulated data sets with $m = 4{,}000$ and $n \approx 1{,}000$ per data set. Scenarios are described in Section 4. For scenarios 1 and 2 $\Delta = 2.8$ and for scenarios 3 to 6 $\Delta = 2.0$ (T = within-trial; S = stratified; IPW = inverse probability of sampling weighted; ESE = Empirical standard error ($\times 100$); ASE = Average estimated standard error ($\times 100$); ECP = Empirical coverage probability; Cov = covariate; Bin = binary; Cont= continuous)

			Bias			**ESE**		**ASE**		**ECP**	
Scenario	Cov.	(β_1,α)	$\hat{\Delta}_T$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPW}$
1	Bin.	(0.4,1)	0.16	0.09	0.09	7.03	7.67	7.73	7.61	0.80	0.77
2	Bin.	(0.6,1)	0.24	0.13	0.13	6.36	6.82	6.62	6.86	0.49	0.52
3	Cont.	(0.4,1)	0.80	0.45	0.40	13.12	16.53	12.88	16.57	0.07	0.32
4	Cont.	(0.6,1)	1.20	0.67	0.60	13.19	17.58	12.90	17.24	<0.01	0.08
5	Cont.	(0.4,2)	1.60	0.89	0.80	17.37	22.12	16.98	22.20	<0.01	0.05
6	Cont.	(0.6,2)	2.39	1.34	1.20	17.49	23.79	17.04	23.32	<0.01	<0.01

Table 3: Estimated difference in means and corresponding 95% confidence intervals (CIs) in two target populations (all men and women combined and all women living with HIV in the US) based on data from AIDS Clinical Trials Group (ACTG) trials. T = within trial; S = stratified; IPW = inverse probability of sampling weighted.

		Difference in Means (95% CI)		
Cohort	**Trial**	$\hat{\Delta}_T$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPW}$
WIHS	320[a]	24 (7, 41)	38 (17, 59)	46 (23, 70)
WIHS	A5202[b]	1 (-35, 37)	-19 (-62, 25)	35 (-45, 115)
CNICS	320	19 (12, 25)	18 (9, 26)	17 (9, 25)
CNICS	A5202	6 (-8, 20)	7 (-18, 32)	-2 (-31, 28)

[a] For 320, the treatment contrast was PI ($X = 1$) vs. no PI ($X = 0$).

[b] For A5202, the treatment contrast was ABC-3TC ($X = 1$) vs. TDF-FTC ($X = 0$).

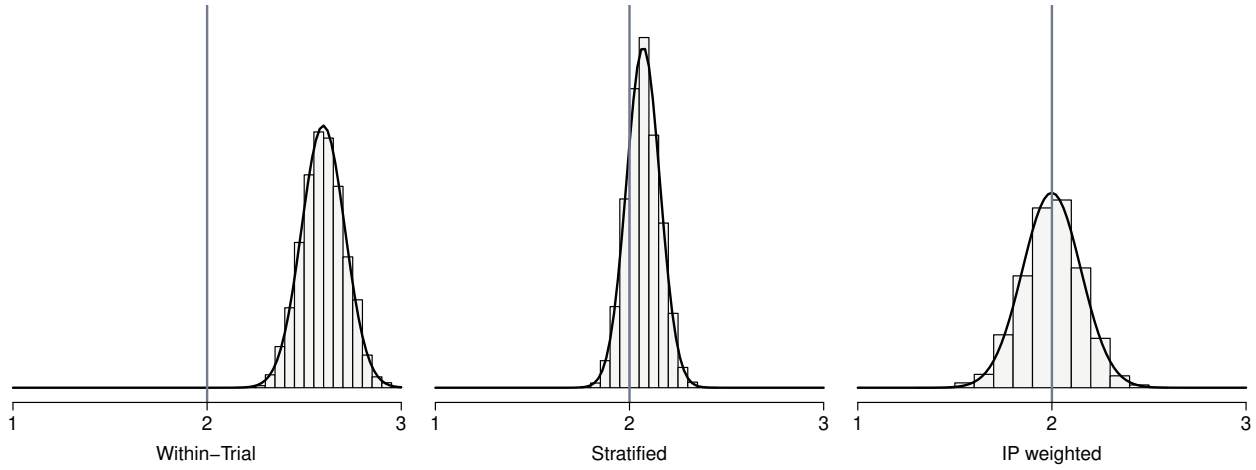

Fig. 1: Comparison of the distributions of within-trial estimator $\hat{\Delta}_T$, stratified estimator $\hat{\Delta}_S$, and inverse probability of sampling weighted estimator $\hat{\Delta}_{IPW}$, based on 5,000 simulated data sets where the sampling score model is correctly specified and $\Delta_0 = 2$ with one continuous covariate, $\beta = (-7, 0.6)$ and $\alpha = 1$ (Scenario 4).

