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

Year 2015 *Paper* 45

Generalizing Evidence from Randomized Trials using Inverse Probability of Sampling Weights

Ashley L. Buchanan[*] Michael G. Hudgens[†] Stephen R. Cole[‡]
Katie Mollan[**] Paul E. Sax[††] Eric Daar[‡‡]
Adaora A. Adimora[§] Joseph Eron[¶] Michael Mugavero[‖]

[*]Harvard University, buchanan@hsph.harvard.edu
[†]University of North Carolina at Chapel Hill, mhudgens@email.unc.edu
[‡]University of North Carolina at Chapel Hill, cole@unc.edu
[**]University of North Carolina at Chapel Hill, kmollan@email.unc.edu
[††]Brigham and Womens Hospital and Harvard Medical School, psax@partners.org
[‡‡]Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, edaar@labiomed.org
[§]University of North Carolina at Chapel Hill, adimora@med.unc.edu
[¶]University of North Carolina at Chapel Hill, joseph_eron@med.unc.edu
[‖]University of Alabama, mmugavero@uab.edu

http://biostats.bepress.com/uncbiostat/art45

Generalizing Evidence from Randomized Trials using Inverse Probability of Sampling Weights

Ashley L. Buchanan, Michael G. Hudgens, Stephen R. Cole, Katie Mollan, Paul E. Sax, Eric Daar, Adaora A. Adimora, Joseph Eron, and Michael Mugavero

Abstract

Results obtained in randomized trials may not generalize to specific target populations. In a randomized trial, the treatment assignment mechanism is known, but assuming participants are a random sample from the target population is often dubious. Lack of generalizability can occur when the distribution of treatment effect modifiers in trial participants differs from the distribution in the target population. We consider an inverse probability of sampling weighted (IPSW) estimator for generalizing trial results to a user-specified target population that differs in important clinical or demographic characteristics from the randomized trial. The IPSW estimator is shown to be consistent and asymptotically normal assuming a model for the sampling score (i.e., the probability of participating in the trial) is correctly specified. Expressions for the asymptotic variance and a consistent sandwich-type estimator of the variance are derived. Simulation results comparing the IPSW estimator and a previously proposed stratified estimator show that the estimators perform similarly when the sampling score model includes a binary covariate. However, with a continuous covariate in the sampling score model, the IPSW estimator is less biased and the corresponding Wald confidence interval has better coverage. The IPSW estimator is employed to generalize results from two randomized trials of HIV treatment conducted by the United States (US) National Institutes of Health AIDS Clinical Trials Group to all people currently living with HIV in the US.

Generalizing Evidence from Randomized Trials using Inverse Probability of Sampling Weights

Ashley L. Buchanan

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

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 Mollan

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

Paul E. Sax

Division of Infectious Diseases and Department of Medicine, Brigham and Womens Hospital, and Harvard Medical School

Eric Daar

Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, California

Adaora A. Adimora

Department of Internal Medicine, University of North Carolina

Joseph J. Eron

Department of Medicine, University of North Carolina

Michael J. Mugavero

School of Medicine, University of Alabama

†

†*Address for correspondence:* 135 Dauer Drive, 3101 McGavran-Greenberg Hall CB 7420, Chapel Hill, NC 27599 − 7420

E-mail: mhudgens@bios.unc.edu

Summary.

Results obtained in randomized trials may not generalize to specific target populations. In a randomized trial, the treatment assignment mechanism is known, but assuming participants are a random sample from the target population is often dubious. Lack of generalizability can occur when the distribution of treatment effect modifiers in trial participants differs from the distribution in the target population. We consider an inverse probability of sampling weighted (IPSW) estimator for generalizing trial results to a user-specified target population that differs in important clinical or demographic characteristics from the randomized trial. The IPSW estimator is shown to be consistent and asymptotically normal assuming a model for the sampling score (i.e., the probability of participating in the trial) is correctly specified. Expressions for the asymptotic variance and a consistent sandwich-type estimator of the variance are derived. Simulation results comparing the IPSW estimator and a previously proposed stratified estimator show that the estimators perform similarly when the sampling score model includes a binary covariate. However, with a continuous covariate in the sampling score model, the IPSW estimator is less biased and the corresponding Wald confidence interval has better coverage. The IPSW estimator is employed to generalize results from two randomized trials of HIV treatment conducted by the United States (US) National Institutes of Health AIDS Clinical Trials Group to all people currently living with HIV in the US.

Keywords: Causal inference; External validity; Generalizability; HIV/AIDS; Inverse probability weights; Propensity score; Target population

1. Introduction

Generalizability is a concern for many scientific studies, including those in public health and medicine. Using information in the study sample, it is often of interest to draw inference about a specified target population. Generalizability is defined as the degree to which an effect estimated in a sample approximates the true measure of effect in the target population. 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. One study highlighted the overrepresentation 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 (Greenblatt, 2011). The Women's Interagency HIV Study (WIHS) is a prospective, observational, multicenter study of women living with HIV and women at risk for HIV infection in the U.S. (Bacon et al., 2005). Another study reviewed eligibility criteria of 20 AIDS Clinical Trial Group (ACTG) studies and found that 28% to 68% of the WIHS cohort would have been excluded (Gandhi et al., 2005).

