ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2018)

Generalizing evidence from randomized trials using inverse probability of sampling weights

Ashley L. Buchanan,

University of Rhode Island, Kingston, USA

Michael G. Hudgens, Stephen R. Cole and Katie R. Mollan,

University of North Carolina at Chapel Hill, USA

Paul E. Sax,

Brigham and Women's Hospital, and Harvard Medical School, Boston, USA

Eric S. Daar,

Los Angeles Biomedical Research Institute at Harbor–University of California at Los Angeles Medical Center, USA

Adaora A. Adimora and Joseph J. Eron

University of North Carolina at Chapel Hill, USA

and Michael J. Mugavero

University of Alabama, Tuscaloosa, USA

[Received September 2016. Final revision December 2017]

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 with a previously proposed stratified estimator. The methods are then utilized to generalize results from two randomized trials of human immunodeficiency virus treatment to all people living with the disease in the USA.

Keywords: Causal inference; External validity–generalizability; Human immunodeficiency virus–acquired immune deficiency syndrome; 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;

Address for correspondence: Ashley L. Buchanan, Pharmacy Practice, University of Rhode Island, 7 Greenhouse Road, Kingston, RI 02881, USA.
E-mail: buchanan@uri.edu

 0964–1998/18/182000

Tipton, 2013; Keiding and Louis, 2016) and economics (Hotz *et al*., 2005; Heckman *et al*., 2006; Allcott, 2011, 2015; Allcott and Mullainathan, 2012; Hartman *et al*., 2015; Muller, 2014; Gechter, 2015). 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 individuals who are infected with the human immunodeficiency virus (HIV), there is often concern that trial participants are not representative of the larger population of HIV positive individuals. Greenblatt (2011) highlighted the overrepresentation of African-American and Hispanic women among HIV cases in the USA and the limited clinical trial participation of members of these groups. The Women's Interagency HIV Study (WIHS) is a prospective, observational, multicentre study considered to be representative of women living with HIV and women at risk for HIV infection in the USA (Bacon *et al*., 2005). However, a review of eligibility criteria of 20 AIDS Clinical Trials Group (ACTG) studies found that 28–68% of the HIV positive women in the WIHS cohort would have been excluded from these trials (Gandhi *et al*., 2005).

There are 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. Generalizability methods employing sampling scores are akin to methods that use treatment propensity scores to adjust for (measured) confounding (Rosenbaum and Rubin, 1983) 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 by using logistic regression and then employed inverse probability of sampling weighted (IPSW) methods to estimate the treatment effect in the target population. The IPSW approach is similar to inverse probability weighting methods that are used in a wide variety of contexts (for example, see Wooldridge (2002), Ding and Lehrer (2010) and Seaman and White (2013)). 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, 2014; 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 generalizability 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 with the stratified estimator. In Section 2, assumptions and notation are discussed. The IPSW estimator and the stratified estimator are described in Section 3.1. 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 USA. Section 6 concludes with a discussion.

2. Assumptions and notation

Suppose that we are interested in drawing inference about the effect of a treatment (e.g. a drug) on an outcome (e.g. a disease) in some target population. Assume that 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. Let $\mu_1 = E(Y^1)$ and $\mu_0 = E(Y^0)$ denote the mean potential outcomes in the target population. The parameter of interest is the population average treatment effect (PATE) $\Delta = \mu_1 - \mu_0$. The goal is to draw inference about Δ in a setting where two data sets are available. Assume that a random sample (e.g. a 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.

Throughout, it is supposed that the stable unit treatment value assumption (Rubin, 1980) holds, i.e. there are no variations of treatment and there is no interference between individuals. Under this assumption, each individual has only two potential outcomes: Y^0 and Y^1. Plausibility of the assumption that there are no variations of treatment will depend on the extent to which the form of treatment (i.e. the delivery mechanism, dose, non-compliance rate and so forth) differs between individuals, in particular between trial and cohort study participants. For example, in a randomized trial, treatment administration may be accompanied by adherence counselling, unlike in a cohort study. Note that the no variations of treatment assumption applies both to the treatment as well as to the control condition, and this would be suspect if there were a placebo effect in the randomized trial but not in the cohort study. The no-interference assumption supposes that the treatment of one individual does not affect the outcome of any other individuals. This assumption will be plausible in many settings but may be questionable in some studies, e.g., in an influenza vaccine trial, whether one individual is vaccinated may affect whether another individual develops flu.

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

Assume that 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 non-participants regarding the treatment–outcome relationship conditional on Z. The set of covariates Z should be chosen such that the ignorable trial participation mechanism is considered plausible. Judging whether a set of covariates Z is sufficient to satisfy this conditional independence assumption may be facilitated by explicitly representing the assumed data-generating mechanism using a directed acyclic graph (Greenland *et al.*, 1999). The ignorable trial participation mechanism assumption can then be verified by inspection of the directed acyclic graph (Pearl and Bareinboim, 2014).

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 $x = 0, 1$. That is, there is a positive probability of being included in the trial for each value of the covariates. Finally, it is assumed that the sampling score model, described in the next section, is correctly specified.