Appendix A: Sandwich-Type Estimators of the Variance for the IPSW Estimator

The empirical sandwich-type estimator can be used to estimate the asymptotic variance of the IPSW estimator. Define the following matrices:

$$\hat{A}^* = (n+m)^{-1} \sum_i \frac{\partial}{\partial \theta^*} \Psi^*_\Delta(Y_i, Z_i, X_i, S_i, \theta^*) \Big|_{\theta^* = \hat{\theta}^*}$$

$$\hat{B}^* = (n+m)^{-1} \sum_i \Psi^*_\Delta(Y_i, Z_i, X_i, S_i, \hat{\theta}^*) \Psi^{*^T}_\Delta(Y_i, Z_i, X_i, S_i, \hat{\theta}^*)$$

Substituting these empirical estimators for their corresponding quantities in (2) yields a consistent estimator of the asymptotic variance of $\hat{\theta}^*$ when β is known. That is, $\hat{\Sigma}^*_\theta = \hat{A}^{*-1}\hat{B}^*\hat{A}^{*-T}$ is a consistent estimator of Σ^*_θ and thus a consistent estimator of the asymptotic variance of $\hat{\Delta}_{IPW}$ is

$$\hat{\Sigma}^*_{IPW} = \hat{\Sigma}^{*(11)}_\theta + \hat{\Sigma}^{*(22)}_\theta - 2\hat{\Sigma}^{*(12)}_\theta$$

The estimated standard error is $\hat{\text{se}}(\hat{\Delta}_{IPW}) = \sqrt{(n+m)^{-1}\hat{\Sigma}^*_{IPW}}$.

Similarly, when the weights are estimated (i.e., β is not known), define the following matrices:

$$\hat{A} = (n+m)^{-1} \sum_i \frac{\partial}{\partial \theta} \Psi_\Delta(Y_i, Z_i, X_i, S_i, \theta) \Big|_{\theta = \hat{\theta}}$$

$$\hat{B} = (n+m)^{-1} \sum_i \Psi_\Delta(Y_i, Z_i, X_i, S_i, \hat{\theta}) \Psi^T_\Delta(Y_i, Z_i, X_i, S_i, \hat{\theta})$$

Substituting these empirical estimators for their corresponding quantities in (4) yields a consistent estimator of the asymptotic variance of $\hat{\theta}$. That is, $\hat{\Sigma}_\theta = \hat{A}^{-1}\hat{B}\hat{A}^{-T}$ is a consistent estimator of Σ_θ and thus a consistent estimator of the asymptotic variance of $\hat{\Delta}_{IPW}$ is

$$\hat{\Sigma}_{IPW} = \hat{\Sigma}^{(11)}_\theta + \hat{\Sigma}^{(22)}_\theta - 2\hat{\Sigma}^{(12)}_\theta$$

The estimated standard error is $\hat{\text{se}}(\hat{\Delta}_{IPW}) = \sqrt{(n+m)^{-1}\hat{\Sigma}_{IPW}}$.

Appendix B: Proof of Efficiency Gain When Sampling Scores are Estimated

First consider the case when β is known. The asymptotic variance of $\hat{\Delta}_{IPW}$ can be expressed as $\Sigma^*_{IPW} = \tau\Sigma^*_\theta\tau^T$ where $\tau = (1, -1)$ and $\Sigma^*_\theta = A(\theta^*)^{-1}B(\theta^*)A(\theta^*)^{-T}$. Next consider the case when β is estimated. Let $d_i(\beta) = \partial w(z_i, \beta)/\partial\beta$ and define

$$E_{\beta\beta} = \lim_{n,m\to\infty} (n+m)^{-1} \sum_i \Pi^{-1}_{S_i}\{d_i(\beta)d_i^T(\beta)\}/w_i(1-w_i)^T$$

$$G_1 = \lim_{n,m\to\infty} (n+m)^{-1} \sum_i E[\{S_iX_i(Y_i - \mu_1)d_i(\beta)\}/w_i^2]$$

$$G_2 = \lim_{n,m\to\infty} (n+m)^{-1} \sum_i E[\{S_i(1-X_i)(Y_i - \mu_0)d_i(\beta)\}/w_i^2],$$

and let $G = (G_1, G_2)$. Then, using block matrix notation note

$$A(\theta) = \begin{pmatrix} A(\theta^*) & -G^T \\ 0_{p\times 2} & -E_{\beta\beta} \end{pmatrix} \text{ and } B(\theta) = \begin{pmatrix} B(\theta^*) & G^T \\ G & E_{\beta\beta} \end{pmatrix}$$

where in general $0_{r\times c}$ is a $r \times c$ matrix of zeros. It follows that

$$\Sigma_\theta = A(\theta)^{-1}B(\theta)A(\theta)^{-T} = \begin{pmatrix} A(\theta^*)^{-1}B(\theta^*)A(\theta^*)^{-T} - A(\theta^*)^{-1}G^T E_{\beta\beta}^{-1} GA(\theta^*)^{-T} & 0_{2\times p} \\ E_{\beta\beta}^{-1}GA(\theta^*)^{-T} - A(\theta^*)^{-T}GE_{\beta\beta}^{-T} & -E_{\beta\beta}^{-T} \end{pmatrix}.$$

Therefore

$$\Sigma_{IPW} = \Sigma_{IPW}^* - ME_{\beta\beta}^{-1}M^T$$

where $M = \tau A(\theta^*)^{-1}G^T$. It is straightforward to show $E_{\beta\beta}^{-1}$ is positive definite, implying $\Sigma_{IPW} \leq \Sigma_{IPW}^*$.