There exist several quantitative methods which employ sampling scores to generalize results from a trial to a target population. Here, the sampling score is defined as the probability of participation in the trial conditional on covariates. These approaches 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 an inverse-probability-of-sampling-weighted Cox proportional hazards model was fit to draw inference about the effect of treatment in the target population. A robust estimate of the variance was employed (Robins, 1998); however, no closed-form expression for the variance was provided. As an alternative, a sampling score stratified estimator was proposed to generalize trial results (Tipton, 2013; O'Muircheartaigh and Hedges, 2013; Tipton et al., 2014). To date there has been no formal studies or derivations of the large sample statistical properties of these generaliziability estimators (i.e., consistency and asymptotic normality).

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 for univariate outcomes to a target population. The inverse weighted estimator is compared to the stratified estimator. In Section 2, the assumptions and notation are discussed. The inverse probability of sampling weighted (IPSW) estimator and the stratified estimator are described in Section 3. In Section 4 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 5. In Section 7 the IPSW estimator is applied to generalize results from two ACTG trials to all people currently living with HIV in the US. Section 7 concludes with a discussion.

2. Assumptions and Notation

Suppose we are interested in drawing inference about the effect of some 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, 1980) holds, i.e., there are no variations of treatment and there is no interference between individuals (i.e., the outcome of one individual is assumed to be unaffected by treatment assignment of others). 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 datasets are available. A random sample (e.g., cohort study) of m individuals is drawn from the near infinite target population. A second sample of n individuals participate in a randomized trial, and the treatment assignment mechanism is known to the analyst. 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. In general, let upper case letters denote random

variables and lower case letters denote realizations of those random variables. Define Z as a $1 \times p$ vector of fixed characteristics 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.

Once in the trial assume participants are randomly assigned to received 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, so $P(S = s|\mathbf{Z}, Y^0, Y^1) = P(S = s|Z)$. In other words, participants in the trial are no different from nonparticipants in regards to the treatment-outcome relationship conditional on $\mathbf{Z}$. Trial participation and treatment positivity are also assumed, so $P(S = s|Z) > 0$ and $P(X = x|Z, S = 1) > 0$ for all $Z = z$. Assume participants in the trial are adherent to their treatment assignment (i.e., ignoring noncompliance issues) and the model for the sampling scores is correctly specified (e.g., correct covariate functional forms).

3. Estimators of the Population Average Treatment Effect

A traditional (i.e., unweighted) approach to estimating treatment effects is a difference in means. 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 can be 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 which do not assume trial participants are a random sample from the target population. Both estimators utilize sampling scores. In practice, the sampling scores are likely unknown and can be estimated using a parametric model. Let $\boldsymbol{\beta}$ be the $1 \times p$ vector of coefficients in the logistic regression model and $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimator of $\boldsymbol{\beta}$. Following Cole and Stuart (2010), the sampling scores $P(S = 1|Z = Z) = \{1 + \exp(-Z\boldsymbol{\beta})\}^{-1}$ are estimated using logistic regression. In particular, let $P(S = 1|Z = Z) = w(z, \boldsymbol{\beta})$, $w_i = w(Z_i, \boldsymbol{\beta})$, and $\hat{w}_i = w(Z_i, \hat{\boldsymbol{\beta}})$. To account for the random sampling of the cohort from the target population when estimating $\boldsymbol{\beta}$, each individual in the cohort is inverse weighted by the sampling fraction $r_i = m/(N - n)$, where N is the size of the target population with $N >> n$ and $N >> m$, and each trial participant is given a weight of $r_i = 1$. Following Cole and Stuart (2010), the IPSW estimator of the PATE is

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

An alternative approach for estimating the PATE uses stratification based on the sampling scores (Tipton, 2013; O'Muircheartaigh and Hedges, 2013; Tipton et al., 2014). This estimator is computed in the following steps. First, $\boldsymbol{\beta}$ is estimated using a logistic regression model and the sampling scores $\hat{w}_i$ are estimated. These estimated sampling scores are used to form L strata according to the distribution in the target population. The distribution of sampling scores in the combined trial and cohort are used to estimate the strata (Tipton, 2013). The difference of sample means within each stratum is computed among those in the trial. Lastly, the PATE is estimated as a weighted sum of the differences of sample means across strata, where the weight ω_l is the proportion of observations in stratum l in the target population. Let n_l be the number in the trial in stratum l and m_l be the number in the cohort in stratum l. Let S_{il} 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

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

where the L stratum are defined by the distribution of the sampling scores in the target population, $l = 1, \ldots, L$ and $i = 1, \ldots, (n + m)$ and $\omega_l = N_l/N$ with N_l as number in stratum l in the target population.