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

$$\hat{\Delta}_{\mathrm{T}} = \frac{\sum\limits_i S_i Y_i X_i}{\sum\limits_i S_i X_i} - \frac{\sum\limits_i S_i Y_i (1 - X_i)}{\sum\limits_i S_i (1 - X_i)},$$

where here and in what follows $\Sigma_i = \Sigma_{i=1}^{n+m}$. If trial participants are assumed to constitute a random sample from the target population, it is straightforward to show that $\hat{\Delta}_{\mathrm{T}}$ is a consistent and asymptotically normal estimator of Δ. In contrast, if we are not willing to assume that trial participants are a random sample from the target population, then $\hat{\Delta}_{\mathrm{T}}$ is no longer guaranteed to be consistent.

Below we consider two estimators of Δ that do not assume that 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 that we assume that the $1 \times p$ vector Z includes 1 as the first component 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

$$\hat{\Delta}_{\mathrm{IPSW}} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{\sum\limits_i S_i Y_i X_i / \hat{w}_i}{\sum\limits_i S_i X_i / \hat{w}_i} - \frac{\sum\limits_i S_i Y_i (1 - X_i) / \hat{w}_i}{\sum\limits_i S_i (1 - X_i) / \hat{w}_i}. \tag{1}$$

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 by 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

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

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

3.2. Large sample properties of the inverse probability of sampling weighted estimator

Because the trial participants are not assumed to be a random sample from the target population, the observed random variables $(S_i, Z_i, S_iX_i, S_iY_i)$ 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 to 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 matrices

$$A_{m,n}(\theta^*) = (n+m)^{-1} \sum_i E\left\{ \frac{\partial}{\partial \theta^*} \Psi^*_\Delta(Y_i, Z_i, X_i, S_i, \theta^*) \right\},$$
$$B_{m,n}(\theta^*) = (n+m)^{-1} \sum_i \mathrm{cov}\{\Psi^*_\Delta(Y_i, Z_i, X_i, S_i, \theta^*)\}.$$

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 that $E\{\Psi^*_\Delta(Y_i, Z_i, X_i, S_i, \theta^*)\} = 0$ for $i = 1, \ldots, n+m$, implying under suitable regularity conditions 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

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

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

$$\Sigma^*_{\mathrm{IPSW}} = \Sigma^{*(11)}_\theta + \Sigma^{*(22)}_\theta - 2\Sigma^{*(12)}_\theta \tag{3}$$

and in general $\Sigma^{(ij)}$ refers to the entry in the ith row and the jth column of the matrix Σ. A consistent estimator of equation (3) is given in the on-line 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

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

(Scott and Wild, 1986). Let $\hat{\theta} = (\hat{\mu}_1, \hat{\mu}_0, \hat{\beta})$ and $\theta = (\mu_1, \mu_0, \beta)$ 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 matrices

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

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

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 that $E\{\Psi_\Delta(Y_i, Z_i, X_i, S_i, \theta)\} = 0$ for $i = 1, \ldots, n+m$, 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

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

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

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

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

A comparison of equations (3) and (5) shows that the variance is smaller when the sampling scores are estimated (see the on-line appendix B). Therefore, even if the correct sampling scores are known, estimation of the sampling scores is preferable because of improved efficiency. This is analogous to a well-known result for inverse-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 by using standard software assuming that the weights are known. This leads to valid but conservative CIs. In the on-line supplementary materials, 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 that 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 $\hat{\Delta}_{\mathrm{S}}$ 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 that $\hat{\Delta}_{\mathrm{S}}$ is asymptotically normal (Lunceford and Davidian, 2004). In practice, it is routine to approximate the sampling variance of $\hat{\Delta}_{\mathrm{S}}$ 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 $\hat{\Delta}_{\mathrm{S}}$ is

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

where $\hat{\sigma}_l^2 = \Sigma_{x=0}^{1} n_{xl}^{-1} s_{xl}^2$, $n_{xl} = \Sigma_{i=1}^{n_l+m_l} S_{il} I(X_{il} = x)$, $s_{xl} = n_{xl}^{-1} \Sigma_{i=1}^{n_l+m_l} S_{il} I(X_{il} = x)(Y_{il} - \bar{Y}_{xl})^2$ and $\bar{Y}_{xl} = n_{xl}^{-1} \Sigma_{i=1}^{n_l+m_l} S_{il} I(X_{il} = x) Y_{il}$ 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 5000 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 3–6, 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 such that the sample size in the trial was approximately $n \approx 1000$. The cohort was a random sample of size $m = 4000$ from the target population (less those selected in the trial). The number of participants in the randomized trial was small compared with the size of the target, so the cohort was essentially a random sample from the target.

