American Journal of Epidemiology
Vol. 172, No. 1
DOI: 10.1093/aje/kwq084
Advance Access publication:
June 14, 2010

Practice of Epidemiology

Generalizing Evidence From Randomized Clinical Trials to Target Populations

The ACTG 320 Trial

Stephen R. Cole* and Elizabeth A. Stuart

* Correspondence to Dr. Stephen R. Cole, Department of Epidemiology, Gillings School of Global Public Health and Center for AIDS Research, CB7435, University of North Carolina, Chapel Hill, NC 27599 (e-mail: cole@unc.edu).

Initially submitted October 21, 2009; accepted for publication March 24, 2010.

Properly planned and conducted randomized clinical trials remain susceptible to a lack of external validity. The authors illustrate a model-based method to standardize observed trial results to a specified target population using a seminal human immunodeficiency virus (HIV) treatment trial, and they provide Monte Carlo simulation evidence supporting the method. The example trial enrolled 1,156 HIV-infected adult men and women in the United States in 1996, randomly assigned 577 to a highly active antiretroviral therapy and 579 to a largely ineffective combination therapy, and followed participants for 52 weeks. The target population was US people infected with HIV in 2006, as estimated by the Centers for Disease Control and Prevention. Results from the trial apply, albeit muted by 12%, to the target population, under the assumption that the authors have measured and correctly modeled the determinants of selection that reflect heterogeneity in the treatment effect. In simulations with a heterogeneous treatment effect, a conventional intent-to-treat estimate was biased with poor confidence limit coverage, but the proposed estimate was largely unbiased with appropriate confidence limit coverage. The proposed method standardizes observed trial results to a specified target population and thereby provides information regarding the generalizability of trial results.

bias; bias (epidemiology); causal inference; external validity; generalizability; randomized trials; standardization

Abbreviations: ACTG, AIDS Clinical Trial Group; AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus.

Properly planned and conducted randomized clinical trials (henceforth referred to as trials) typically provide stronger internal validity than observational study designs, such as prospective cohort studies. Such trials accomplish heightened internal validity by ensuring that the conditions necessary for proper inference are met. Specifically, trials ensure consistency (1–3) and positivity (1, 4) by design and no unmeasured confounding in expectation by randomization (5, 6). Trials and cohort studies constrain the amount of selection bias (7) due to dropout when near-complete patient follow-up is attained. However, even such trials are susceptible to a lack of external validity, or generalizability (8, 9), as recently discussed (10–12). This susceptibility is a function of the extent to which trial participants do not represent the target population. For an example of when trials might selectively enroll from the target population, a recent study (13) applied eligibility criteria from 32 human immunodeficiency virus (HIV) trials (largely funded by the National Institutes of Health) to the Women's Interagency HIV Study (14) (the largest observational cohort of HIV-infected women in the United States) and found that, across trials, a median of 58% of women would have been eligible for a given trial (range, 32.4%–100%).

In simple settings, trial results may be mapped to a target population by using nonparametric direct standardization (15, 16 (p. 49)). However, when there are many covariates, or some covariates are continuous, direct standardization will fail. Here, we illustrate a model-based method to standardize observed trial results to a specified target population. Thereby, this method provides information regarding

the generalizability of the trial results to the specified target population. We apply the method to the AIDS (acquired immunodeficiency syndrome) Clinical Trial Group (ACTG) 320 study (17), a landmark trial in HIV care that compared a novel highly active antiretroviral therapy combination (henceforth referred to as treatment) with a largely ineffective existing therapy combination (henceforth referred to as control). In addition, we provide a limited Monte Carlo simulation evaluation of the proposed method.

MATERIALS AND METHODS

Study population

Between January 1996 and January 1997, 1,156 patients were recruited from 33 AIDS clinical trial units and 7 National Hemophilia Foundation sites in the United States and Puerto Rico (17). Eligible patients were 1) at least 16 years of age, 2) HIV positive, 3) immunosuppressed (i.e., CD4 cell count <201 cells/mm^3), 4) experienced with antiretroviral therapy (i.e., at least 3 months of prior zidovudine use), and 5) able to care for themselves (i.e., Karnofsky performance test score ≥ 70). Patients were excluded if they had a week or more prior treatment with the nucleoside reverse transcriptase inhibitor lamivudine or had any prior treatment with protease inhibitors. Institutional review boards at each of the participating institutions approved the study protocol, and written informed consent was given by all study patients. The public-use ACTG 320 data set was used in the present study and is available from the National Technical Information Service (http://www.ntis.gov/).