4. Large Sample Properties of the IPSW Estimator

Let Δ_0 be the true value of Δ. Let $w_0 = w(Z_i, \boldsymbol{\beta}_0)$ be the true weight. Because the trial participants are not assumed to be a random sample from the target population, the observed data $(S_i, Z_i, S_i X_i, S_i Y_i)$ are assumed to be independent but not necessarily identically distributed. Below we express the IPSW estimator 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 $\boldsymbol{\beta}$ is known, so the solution does not require a score equation for the sampling score model. Let $\hat{\boldsymbol{\theta}}^* = (\hat{\mu}_1, \hat{\mu}_0)$ and $\boldsymbol{\theta}_0^* = (\mu_1, \mu_0)$, then $\hat{\theta}^*$ is the solution to the estimating equation

$$\sum_i \Psi_\Delta^*(Y_i, \mathbf{Z}_i, X_i, S_i, \theta_0^*) = \begin{pmatrix} \sum_i [S_i X_i (Y_i - \mu_1)]/w_i \\ \sum_i [S_i (1 - X_i)(Y_i - \mu_0)]/w_i \end{pmatrix} = 0$$

The expectation of $\Psi_\Delta^*(Y_i, \mathbf{Z}_i, X_i, S_i, \theta_0^*)$ is zero at the true value Δ_0, so $\hat{\theta}^*$ converges in probability to $\boldsymbol{\theta}_0^*$ (Stefanski and Boos, 2002). Define the following matrices:

$$\mathbf{A}(\boldsymbol{\theta}_0^*) = (n + m)^{-1} \sum_i E\left[\partial / \partial \boldsymbol{\theta}_0^* \Psi_\Delta^*(Y_i, \mathbf{Z}_i, X_i, S_i, \theta_0^*)\right]$$

$$\mathbf{B}(\boldsymbol{\theta}_0^*) = (n + m)^{-1} \sum_i E\{\mathrm{cov}\left[\Psi_\Delta^*(Y_i, \mathbf{Z}_i, X_i, S_i, \theta_0^*)\right]\}$$

Then, $\hat{\boldsymbol{\theta}}^*$ is asymptotically normally distributed with mean $\boldsymbol{\theta}_0^*$ and covariance matrix $\Sigma_\theta^* = (n+m)^{-1}\mathbf{A}^{-1}(\boldsymbol{\theta}_0^*)\mathbf{B}(\boldsymbol{\theta}_0^*)\mathbf{A}^{-T}(\boldsymbol{\theta}_0^*)$. Because $\hat{\Delta}_{IPW}$ is a linear combination of $\hat{\boldsymbol{\theta}}^*$, $\hat{\Delta}_{IPW}$ is a consistent estimator of Δ_0. Furthermore, $(n+m)^{1/2}(\hat{\Delta}_{IPW} - \Delta_0)$ converges in distribution to $N(0, \Sigma_{IPW}^*)$ (Carroll et al. 2010, Appendix A.6) and the sandwich-type estimator of the variance $\hat{\Sigma}_{IPW}^*$ (see Appendix) is consistent for Σ_{IPW}^*, under the suitable regularity conditions as $n, m \to \infty$ and $n/(n+m) \to c$ with $0 < c \leq 1$. When $\boldsymbol{\beta}_{1\times p}$ is known, by the delta method, it follows that $\hat{\Delta}_{IPW}$ is asymptotically normal with asymptotic variance

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

where $\Sigma^{*(ij)}$ refers to the i^{th} row and the j^{th} column of the matrix Σ_{IPW}^*.

In the more likely case that $\boldsymbol{\beta}$ is unknown, an additional estimating equation for each element of $\boldsymbol{\beta}$ is needed. Using M-estimation, this suggests that the estimating equation based on the score function of the logistic regression model can be used to obtain the consistent sandwich-type estimator of the variance (Carroll et al., 2010; Stefanski and Boos, 2002). The vector of parameters $\boldsymbol{\beta}$ can be consistently estimated by solving the estimating equations

$$\sum_i \psi_\beta(S_i, \mathbf{Z}_i, \boldsymbol{\beta}) = \sum_i r_i^{-1} \frac{S_i - w_i}{w_i(1-w_i)} \frac{\partial}{\partial \boldsymbol{\beta}} w_i = \mathbf{0}$$

(Manski and Lerman, 1977; Scott and Wild, 1986, 2002). Let $\hat{\boldsymbol{\theta}} = (\hat{\mu}_1, \hat{\mu}_0, \hat{\boldsymbol{\beta}})$ and $\boldsymbol{\theta}_0 = (\mu_1, \mu_0, \boldsymbol{\beta})$. $\hat{\boldsymbol{\theta}} = (\hat{\mu}_1, \hat{\mu}_0, \hat{\boldsymbol{\beta}})$ is the solution to the estimating equation

$$\sum_i \Psi_\Delta(Y_i, \mathbf{Z}_i, X_i, S_i, \Delta, \boldsymbol{\beta}) = \begin{pmatrix} \sum_i [S_i X_i (Y_i - \mu_1)]/w_i \\ \sum_i [S_i (1-X_i)(Y_i - \mu_0)]/w_i \\ \sum_i \psi_\beta(S_i, \mathbf{Z}_i, \boldsymbol{\beta}) \end{pmatrix}$$