For those who were 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 1–4, $(\nu_0, \nu_1, \xi, \alpha) = (0, 1, 2, 1)$. For scenarios 5 and 6, $(\nu_0, \nu_1, \xi, \alpha) = (0, 1, 2, 2)$. Two sampling score models were considered: $\beta = (-7, 0.4)$ for scenarios 1, 3 and 5; $\beta = (-7, 0.6)$ for scenarios 2, 4 and 6. The truth was calculated for each scenario on the basis of 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 3–6. To estimate the sampling scores, the combined trial ($S_i = 1$) and cohort ($S_i = 0$) data were 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 $\hat{\Delta}_{\mathrm{T}}$ was biased for all scenarios and had low coverage (the results are not shown). For all scenarios, $\hat{\Delta}_{\mathrm{IPSW}}$ was unbiased. For scenarios 1 and 2, $\hat{\Delta}_{\mathrm{S}}$ was unbiased and standard errors were comparable with $\hat{\Delta}_{\mathrm{IPSW}}$. For scenarios 3–6, $\hat{\Delta}_{\mathrm{S}}$ was biased, possibly because of 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 $\hat{\Delta}_{\mathrm{IPSW}}$. With a continuous covariate, the Wald CI of the

Table 1. Simulation results for estimators of Δ when the sampling score model was correctly specified†

Scenario	*Covariate*	*(β_1, α)*	*Bias*			*Empirical standard error (× 100)*		*Average estimated standard error (× 100)*		*Empirical coverage probability*	
			$\hat{\Delta}_{\mathrm{T}}$	$\hat{\Delta}_{\mathrm{S}}$	$\hat{\Delta}_{\mathrm{IPSW}}$	$\hat{\Delta}_{\mathrm{S}}$	$\hat{\Delta}_{\mathrm{IPSW}}$	$\hat{\Delta}_{\mathrm{S}}$	$\hat{\Delta}_{\mathrm{IPSW}}$	$\hat{\Delta}_{\mathrm{S}}$	$\hat{\Delta}_{\mathrm{IPSW}}$
1	Binary	(0.4,1)	0.07	0.00	0.00	6.2	7.1	7.1	7.3	0.98	0.95
2	Binary	(0.6,1)	0.11	0.00	0.00	6.3	7.1	6.6	7.1	0.96	0.95
3	Continuous	(0.4,1)	0.20	0.04	0.00	8.1	13.4	7.9	13.4	0.91	0.95
4	Continuous	(0.6,1)	0.60	0.07	0.00	8.6	15.0	8.6	14.9	0.88	0.95
5	Continuous	(0.4,2)	0.80	0.09	0.00	9.4	17.2	8.9	17.2	0.81	0.95
6	Continuous	(0.6,2)	1.20	0.14	0.00	10.1	19.9	9.8	19.6	0.70	0.95

†For 5000 simulated data sets with $m = 4000$ and $n \approx 1000$ per data set. The scenarios are described in Section 4. For scenarios 1 and 2 $\Delta = 2.2$ and for scenarios 3–6 $\Delta = 2.0$. (T, within trial; S, stratified.)

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 Fig. 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 3–6, $Z_{2i} \sim N(0, 1)$. 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 1–4, $(\nu_0, \nu_1, \nu_2, \xi, \alpha_1, \alpha_2) = (0, 1, 1, 2, 1, 1)$. For scenarios 5 and 6, $(\nu_0, \nu_1, \nu_2, \xi, \alpha_1, \alpha_2) = (0, 1, 1, 2, 2, 2)$. The estimated sampling scores were computed on the basis of a misspecified logistic regression with Z_{1i} as the only covariate, i.e. $\hat{w}_i = \{1 + \exp(-\hat{\beta}_0 - \hat{\beta}_1 Z_{1i})\}^{-1}$. Two sampling score models were considered: scenarios 1, 3 and 5 set $\beta = (-7, 0.4, 0.4)$; scenarios 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 3–6.

Comparisons between the IPSW and stratified estimators when the sampling score model was misspecified are summarized in the on-line appendix C Table 1. The bias was reduced by approximately half when either the IPSW or the stratified estimator was employed as compared with 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.

Lastly, simulations were also performed with reduced overlap in the distribution of Z in the trial and target population. Specifically, the simulation study that was described above with correct specification of the sampling score model was repeated, except that $\beta_1 = 1$ in scenarios 1, 3 and 5, and $\beta_1 = 2$ in scenarios 2, 4 and 6. Thus, there was a stronger association between the covariate Z_1 and trial participation than in the original set of simulations, leading to greater differences in the covariate distributions between trial participants and the cohort. For example, in scenario 1, $P(Z_1 = 1|S = 1) = 0.40$ and $P(Z_1 = 1|S = 0) = 0.20$ when $\beta_1 = 1$, compared with $P(Z_1 = 1|S = 1) = 0.26$ and $P(Z_1 = 1|S = 0) = 0.20$ when $\beta_1 = 0.4$. Results from this last simulation study are summarized in the on-line appendix C Table 2. The IPSW estimator was unbiased for all scenarios, with the corresponding CI coverage approximating the nominal level except in scenarios 4 and 6. The stratified estimator was biased (although less than the within-trial estimator) and the corresponding CIs did not cover at the nominal level except in scenario 1. Because of the reduced overlap in the covariate distributions between the trial and cohort, both the IPSW and the stratified estimators were more variable relative to the simulation results in Table 1.

5. Applications

5.1. Trials and cohorts

In this section, the methods that were described in Section 3.1 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 USA and all people currently living with HIV in the USA.