References

Bacon, M. C., von Wyl, V., Alden, C., Sharp, G., Robison, E. and Hessol, N. (2005) The Women's Interagency HIV Study: An observational cohort brings clinical sciences to the bench. *Clinical and Diagnostic Laboratory Immunology*, **12**, 1013–1019.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2010) *Measurement Error in Nonlinear Models: A Modern Perspective*. New York: CRC Press.

Cole, S. R. and Stuart, E. A. (2010) Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, **172**, 107–115.

Centers for Disease Control and Prevention (2012) Diagnoses of HIV infection and AIDS in the United States and dependent areas. *HIV Surveillance Report*, **17**.

Gandhi, M., Ameli, N., Bacchetti, P., Sharp, G. B., French, A. L. and Young, M. (2005) Eligibility criteria for HIV clinical trials and generalizability of results: The gap between published reports and study protocols. *AIDS*, **19**, 1885–1896.

Greenblatt, R. M. (2011) Priority issues concerning HIV infection among women. *Women's Health Issues*, **21**, S266–S271.

Hammer, S. M., Squires, K. E., Hughes, M. D., Grimes, J. M., Demeter, L. M. and Currier, J. S. (1997) A controlled trial of two nucleoside analogues plus indinavir in persons with HIV infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine*, **337**, 725–733.

Hernan, M. A. and VanderWeele, T. J. (2011) Compound treatments and transportability of causal inference. *Epidemiology*, **22**, 368–377.

Hirano, K., Imbens, G. W. and Ridder, G. (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, **71**, 1161–1189.

Keiding, N. and Louis, T. A. (2016) Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **179**, 319–376.

Kitahata, M. M., Rodriguez, B., Haubrich, R., Boswell, S., Mathews, W. C. and Lederman, M. M. (2008) Cohort profile: The Centers for AIDS Research Network of Integrated Clinical Systems. *International Journal of Epidemiology*, **37**, 948–955.

Lunceford, J. K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, **23**, 2937–2960.

O'Muircheartaigh, C. and Hedges, L. V. (2013) Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**, 195–210.

Robins, J. M., Mark, S. D. and Newey, W. K. (1992) Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, **48**, 479–495.

Rubin, D. B. (1978) Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, **7**, 34–58.

— (1980) Comment on "Randomization analysis of experimental data in the Fisher randomization test" by Basu. J. *American Statistical Association*, **75**, 591–593.

Sax, P. E., Tierney, C., Collier, A. C., Daar, E. S., Mollan, K., Budhathoki, C., Godfrey, C., Jahed, N. C., Myers, L., Katzenstein, D. et al. (2011) Abacavir/lamivudine versus tenofovir DF/emtricitabine as part of combination regimens for initial treatment of HIV: Final results. *Journal of Infectious Diseases*, **204**, 1191–1201.

Sax, P. E., Tierney, C., Collier, A. C., Fischl, M. A., Mollan, K., Peeples, L., Godfrey, C., Jahed, N. C., Myers, L., Katzenstein, D. et al. (2009) Abacavir–lamivudine versus tenofovir–emtricitabine for initial HIV-1 therapy. *New England Journal of Medicine*, **361**, 2230–2240.

Scott, A. J. and Wild, C. (1986) Fitting logistic models under case control or choice based sampling. *Journal of the Royal Statistical Society. Series B. Methodological*, **48**, 170–182.

Stuart, E. A., Bradshaw, C. P. and Leaf, P. J. (2015) Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, **16**, 475–485.

Stuart, E. A., Cole, S. R., Bradshaw, C. P. and Leaf, P. J. (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174**, 369–386.

Tipton, E. (2013) Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, **38**, 239–266.

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K. and Caverly, S. (2014) Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, **7**, 114–135.

Westreich, D. and Cole, S. R. (2010) Invited commentary: Positivity in practice. *American Journal of Epidemiology*, **171**, 674–677.

Westreich, D., Lessler, J. and Funk, M. J. (2010) Propensity score estimation: Neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, **63**, 826–833.

Wooldridge, J. M. (2007) Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, **141**, 1281–1301.

Zhang, M., Tsiatis, A. A. and Davidian, M. (2008) Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, **64**, 707–715.