Define the following matrices: $\mathbf{A}(\boldsymbol{\theta}_0) = (n+m)^{-1}\sum_i E\left[\partial/\partial\boldsymbol{\theta}_0 \Psi_\Delta(Y_i, \mathbf{Z}_i, X_i, S_i, \Delta)\right]$ and $\mathbf{B}(\boldsymbol{\theta}_0) = (n+m)^{-1}\sum_i E\{\mathrm{cov}\left[\Psi_\Delta(Y_i, \mathbf{Z}_i, X_i, S_i, \Delta)\right]\}$. Then, $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and covariance matrix $\Sigma_\theta = (n+m)^{-1}\mathbf{A}^{-1}(\boldsymbol{\theta}_0)\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}^{-T}(\boldsymbol{\theta}_0)$. Therefore, by the delta method, it follows that $\hat{\Delta}_{IPW}$ is asymptotically normal with asymptotic variance

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

where $\Sigma^{(ij)}$ refers to the i^{th} row and the j^{th} column of the matrix Σ_{IPW}.

By comparison of equations (2) and (3), it follows that the variance is smaller when the sampling scores are estimated because $\Sigma_\theta^{(12)}$ is larger than $\Sigma_\theta^{*(12)}$. 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). Even if the correct sampling scores are known, estimation of the sampling scores is preferable due to improved efficiency. It is common practice to compute the variance using standard software assuming the weights are known. This leads to valid, but conservative confidence intervals. The consistent sandwich-type estimators of the variance of $\hat{\Delta}_{IPW}$ are provided in the Appendix.

In the Electronic Supplement, an R function is provided to compute the IPSW estimator and its corresponding sandwich-type estimator of the variance.

5. Estimator of the Variance of the Stratified Estimator

One approach to obtain an estimator of the variance of the stratified estimator is to employ estimating equations, which include an estimating equation for the means, quintiles, proportion in each quintile, and each element of β. This approach also demonstrates that the estimator is asymptotically normal; however, this property depends critically on the density of the propensity score (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). Define the quintiles of $\hat{w}_i$, where the l^{th} sample quintile is $\hat{q}_l$, $l = 1, \ldots, L$, such that the proportion of $\hat{w}_i \leq \hat{q}_l$ is roughly l/L in the target. Since we assume the trial and cohort both arise from the same target, the distribution of sampling scores in the combined trial and target are used to estimate the quintiles (Tipton, 2013). In practice, the cohort data will need to be weighted to get the correct distribution of the sampling scores in the target. Let $\hat{q}_0 = 0$ and $\hat{q}_L = 1$. Define $\hat{Q}_l = (\hat{q}_{l-1}, \hat{q}_l)$. Let $N_l = \sum_{i=1}^{N} I(\hat{w}_i \in \hat{Q}_l)$ be the number of individuals in stratum l in the target. Let $n_l = \sum_{i=1}^{n+m} S_i I(\hat{w}_i \in \hat{Q}_l)$ be the number of individuals in stratum l who are selected into the trial. Let $n_{1l} = \sum_{i=1}^{n+m} S_i X_i I(\hat{w}_i \in \hat{Q}_l)$ be the number of individuals in stratum l who are selected into the trial and randomized to treatment. The approximate variance of $\hat{\Delta}_S$ is

$$L^{-2} \sum_{l=1}^{L} \hat{\sigma}_l^2$$

assuming an equal number of participants in each stratum (Tipton, 2013), where $\hat{\sigma}_l^2 = n_{1l}^{-1} s_{1l}^2 + (n_l - n_{1l})^{-1} s_{0l}^2$, $s_{1l}^2 = n_{1l}^{-1} \sum_{i=1}^{n} I(\hat{w}_i \in \hat{Q}_l)(X_i Y_i - \bar{y}_{1l})^2$, $s_{0l}^2 = (n_l - n_{1l})^{-1} \sum_{i=1}^{n} I(\hat{w}_i \in \hat{Q}_l)((1-X_i)Y_i - \bar{y}_{0l})^2$, $\bar{y}_{1l} = n_{1l}^{-1} \sum_{i=1}^{n} I(\hat{w}_i \in \hat{Q}_l) X_i Y_i$, and $\bar{y}_{0l} = (n_l - n_{1l})^{-1} \sum_{i=1}^{n} I(\hat{w}_i \in \hat{Q}_l)(1 - X_i) Y_i$. This estimator of the variance is conservative because it does not account for estimation of the additional parameters β, Q, or n_{1l} (Lunceford and Davidian, 2004).

6. Simulations

A simulation study was conducted to compare the performance of the IPSW and stratified estimators and included scenarios with a continuous or discrete covariate and a continuous response. The following quantities were computed in the simulated datasets: the bias for each estimator, which was the difference between the average of the estimated difference in means and the true difference in means, standard error, which was the average of the estimated standard errors, Monte Carlo standard error, which was the standard deviation of the estimated difference in means, and empirical coverage probability, which was the proportion of times the 95% confidence interval contained the true difference in means.

A total of 5,000 datasets per scenario were simulated as follows. There were $N = 10^6$ observations in the target population and each had (Z_{1i}, w_i), where the true sampling score was $w_i = \{1 + \exp(-\beta_0 - \beta_1 Z_{1i})\}^{-1}$. In the first two scenarios, one binary covariate $Z_{1i} \sim \text{Bern}(0.2)$ 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 to ensure that the probability of sampling into the trial was a rare event (i.e., the size of the trial was approximately $n \approx 1{,}000$). 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.