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 1156 participants were enrolled in the ACTG 320 trial between January 1996 and January 1997 and were recruited from 33 acquired immune deficiency syndrome (AIDS) clinical trial units and seven National Hemophilia Foundation sites in the USA and Puerto Rico. These participants were HIV positive

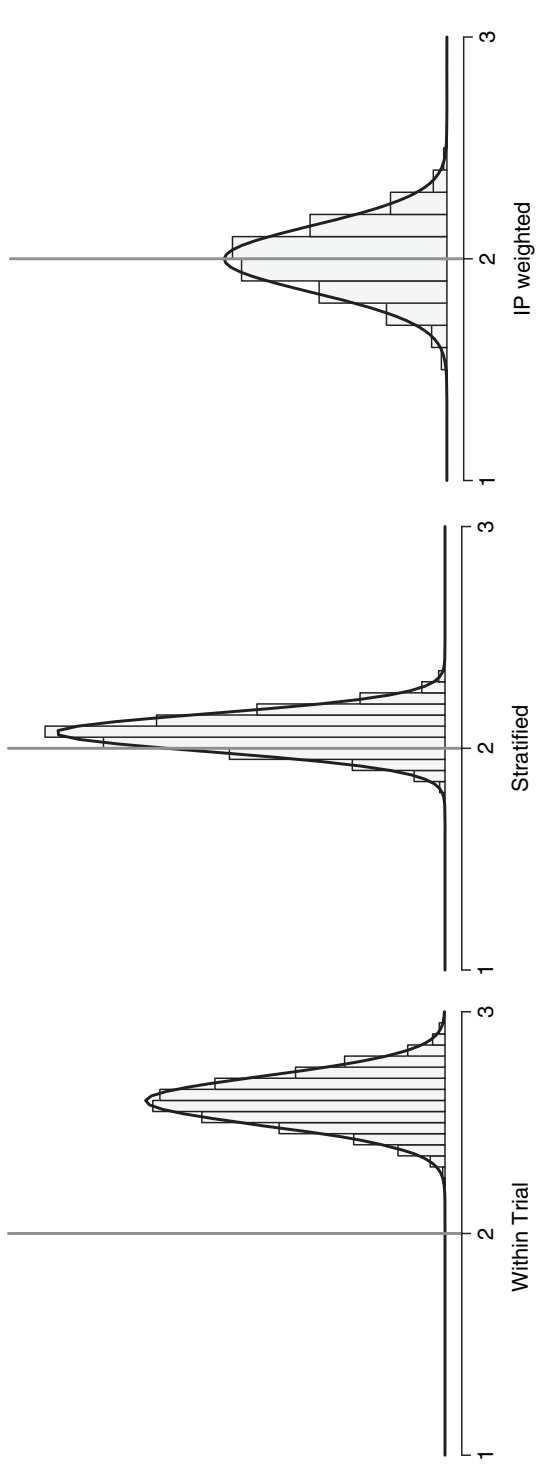

Fig. 1. Comparison of the distributions of (a) the within-trial estimator $\hat{\Delta}_T$, (b) stratified estimator $\hat{\Delta}_S$ and (c) IPSW estimator $\hat{\Delta}_{IPSW}$, based on 5000 simulated data sets where the sampling score model is correctly specified and $\Delta = 2$ with one continuous covariate, $\beta = (-7, 0.6)$ and $\alpha = 1$ (scenario 4)

and highly active anti-retroviral therapy (HAART) naive, and had CD4 T-cell counts of 200 cells mm^{-3} or fewer at screening. Of the 1156 participants, 200 were women (Hammer *et al.*, 1997). Among ACTG 320 trial participants, 116 (10%) were missing the outcome of CD4 cell count at week 4, so they were excluded from the analysis below. The baseline characteristics of the ACTG 320 trial participants are shown in Table 2.

The ACTG A5202 trial assessed equivalence of abacavir–lamivudine (known as 'ABC–3TC') or tenofovir disoproxil fumarate–emtricitabine (known as 'TDF–FTC') plus efavirenz or ritonavir-boosted atazanavir. A total of 1857 participants were enrolled in trial A5202 between September 2005 and November 2007 and were recruited from 59 ACTG sites in the USA and Puerto Rico. These participants were HIV positive and anti-retroviral (ART) naive and had a viral load greater than 1000 copies ml^{-1} at screening. Of the 1857 participants, 322 were women (Sax *et al.*, 2009, 2011). Among A5202 trial participants, 417 (22%) were missing the outcome of CD4 cell count at week 48, so they were excluded from the analysis below. The baseline characteristics of the A5202 trial participants are shown in Table 3.