At enrollment, patients were stratified by CD4 cell count (i.e., 0–50 vs. 51–200 cells/mm^3) and were randomly assigned with equal allocation to the treatment group ($n = 577$) or control group ($n = 579$) (17). The therapy for the control group consisted of the 2 nucleoside reverse transcriptase inhibitors zidovudine and lamivudine, whereas the therapy for the treatment group consisted of the same 2 nucleoside reverse transcriptase inhibitors plus the protease inhibitor indinavir. Characteristics of the 1,156 trial patients are given in Table 1.

After randomization, patients were monitored with study visits at weeks 4, 8, and 16, and every 8 weeks thereafter, until a first occurrence of an AIDS-defining illness, death, or the planned end of follow-up at 52 weeks. Fifty-one of 1,156 patients (4%) dropped out during follow-up. Of the 51 dropouts, 20 and 31 were in the treatment and control groups, respectively. Ninety-six of 1,156 patients (8%) incurred endpoints: 70 developed AIDS, and 26 died. Of the 96 endpoints, 33 were observed in the treatment group and 63 in the control group. Noncompliance is ignored here; it was previously described (17), and methods to account for noncompliance (18, 19) revealed only a modest difference from the hazard ratio obtained by intent-to-treat (20).

Target population

For illustrative purposes, we chose as the target population the US estimate of the number of people infected with HIV in 2006. This estimate was provided by the Centers for Disease Control and Prevention (21, 22). HIV incidence is not directly measured in the United States. However, innovative immunoassays are able to distinguish between recent and established infections, allowing estimates of HIV incidence (23–25). Information on newly diagnosed HIV cases in 22 states was reported to the Centers for Disease Control and Prevention for 2006. Remnant diagnostic serum specimens from patients aged 13 years or older were tested with an immunoassay to classify infections as recent or established. HIV incidence was estimated by using a statistical approach with adjustment for testing frequency and was extrapolated to the United States (26). Characteristics of the target population are also provided in Table 1. For this target population, we did not have individual-level data but did have the joint distribution (i.e., cross-classification) of select characteristics, namely, sex, race, and age groups.

Table 1. Characteristics of 1,156 HIV-infected Patients in the AIDS Clinical Trial Group 320 Study in 1996–1997 Followed for 1 Year and of the Estimated 54,220 HIV-infected Individuals in the United States in 2006

Characteristic[a]	Trial Patients		US Population	
	No.	%	No.	%
Age, years	38 (33, 44)		NA	
Age group, years[b]				
13–29	106	09	18,500	34
30–39	515	45	16,740	31
40–49	388	34	13,370	25
≥50	147	13	5,610	10
Male sex	956	83	39,810	73
Race				
White, non-Hispanic	623	54	19,580	36
Black, non-Hispanic	328	28	24,920	46
Hispanic	205	18	9,720	18
CD4 count (cells/mm^3)[c]	75 (33, 137)		NA	

Abbreviations: AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus; NA, not available.

[a] Values are expressed as median (quartiles) or percentage; percentages may not sum to 100 because of rounding.

[b] Youngest and oldest patients in the trial were aged 16 years and 75 years, respectively.

[c] CD4 cell count was missing for 1 trial patient.

Statistical methods

We begin with a description of the notation we will use. Uppercase letters denote random variables and lowercase letters denote realizations of random variables, or constants. Let T_i^* be $T_i \wedge C_i$, where T_i and C_i are positive, real valued times to the event of interest and right censoring, respectively, for population member $i = 1$ to n. We assume here that right censoring is not informative, or formally that $f(T^*) = f(T^* \mid T, C)$, where $f(.|.)$ is the conditional density function. Let $Y_i = 1$ denote the occurrence of the event of interest (i.e., $T_i^* = T_i$).

A population-level treatment effect is a comparison of potential event times across different levels of a treatment,

say $X = x$. Formally, this is a comparison of the distribution of T_i^1 and T_i^0, where T^x is the potential event time under treatment x. One way to quantify this comparison is to imagine a Cox proportional hazards model (27) on the potential event times as $h_{T^x}(t) = h_0(t) \times \exp(\alpha x)$, where the estimand $\exp(\alpha)$ is the ratio of the hazard had the population been exposed to the treatment to the hazard had the population been exposed to the control.

Let $S_i = 1$ denote selection from the target population into the trial sample of $\sum_{i=1}^{n} S_i$ patients. Where $S_i = 1$, let $X_i = 1$ denote random assignment to the treatment group and 0 to the control group. Typically, a treatment effect is estimated in the trial sample, perhaps by using an analogous Cox model, $h_T(t) = h_0(t) \times \exp(\beta X_i)$, where estimation is by Cox's partial likelihood (28). The log hazard ratio in the trial sample β will not generally equal the population estimand α, except under conditions described in the Appendix. A derivation of the bias in the trial sample estimate of the population effect, defined as differences in means or proportions, is also given in the Appendix. Next, we describe the use of inverse probability-of-selection weights, which are an extension of Horvitz-Thompson weights (29) and have been used extensively in survey sampling (30–32), for confounder control (33), and have been discussed in the context of selection bias (7, 19, 34, 35) or response bias in 2-phase studies (36).