To estimate the weights, the combined trial ($S_i = 1$) and cohort data ($S_i = 0$) was used to fit a (weighted) logistic regression model with S_i as the outcome and the covariate Z_{1i}. To account for the sampling of the cohort from the target, each participant in the cohort was inverse weighted by $\hat{r}_i = m/(N-n)$. Each trial participant was given a weight of $\hat{r}_i = 1$ in the logistic model. A weighted score equation for the logistic regression model was included in the computation of the sandwich-type estimator of the variance for the IPSW estimator. This allowed for unbiased estimation of the parameters in the logistic regression model, as well as the correct information for computation of the variance estimator of $\hat{\Delta}_{IPW}$.

For the stratified estimator, the distribution of the sampling scores in the target population was needed. The quintiles and number within each sampling score stratum were obtained from the inverse weighted data. The approximate estimator of the variance was employed (i.e., the average variance across sampling score strata).

For those included in the randomized trial ($S_i = 1$), X_i was generated as Bern(0.5) and the response 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 (i.e., weak or moderate Z and S association): Scenario 1, 3, and 5 set $\boldsymbol{\beta} = (-7, 0.4)$; Scenario 2, 4, and 6 set $\boldsymbol{\beta} = (-7, 0.6)$. The truth was calculated for each scenario using the distribution of Z in the target population. The truth was $\Delta_0 = 2.2$ for scenarios 1 and 2 and $\Delta_0 = 2$ for scenarios 3 through 6.

Comparisons between the IPSW and stratified estimator when the sampling score model is correctly specified are summarized in Table 1. The estimated sampling scores were computed using logistic regression with the covariate Z_{1i}. The within trial estimator was biased for all scenarios (Table 1) and had low coverage (results not shown). Depending on the scenario, the size of the trial ranged from $n = 987$ to $n = 1{,}091$ participants on average over the simulations for each scenario. For all scenarios, $\hat{\Delta}_{IPW}$ was unbiased. For scenarios 1 to 2, $\hat{\Delta}_S$ was unbiased and standard errors

were comparable for the two estimators. 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 Monte Carlo standard error, supporting the derivations of the sandwich-type estimator of the variance. Coverage was around 95% for the Wald confidence interval of $\hat{\Delta}_{IPW}$ for all scenarios. With a continuous covariate, the Wald confidence interval of the stratified estimator had poor coverage, particularly in the presence of stronger effect modification (e.g., scenarios 5 and 6). Upon visual inspection, the IPSW estimator appeared to be normally distributed (Figure 1).

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, $Z_{2i} \sim \text{Bern}(0.6)$, 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 Bern(0.5) and the response 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 (i.e., weak (w) or moderate (m) Z and S association): Scenario 1, 3, and 5 set $\boldsymbol{\beta} = (-7, 0.4)$; Scenario 2, 4, and 6 set $\boldsymbol{\beta} = (-7, 0.6)$. The truth was calculated for each scenario using the distribution of $\boldsymbol{Z}$ in the target population. The truth was $\Delta_0 = 2.8$ for scenarios 1 and 2 and $\Delta_0 = 2$ for scenarios 3 through 6.

When the sampling score model is misspecified, comparisons between the IPSW and stratified estimator are summarized in Table 2. 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 empirical sandwich-type estimator of the variance of the IPSW estimator performed reasonably well when the sampling score model was misspecified; however, coverage was below the nominal level.

7. Applications

In this section, the methods described in the previous sections were applied to generalize results from two different ACTG randomized clinical trials, ACTG 320 and ACTG A5202. The methods in this paper were developed for continuous outcomes, so this application focused on generalizing results for continuous outcomes in the trials. Results from these two trials were generalized to two different target populations, 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 (Hammer et al., 1997). In ACTG 320, 200 women were enrolled (Hammer et al., 1997). The baseline characteristics of these women and all participants are shown in Supplemental Tables 1 and 2, respectively.

The ACTG A5202 trial examined 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 (Sax et al., 2009, 2011). 322 women were enrolled in ACTG A5202 (Sax et al., 2009, 2011). The baseline characteristics are shown in Supplemental Table 3 among women and in Supplemental Table 4 among all participants.

WIHS and Center for AIDS Research Network of Integrated Clinical Systems (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). The analysis for ACTG 320 only included cohort participants who were HIV-positive, highly active antiretroviral therapy (HAART) naive, and had CD4 cell counts ≤ 200 cells/mm^3 at the previous visit ($m = 493$ and $m = 6{,}158$, respectively). The analysis for A5202 included cohort participants who were HIV-positive, antiretroviral (ART) naive, and had viral load $> 1{,}000$ copies/ml at the previous visit ($m = 1{,}012$ and $m = 12{,}302$, respectively). The WIHS is a prospective, observational, multicenter study of women living with HIV and women at risk for HIV infection in the U.S. (Bacon et al., 2005). A total of 4,129 women (1,065 HIV-uninfected) were enrolled between October 1994 and December 2012 at six US sites. Supplemental Table 1 displays the characteristics of the women in the WIHS sample for ACTG 320. Supplemental Table 3 displays the characteristics of the women in the WIHS sample for ACTG A5202.