Data from two cohort studies, the WIHS and Center for AIDS Research (CFAR) Network of Integrated Clinical Systems (known as the 'CNICS'), are used in the analysis below to generalize the ACTG 320 and A5202 trial results. Participants in the WIHS and CNICS study were considered to be representative samples of the target populations, i.e. all women living with HIV in the USA and all people living with HIV in the USA respectively. A total of 4129 women (1065 HIV uninfected) were enrolled in the WIHS between October 1994 and December 2012

Table 2. Characteristics of WIHS participants, CNICS study participants and ACTG 320 trial participants at baseline†

Variable	*Results for WIHS (m = 493)*	*Results for ACTG 320 women (n = 200)*	*Results for CNICS study (m = 6158)*	*Results for ACTG 320 men and women (n = 1156)*
Male sex (%)	0	0	80	83
Race or ethnic group (%)				
White, non-Hispanic	18	31	40	52
Black, non-Hispanic	55	48	44	28
Hispanic	25	21	12	18
Asian or other	2	1	5	2
Median age (years) (quartile 1–quartile 3)‡	40 (35–45)	36 (30–42)	41 (34–47)	38 (33–44)
Age group (%)				
[16, 30) years	7	23	12	12
[30, 40) years	43	44	34	47
[40, 50) years	40	27	37	30
[50, ·) years	10	7	17	11
Injection drug use (%)	37	18	20	16
Median CD4 cell count‡ (quartile 1–quartile 3)	108 (41–172)	82 (26–139)	89 (27–172)	75 (23–137)
Baseline CD4 cell count (%)				
(0, 50) cells mm^{-3}	30	36	36	39
[50, 100) cells mm^{-3}	17	22	17	22
[100, 200) cells mm^{-3}	37	37	30	32
[200, ·) cells mm^{-3}	16	6	17	7

†*m* is the number of participants in the cohort study. *n* is the number of participants in the trial.
‡One ACTG 320 trial participant was missing a CD4 cell count.

Table 3. Characteristics of WIHS participants, CNICS study participants and ACTG A5202 trial participants at baseline†

Variable	*Results for WIHS ($m = 1012$)*	*Results for ACTG A5202 women ($n = 322$)*	*Results for CNICS study ($m = 12302$)*	*Results for ACTG A5202 men and women ($n = 1857$)*
Male sex (%)	0	0	82	83
Race or ethnic group‡ (%)				
White, non-Hispanic	17	18	45	40
Black, non-Hispanic	58	53	38	33
Hispanic	22	26	12	23
Asian or other	3	3	5	3
Median age (years) (quartile 1–quartile 3)	39 (33–44)	39 (31–46)	39 (31–46)	38 (31–45)
Age group (%)				
[16, 30) years	12	18	20	22
[30, 40) years	43	34	34	34
[40, 50) years	34	33	32	31
[50, ·) years	11	15	14	14
Injection drug use (%)	38	6	17	9
Hepatitis B or C (%)	35	8	18	9
AIDS diagnosis (%)	37	19	23	17
CD4 cell count§ (%)				
(0, 50) cells mm^{-3}	10	19	16	18
[50, 100) cells mm^{-3}	6	7	7	8
[100, 200) cells mm^{-3}	16	17	14	17
[200, 350) cells mm^{-3}	29	40	27	35
[350, ·) cells mm^{-3}	39	16	36	22
Median CD4 cell count (quartile 1–quartile 3)	290 (162–423)	226 (87–313)	271 (109–427)	230 (90–334)
Viral load (%)				
[0, 50000) copies ml^{-1}	55	58	52	54
[50000, 100000) copies ml^{-1}	14	19	15	21
[100000, 300000) copies ml^{-1}	19	12	18	11
[300000, 500000) copies ml^{-1}	5	3	6	4
[500000, ·) copies ml^{-1}	7	8	8	10
Median $\log_{10}$ viral load (quartile 1–quartile 3)	4.61 (4.04–5.11)	4.58 (4.07–4.93)	4.64 (3.95–5.18)	4.66 (4.33–5.01)

†m is the number of participants in the cohort study. n is the number of participants in the trial.
‡Five A5202 trial participants were missing race.
§One A5202 trial participant was missing a CD4 cell count.

at six US sites (Bacon *et al.*, 2005). The CNICS study 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 27000 HIV-infected adults (at least 18 years of age) engaged in clinical care since January 1995 at eight CFAR sites in the USA.

For generalizing results from the ACTG 320 trial, the analysis included cohort participants who were HIV positive and HAART naive, and had CD4 cell counts of 200 cells mm^{-3} or fewer at the previous visit ($m = 493$ women and $m = 6158$ men and women combined). For generalizing results from the A5202 trial, the analysis included cohort participants who were HIV positive and ART naive, and had a viral load of 1000 copies ml^{-1} at the previous visit ($m = 1012$ women and $m = 12302$ men and women combined). Table 2 displays the characteristics of

the women in the WIHS sample and the participants in the CNICS sample that were used to generalize results from the ACTG 320 trial. Likewise, the characteristics of the women in the WIHS sample and participants in the CNICS sample that were used to generalize results from the ACTG A5202 trial are displayed in Table 3.

5.2. Analysis

The IPSW and stratified estimators were employed to generalize the difference in the average change in CD4 cell count from baseline between treatment groups observed among women in the trials to all women currently living with HIV in the USA and among all participants in the trials to all people currently living with HIV in the USA. On the basis of 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 PATE was estimated by 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 analysed 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 the ACTG 320 trial included sex, race or ethnicity, age, history of injection drug use (IDU) and baseline CD4 cell count, and for the ACTG A5202 trial included sex, race or ethnicity, age, history of IDU, hepatitis B or C, AIDS diagnosis, baseline CD4 cell count and baseline $\log_{10}$ viral load. The variable hepatitis B or C was binary, indicating infection with hepatitis B or hepatitis C or both. 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 4. Among all participants and among just women in the ACTG 320 trial, there was a significant difference in the change in CD4 cell count from baseline to 4 weeks between the PI and non-PI groups. Among women in the A5202 trial at week 48, those randomized to ABC–3TC had an average change in CD4 cell count that was comparable with those randomized to a regimen with TDF–FTC. Among all participants in the A5202 trial, those randomized to ABC–3TC had an average change in CD4 cell count that was slightly higher than those randomized to a regimen with TDF–FTC, but this did not achieve statistical significance.

Table 4 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 USA, the IPSW estimate was approximately double the within-trial estimate ($\hat{\Delta}_{\mathrm{IPSW}} = 46$ compared with $\hat{\Delta}_{\mathrm{T}} = 24$), suggesting that the within-trial result may underestimate the effects of PIs in all HIV-infected women in the USA. The IPSW estimator also indicated a much stronger protective effect of ABC–3TC (*versus* TDF–FTC) in the target population of all HIV-infected women in the USA ($\hat{\Delta}_{\mathrm{IPSW}} = 35$ compared with $\hat{\Delta}_{\mathrm{T}} = 1$), 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 USA, the IPSW estimates were comparable with the within-trial effect estimates, suggesting that both the effect of PIs and the effect of the ART combination ABC–3TC (*versus* TDF–FTC) from the trials may be generalizable to all people living with HIV

Table 4. Estimated difference in mean change in CD4 cell count and corresponding 95% CIs in two target populations (all men and women combined and all women living with HIV in the USA) based on data from ACTG trials, the WIHS and CNICS†

Cohort	*m*	*Trial*	*n*	*Difference in mean change (95% CI)*		
				$\hat{\Delta}_T$	$\hat{\Delta}_S$	$\hat{\Delta}_{IPSW}$
WIHS	493	ACTG 320‡	200	24 (7, 41)	38 (17, 59)	46 (23, 70)
WIHS	1012	ACTG A5202§	322	1 (−35, 37)	−19 (−62, 25)	35 (−45, 115)
CNICS	6158	ACTG 320	1156	19 (12, 25)	18 (9, 26)	17 (9, 25)
CNICS	12302	ACTG A5202	1857	6 (−8, 20)	7 (−18, 32)	−2 (−31, 28)

†For the ACTG 320 trial, the outcome was change in CD4 cell count from baseline to week 4. For the A5202 trial, the outcome was change in CD4 cell count from baseline to week 48.
‡For the ACTG 320 trial, the treatment contrast was PI ($X = 1$) *versus* no PI ($X = 0$).
§For the A5202 trial, the treatment contrast was ABC–3TC ($X = 1$) *versus* TDF–FTC ($X = 0$) plus efavirenz or ritonavir-boosted atazanavir.

in the USA. In summary, these results suggest that 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 by 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 estimator outperformed the stratified estimator when the sampling score was correctly specified. The IPSW estimator was unbiased for all scenarios and the CIs exhibited coverage that was approximately at the nominal level, except when there was limited overlap in the distribution of covariates in the trial compared with the target population. 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 and A5202 trial results appear to be generalizable to all people living with HIV in the USA. In contrast, the within-trial effect estimates among women in the two ACTG trials were not comparable with the effect estimates in the target population of women. This lack of comparability may be explained by differences in the distribution of certain effect modifiers between the trial and the target population. Figs 1–4 in the on-line appendix D show within-trial subgroup effect estimates and CIs for both trials. Among women in the A5202 trial, the results in Fig. 3 of appendix D suggest that hepatitis B or C, IDU and age were possible effect modifiers. These three covariates were also associated with trial participation among women, and thus may explain why the A5202 within-trial effect estimate among women was not similar to the IPSW effect estimate in the target population of women. In particular, women with hepatitis B or C were less likely to participate in the A5202 trial and tended to have a greater mean change in CD4 cell count. Thus, by accounting for hepatitis B or C, we would expect the IPSW estimate to be greater than the within-trial estimate. Likewise, women who were younger or had a history of IDU were also less likely to participate in the A5202 trial and tended to have a greater mean change in CD4 cell count than older women or those without a history of IDU respectively. Results from both the ACTG A5202 and the ACTG 320 trial 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 (the results are not shown). In the data example, a complete-case analysis was performed; however, in practice, one would want to address the possibility that the missingness was not completely at random.

When applying these methods, the analysis is subject to the following considerations. First, the ignorable trial participation mechanism is a key assumption which supposes that participants in the ACTG trials are no different from individuals in the WIHS and the CNICS study with respect to the treatment–outcome relationship conditional on observed covariates. However, it is plausible that there are unmeasured covariates that are associated with trial participation and the outcome which confound the association between treatment and outcome even after conditioning on the observed covariates. The methods considered in this paper also assume that there are no variations in treatment. Within the context of the HIV treatment trial analysis, this assumption supposes that ART adherence rates were similar between those in the target population and participants in the ACTG trials. This assumption could be assessed if data related to adherence had been collected in the WIHS, the CNICS study and the ACTG trials. The no variations of treatment assumption additionally supposes that there are no other behavioural responses or contextual effects (Ding and Lehrer, 2015) of study participation that would not remain if the treatment were adopted in the target population.