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). For this analysis, CNICS is considered to be representative of all people living with HIV and in clinical care in the US. The CNICS cohort includes over 27,000 HIV-infected adults (at least 18 years of age) engaged in clinical care since January 1, 1995 at eight CFAR sites in the US. Supplemental Table 2 displays the characteristics of participants in the CNICS sample for ACTG 320. Supplemental Table 4 displays the characteristics of participants in the CNICS sample for ACTG A5202.

The IPSW estimator was employed to assess the generalizability of 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 U.S. and among all participants in the trials to all people currently living with HIV in the U.S. Based on CDC 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 (CDC, 2012).

First, presence of conditions that could induce a lack of generalizability was assessed in the datasets. Namely, variables associated with trial participation that are also treatment effect mod-

ifiers were identified. Among women in ACTG 320 and WIHS participants, CD4 at baseline was both associated with trial participation (P = 0.003) and an effect modifier (P = 0.003). There were differences in the point estimates of treatment effects across levels of all four covariates (Supplemental Figure 1). Among ACTG 320 participants and CNICS participants, race/ethnicity was both associated with trial participation (P < 0.001) and an effect modifier (P = 0.05). There were differences in the point estimates of treatment effects across levels of all five covariates (Supplemental Figure 2). Among women in ACTG A5202 and WIHS participants, age, history of IDU, and hepatitis were both associated with trial participation (P < 0.001, P < 0.001, and P = 0.003, respectively) and effect modifiers (P = 0.02, P = 0.03, and P = 0.04, respectively). There were differences in the point estimates of treatment effects across levels of all seven covariates (Supplemental Figure 3). Among ACTG A5202 participants and CNICS participants, history of IDU and baseline CD4 were both associated with trial participation (P < 0.001 for each variable) and effect modifiers (P = 0.007 and P = 0.05, respectively). There were differences in the point estimates of treatment effects across levels of all covariates, except AIDS diagnosis (Supplemental Figure 4).

Second, the within-trial treatment effects were computed separately among women only and all participants (Table 3). This was an as-treated analysis and ignored treatment compliance issues. Among participants and among women in ACTG 320, there was a significant increase in CD4 at week 4. 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.

Third, 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 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, and overlap between the trial and cohort was assumed to be negligible. A logistic regression model was fit on the combined trial and weighted cohort data. 116 (10%) of ACTG 320 participants were missing CD4 count at week 4, so they were excluded. 417 (22%) of ACTG A5202 trial participants were missing CD4 count at week 48, so they were excluded. Cohort participants were inverse weighted by the size of the cohort divided by the size of the target and trial participants were given a weight of 1. The outcome was trial participation and the possible covariates were sex, race/ethnicity, age, history of IDU, and baseline CD4 for ACTG 320 and sex, race/ethnicity, age, history of IDU, hepatitis B/C, AIDS diagnosis, baseline CD4 and baseline $\log_{10}$ viral load for ACTG A5202. Variables associated with trial participation, the outcome, or effect modifiers, as well as all pairwise interactions, were included in the sampling score model. Due to positivity, sex was excluded from the analysis generalizing the trial results among women.

Table 3 displays the results for the two ACTG trials generalized to both target populations.

Among women in ACTG 320 at week 4, those randomized to the regimen with a PI had an average change in CD4 cell count 46 cells/mm^3 higher than women randomized to regimen without a PI (95% confidence interval (CI) = $(23, 70)$). Among all participants in ACTG 320 at week 4, those randomized to a regimen with a PI had a change in an average CD4 cell count 17 cells/mm^3 higher than those randomized to a regimen without a PI (95% CI = $(9, 25)$). In ACTG A5202, women randomized to ABC-3TC had an average change in CD4 cell count 35 cells/mm^3 higher than women randomized to TDF-FTC (95% CI = $(-45, 115)$). Among all participants in ACTG A5202, those randomized to ABC-3TC had an average change in CD4 cell count 2 cells/mm^3 lower than those randomized to TDF-FTC (95% CI = $(-31, 28)$).

8. Discussion

In this paper, we considered an estimator using inverse probability of sampling weights to generalize results from a randomized trial to a specific target population. The IPSW estimator and corresponding confidence interval provide inference about the effect of treatment in the target population, i.e., a contrast in the average outcome had (contrary to fact) everyone in the target population received treatment compared to if everyone in the target population did not receive treatment. The IPSW estimator was shown to be consistent and asymptotically normal and a consistent sandwich-type estimator of the variance was provided. In the illustrative example, the IPSW estimator was employed to generalize results from the ACTG to all people currently living with HIV in the US. For ACTG 320, the within-trial effect was comparable to the effect estimated with the IPSW, so the results appear to be generalizable to all people living with HIV in the US. On the other hand, the within trial effect from ACTG A5202 was not comparable to the effect estimate based on the IPSW estimator. For the A5202 results among women, the difference in the effect estimates is primarily due to hepatitis, which was negatively associated with participation in the trial and a treatment effect modifier. Results from both ACTG A5202 and ACTG 320 were not sensitive to the specification of the size of the target population; however, some results were sensitive to the specification of the sampling score model. For the sake of focusing on generaliziability in the example, the missing information on the outcome was ignored in the analysis; however, in practice, one would want to address the possibly not missing (completely) at random data.