In addition, the sampling score model was assumed to be correctly specified (e.g. correct covariate functional forms). Because some degree of model misspecification is inevitable, a sensitivity analysis of inferences about the treatment effect in the population to the sampling score model specification is recommended. Similarly, 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).

The inferential methods considered in this paper assume that the cohort is a random sample (i.e. representative) of the target population. In the HIV application, participants in the WIHS (or CNICS study) are assumed to constitute a random sample of women (men and women) living with HIV in the USA. This assumption would be violated if cohort participation is associated with individual characteristics, such as age, living in an urban area, income and employment status. In the context of the HIV application, this assumption might be considered more plausible if the target population were instead defined with greater specificity, e.g. as all women (and men) living with HIV who are in care at the geographical locations which have sites in the cohort studies. If the cohort is not considered representative of the target population, one possibility is weighting the cohort data to the distribution of covariates in a census (e.g. Centers for Disease Control and Prevention estimates). A limitation of this approach is that the census may not have covariate information as rich as the cohort data. The Centers for Disease Control and Prevention estimates that were 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 USA could further sharpen the information about the target population.

In this paper, we consider randomized trials where individuals are independently randomized to treatment or control. Future research could entail extensions to cluster randomized trials wherein clusters of individuals are independently randomized to treatment or control, with all individuals in the same cluster receiving the same randomization assignment. Causal inference methods for individually randomized studies are not necessarily valid for cluster randomized

trials (Middleton and Aronow, 2015), such that the methods that are considered in this paper may not be directly applicable to cluster randomized trials.

Future research could also entail the use of machine learning methods (Westreich *et al*., 2010), maximum entropy (Hartman *et al*., 2015) or flexible regression methods such as Bayesian adaptive tree regression (Chipman *et al*., 2010) instead of weighted logistic regression to estimate the sampling scores. The IPSW estimator considered in this paper may be highly variable when there is limited overlap in the distribution of the covariates in the trial compared with the target population. Thus, alternative estimators should be developed for settings where there is limited covariate distribution overlap. Formal sensitivity analysis methods could be developed to assess the extent to which violations of key assumptions, such as the ignorable trial participation mechanism assumption, potentially affect inference about the treatment effect in the target population. Alternatively, bounds could be derived (as in Gechter (2015)) under weaker assumptions which only partially identify the PATE. For the methods that were considered in this paper, no information on the exposure or outcome is required from the cohort study. In settings where the outcome data are available in the cohort, approaches similar to Hotz *et al*. (2005) and Hartman *et al*. (2015) could be developed to test the ignorable trial participation mechanism assumption.

Additional extensions might be considered based on the types of data that are typical of biomedical, public health and econometric studies. For example, time-to-event end points are common in HIV and AIDS trials, so extensions to accommodate right-censored outcomes could be considered. Lastly, in some settings such as infectious disease studies, the treatment or exposure of one individual may affect the outcome of another individual; extensions of existing generalizability methods, such as using inverse probability of sampling weights, to allow for interference would have utility in such settings.

Acknowledgements

These findings are presented on behalf of the WIHS, the CNICS and the ACTG. We 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 grants R01AI100654, R01AI085073, U01AI042590, U01AI069918, R56AI102622, 5 K24HD059358-04, 5 U01AI103390-02 (WIHS), R24AI067039 (CNICS) and P30AI50410 (University of North Carolina CFAR). The views and opinions of the authors that are expressed in this paper do not necessarily state or reflect those of the National Institutes of Health. We thank the reviewers for helpful comments that greatly improved this work.

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, University of North Carolina WIHS (Adaora Adimora), U01-AI-103390, Connie Wofsy Women's 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, with additional co-funding from the Eunice Kennedy Shriver National Institute of Child Health and Human

Development, the National Cancer Institute, the National Institute on Drug Abuse and the National Institute on Mental Health. Targeted supplemental funding for specific projects is also provided by the National Institute of Dental and Craniofacial Research, the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Deafness and other Communication Disorders and the National Institutes of Health Office of Research on Women's Health. WIHS data collection is also supported by UL1-TR000004 (University of California, San Francisco, Clinical and Translational Science Institute) and UL1-TR000454 (Atlanta Clinical and Translational Science Institute).

The data and computer code used to generate the results in the manuscript will be provided on request but 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 that are specific to the studies analysed in this paper or otherwise at the National Institutes of Health. A request for data and computer code can be initiated by contacting the author for correspondence.