When applying this method, the analysis is subject to the following considerations. The absence of unmeasured covariates associated with the trial participation mechanism and treatment effect modifiers is an untestable assumption. Treatment compliance issues were ignored in this method; however, this issue should be considered in analyses. The sampling score model was assumed to be correct (i.e., correct covariate functional forms); however, this is not guaranteed in practice. The stratified estimator (Tipton et al., 2014; O'Muircheartaigh and Hedges, 2013) requires that individuals sharing the same stratum of the distribution of sampling scores can be identified, which may be

difficult in practice. This estimator may be biased when there is residual confounding within strata and, therefore, is not a consistent estimator of the PATE in some cases (e.g., a continuous covariate in the sampling score model) (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 we do not believe the cohort is representative, one possibility is weighting the cohort data to the distribution of covariates in a census (e.g., CDC estimates). The downside of this is the census may not have covariate information as rich as the cohort data. The CDC estimates used in the example were for all people living with HIV. A future refinement could be to use surveillance studies that would report on the number of ART and HAART naive HIV patients in the U.S.

Weighted logistic regression was used as an approach to consistently estimate the parameters of the logistic regression model (e.g., the intercept); however, other approaches may be possible. Additional research to develop an augmented estimator could improve efficiency (Zhang et al., 2008). This method could be extended to accommodate the presence of interference. Lastly, this method holds for continuous outcomes. Further results are needed for estimation with right-censored data.

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 Womens Interagency HIV Study (WIHS). The contents of this publication are solely the responsibility of the authors and do not represent the official views of the National Institutes of Health (NIH). WIHS (Principal Investigators): UAB-MS WIHS (Michael Saag, MirjamColette 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).

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 samples with m = 4,000 and $n \approx$ 1,000. Scenarios are described in Section 6. For scenarios 1 and 2, $\Delta_0 = 2.2$ and, for scenarios 3 to 6, $\Delta_0 = 2.0$ (T = within trial; S = stratified; $IPSW$ = inverse probability of sampling weighted; ESE = Empirical standard error (×100); ASE = Average standard error (×100); ECP = Empirical coverage probability)

			Bias			**ESE**		**ASE**		**ECP**	
Scenario	Cov.	(β_1,α)	$\hat{\Delta}_T$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPSW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPSW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPSW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPSW}$
1	Bin.	(0.4,1)	0.07	2e-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-3	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 samples with m = 4,000 and $n \approx$ 1,000. Scenarios are described in Section 6. For scenarios 1 and 2, $\Delta_0 = 2.8$ and, for scenarios 3 to 6, $\Delta_0 = 2.0$ (T = within trial; S = stratified; $IPSW$ = inverse probability of sampling weighted; ESE = Empirical standard error (×100); ASE = Average standard error (×100); ECP = Empirical coverage probability)

			Bias			**ESE**		**ASE**		**ECP**	
Scenario	Cov.	(β_1,α)	$\hat{\Delta}_T$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPSW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPSW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPSW}$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPSW}$
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: Results for continuous outcomes in two AIDS Clinical Trials Group (ACTG) trials where the sampling score model included all variables associated with trial participation, the outcome, or effect modifiers (with a linear term for continuous variables) and all pairwise interactions (T = within trial; S = stratified; IPSW = inverse probability of sampling weighted).

		Difference in Means (95 % CI)		
Cohort	**Trial**	$\hat{\Delta}_T$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPSW}$
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 ACTG 320, the treatment contrast was PI vs. no PI.

[b]For A5202, the treatment contrast was ABC-3TC vs. TDF-FTC.

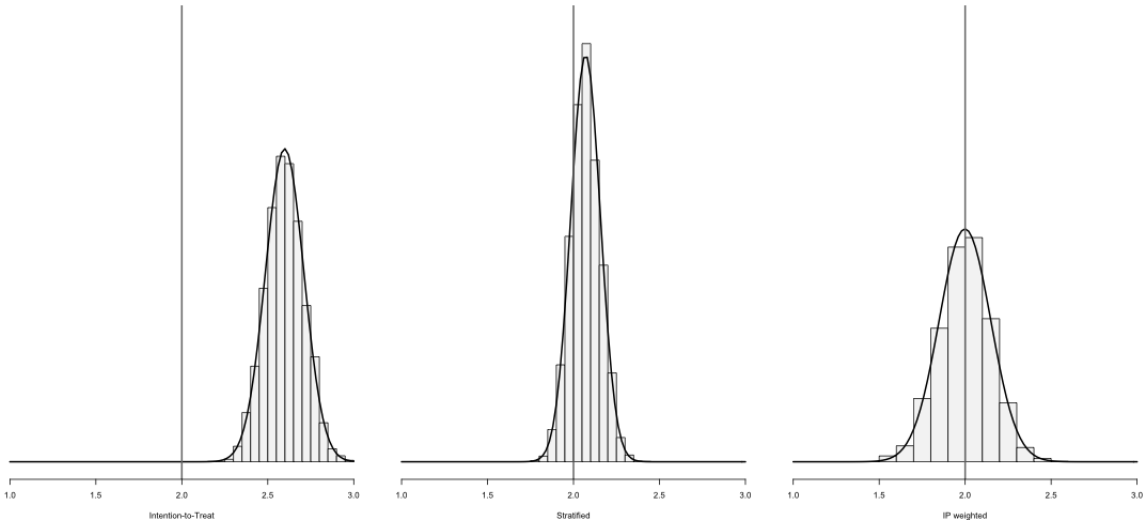

Fig. 1: Comparison of the distributions of within-trial estimator $\hat{\Delta}_T$, stratified estimator $\hat{\Delta}_S$, and inverse probability of sampling weighted estimator $\hat{\Delta}_{IPSW}$, based on 5,000 simulated datasets where the sampling score model is correctly specified and $\Delta_0 = 2.0$ with one continuous covariate, $\boldsymbol{\beta} = (-7, 0.6)$ and $\alpha = 1$.

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

The empirical sandwich-type estimator is used to estimate the asymptotic variance of the IPSW estimator. Substituting the following empirical estimates for their corresponding quantities in equation (2) produces a consistent sandwich-type estimator of the variance when $\boldsymbol{\beta}$ is known. Define the following matrices: $\hat{\mathbf{A}}^* = (n+m)^{-1}\sum_i \partial/\partial\boldsymbol{\theta}_0^*\Psi_\Delta^*(Y_i, \mathbf{Z}_i, X_i, S_i, \hat{\boldsymbol{\theta}}^*)$ and $\hat{\mathbf{B}}^* = (n+m)^{-1}\sum_i \Psi_\Delta^*(Y_i, \mathbf{Z}_i, X_i, S_i, \hat{\boldsymbol{\theta}}^*)\Psi_\Delta^{*^T}(Y_i, \mathbf{Z}_i, X_i, S_i, \hat{\boldsymbol{\theta}}^*)$. $\hat{\boldsymbol{\theta}}^*$ is asymptotically normally distributed with mean $\boldsymbol{\theta}_0^*$ and covariance matrix $\hat{\Sigma}_\theta^* = \hat{\mathbf{A}}^{*^{-1}}\hat{\mathbf{B}}^*\hat{\mathbf{A}}^{*-T}$. When $\boldsymbol{\beta}$ is known, the estimator of the asymptotic variance of $\hat{\Delta}_{IPW}$ is

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

where $\hat{\Sigma}_\theta^{*(ij)}$ refers to the i^{th} row and the j^{th} column of the matrix $\hat{\Sigma}_{IPW}^*$. The estimated standard error is $\hat{\text{se}}(\hat{\Delta}) = \sqrt{(n+m)^{-1}\hat{\Sigma}_{IPW}^*}$.

Similarly, when the weights are estimated, the following expressions can be used to obtain a consistent sandwich-type estimator of the variance. Let $\hat{\boldsymbol{\theta}} = (\hat{\mu}_1, \hat{\mu}_0, \hat{\boldsymbol{\beta}})$ and $\boldsymbol{\theta}_0 = (\mu_1, \mu_0, \boldsymbol{\beta}_0)$. Define the following matrices: $\hat{\mathbf{A}} = (n+m)^{-1}\sum_i \frac{\partial}{\partial\boldsymbol{\theta}_0}\Psi_\Delta(Y_i, \mathbf{Z}_i, X_i, S_i, \hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{B}} = (n+m)^{-1}\sum_i \Psi_\Delta(Y_i, \mathbf{Z}_i, X_i, S_i, \hat{\boldsymbol{\theta}})\Psi_\Delta^{*^T}(Y_i, \mathbf{Z}_i, X_i, S_i, \hat{\boldsymbol{\theta}})$. $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and covariance matrix $\hat{\Sigma}_\theta = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-T}$. When $\boldsymbol{\beta}$ is not known, the estimator of the asymptotic variance of $\hat{\Delta}_{IPW}$ is

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

where $\hat{\Sigma}_\theta^{(ij)}$ refers to the i^{th} row and the j^{th} column of the matrix $\hat{\Sigma}_{IPW}$. The estimated standard error is $\hat{\text{se}}(\hat{\Delta}) = \sqrt{(n+m)^{-1}\hat{\Sigma}_{IPW}}$.

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.

CDC (2012) Diagnoses of HIV infection and AIDS in the United States and dependent areas. *HIV Surveillance Report*, **17**.

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.

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.

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

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.

Manski, C. F. and Lerman, S. R. (1977) The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, 1977–1988.

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. (1998) Marginal structural models. In *Proceedings of the Section on Bayesian Statistical Science*, 1–10. Alexandria, VA.

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

Rubin, D. B. (1980) Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the 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. and Wild, C. (2002) On the robustness of weighted methods for fitting models to case–control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 207–219.

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.

Stefanski, L. A. and Boos, D. D. (2002) The calculus of M-estimation. *The American Statistician*, **56**, 29–38.

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.

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.