References

Allcott, H. (2011) Social norms and energy conservation. *J. Publ. Econ.*, **95**, 1082–1095.
Allcott, H. (2015) Site selection bias in program evaluation. *Q. J. Econ.*, **130**, 1117–1165.
Allcott, H. and Mullainathan, S. (2012) External validity and partner selection bias. *Working Paper 18373*. National Bureau of Economic Research, Cambridge.
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. *Clin. Diag. Lab. Immun.*, **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.
Centers for Disease Control and Prevention (2012) Diagnoses of HIV infection and AIDS in the United States and dependent areas. *HIV Surveillance Report 17*. Centers for Disease Control and Prevention, Atlanta.
Chipman, H. A., George, E. I. and McCulloch, R. E. (2010) BART: Bayesian additive regression trees. *Ann. Appl. Statist.*, **4**, 266–298.
Cole, S. R. and Stuart, E. A. (2010) Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am. J. Epidem.*, **172**, 107–115.
Ding, W. and Lehrer, S. F. (2010) Estimating treatment effects from contaminated multi-period education experiments: the dynamic impacts of class size reductions. *Rev. Econ. Statist.*, **92**, 31–42.
Ding, W. and Lehrer, S. F. (2015) Estimating context-independent treatment effects in education experiments. *Working Paper 3788*. Queen's University, Kingston.
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.
Gechter, M. (2015) Generalizing the results from social experiments: theory and evidence from Mexico and India. *Manuscript*. Department of Economics, Pennsylvania State University, State College.
Greenblatt, R. M. (2011) Priority issues concerning HIV infection among women. *Wmn's Hlth Iss.*, **21**, suppl., S266–S271.
Greenland, S., Pearl, J. and Robins, J. M. (1999) Causal diagrams for epidemiologic research. *Epidemiology*, **10**, 37–48.
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 Engl. J. Med.*, **337**, 725–733.
Hartman, E., Grieve, R., Ramsahai, R. and Sekhon, J. S. (2015) From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *J. R. Statist. Soc.* A, **178**, 757–778.
Heckman, J. J., Urzua, S. and Vytlacil, E. (2006) Understanding instrumental variables in models with essential heterogeneity. *Rev. Econ. Statist.*, **88**, 389–432.
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.
Hotz, V. J., Imbens, G. W. and Mortimer, J. H. (2005) Predicting the efficacy of future training programs using past experiences at other locations. *J. Econmetr.*, **125**, 241–270.

Keiding, N. and Louis, T. A. (2016) Perils and potentials of self-selected entry to epidemiological studies and surveys (with discussion). *J. R. Statist. Soc.* A, **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. *Int. J. Epidem.*, **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. *Statist. Med.*, **23**, 2937–2960.
Middleton, J. A. and Aronow, P. M. (2015) Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statist. Polit. Poly*, **6**, 39–75.
Muller, S. (2014) Randomised trials for policy: a review of the external validity of treatment effects. *Working Paper 127*. Southern Africa Labour and Development Research Unit, University of Cape Town, Cape Town.
O'Muircheartaigh, C. and Hedges, L. V. (2014) Generalizing from unrepresentative experiments: a stratified propensity score approach. *Appl. Statist.*, **63**, 195–210.
Pearl, J. and Bareinboim, E. (2014) External validity: from do-calculus to transportability across populations. *Statist. Sci.*, **29**, 579–595.
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.
Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Rubin, D. B. (1980) Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *J. Am. Statist. Ass.*, **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., Farajallah, A., Rooney, J. F., Ha, B., Woodward, W. C., Feinberg, J., Tashima, K., Murphy, R. L. and Fischl, M. A. on behalf of the AIDS Clinical Trials Group Study A5202 Team (2011) Abacavir/lamivudine versus tenofovir DF/emtricitabine as part of combination regimens for initial treatment of HIV: final results. *J. Infect. Dis.*, **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., Farajallah, A., Rooney, J. F., Ha, B., Woodward, W. C., Koletar, S. L., Johnson, V. A., Geiseler, P. J. and Daar, E. S. for the AIDS Clinical Trials Group Study A5202 Team (2009) Abacavir–lamivudine versus tenofovir–emtricitabine for initial HIV-1 therapy. *New Engl. J. Med.*, **361**, 2230–2240.
Scott, A. J. and Wild, C. J. (1986) Fitting logistic models under case-control or choice based sampling. *J. R. Statist. Soc.* B, **48**, 170–182.
Seaman, S. R. and White, I. R. (2013) Review of inverse probability weighting for dealing with missing data. *Statist. Meth. Med. Res.*, **22**, 278–295.
Stuart, E. A., Bradshaw, C. P. and Leaf, P. J. (2015) Assessing the generalizability of randomized trial results to target populations. *Prevn Sci.*, **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. *J. R. Statist. Soc.* A, **174**, 369–386.
Tipton, E. (2013) Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. *J. Educ. Behav. Statist.*, **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. *J. Res. Educ. Effect.*, **7**, 114–135.
Westreich, D. and Cole, S. R. (2010) Invited commentary: positivity in practice. *Am. J. Epidem.*, **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. *J. Clin. Epidem.*, **63**, 826–833.
Wooldridge, J. M. (2002) Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Port. Econ. J.*, **1**, 117–139.
Wooldridge, J. M. (2007) Inverse probability weighted estimation for general missing data problems. *J. Econmetr.*, **141**, 1281–1301.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for online publication for "Generalizing evidence from randomized trials using inverse probability of sampling weights"'.