Define the inverse probability-of-selection weight as

$$W_i = \begin{cases} \dfrac{P(S_i = 1)}{P(S_i = 1 \mid \mathbf{Z}_i)}, & S_i = 1, \\ 0, & S_i = 0, \end{cases}$$

where $P(.|.)$ is the conditional probability function. Let $\mathbf{Z}$ be an n-by-p matrix of discrete or continuous variables that describe the composition of the target population. For instance, in the simplest form, say the target population may be described by only a single binary characteristic such as sex, $\mathbf{Z} = Z_i = 0$ or 1. In our example, the target population is described by the complete cross-classification of sex, race, and age groups.

From the weight definition above, zero weights are given to target population members who are not selected into the trial sample, and real-valued positive weights are given to members who are selected. For selected members, the numerator of the weights, which is an estimate of the marginal probability of being selected, implies that $E(W_i \mid S_i = 1) = 1$, where $E(.|.)$ is the conditional expectation. The numerator is used to ensure that the weighted sample remains the same size as the observed sample in expectation. For selected members, the denominator of the weights is an estimate of the probability of being selected into the sample conditional on a vector of measured characteristics $\mathbf{Z}$. The weights W_i are therefore inversely proportional to an estimate of the conditional probability of being selected. On the basis of the findings given in the Appendix, the collection of characteristics $\mathbf{Z}$ is chosen, based on prior knowledge and data exploration, to include factors 1) on which the trial sample differs from the target population and 2) for which there is heterogeneity in the effect of treatment on the outcome of interest.

The conditional probabilities for both the numerator and denominator of the weights were obtained by using linear-logistic regression models, specifically,

$$\log\frac{P(S_i = 1)}{1 - P(S_i = 1)} = \delta \quad \text{and} \quad \log\frac{P(S_i = 1 \mid \mathbf{Z}_i)}{1 - P(S_i = 1 \mid \mathbf{Z}_i)} = \mathbf{Z}_i\phi,$$

where $1/[1 + \exp(-\delta)]$ is the marginal probability of being selected into the trial sample from the target population; $\mathbf{Z}_i$ includes a column of 1's for the intercept; and $\exp(\varphi_k)$, for $k = 1$ to p, are the log odds ratios for being selected for each component of the n-by-p covariate matrix $\mathbf{Z}$. In the models used for the ACTG 320 trial data, we included the characteristics themselves, as well as product terms to account for the joint distribution.

An inverse probability-of-selection-weighted Cox proportional hazards model may be fit by using the following weighted partial likelihood:

$$L(\gamma) = \prod_{i=1}^{n} \left[\frac{\exp(\gamma X_i) \times W_i}{\sum_{k=1}^{n} R_k(t_i) \times \exp(\gamma X_k) \times W_k} \right]^{Y_i},$$

where $R_k(t_i) = 1$ if patient k is at risk for the event at the event time for patient i, namely t_i. The resultant log hazard ratio, γ, provides a consistent, asymptotically normal estimate of the population treatment effect α under the assumption that the model for the denominator of the selection weight includes all characteristics that both 1) differ between trial sample and target population and 2) demonstrate heterogeneity in the treatment effect. A proof of the consistency of the proposed method for the special case of the difference in means or proportions is provided in the Appendix. Throughout, hazard ratios are used to measure the strength of association, 95% confidence limits are used to measure precision, robust variances (37–39) are used in conjunction with weighted Cox models (40), and confidence limit ratios are used to compare precision across estimates. The confidence limit ratio is simply the ratio of the upper to the lower confidence limit. The proportional hazards assumption appeared reasonable in these data under the original intent-to-treat analysis (P for heterogeneity = 0.263) and the proposed weighted analysis (P for heterogeneity = 0.211).

RESULTS

The intent-to-treat analysis of the ACTG 320 trial found a hazard of AIDS or death of 0.51 (95% confidence limits: 0.33, 0.77) for the 577 people randomly assigned to the treatment group relative to the 579 randomly assigned to the control group. In the trial, older age was associated with higher incidence of AIDS or death (P for trend = 0.0315); compared with the age group 13–29 years, the hazard ratios for the age groups 30–39, 40–49, and ≥50 were 1.33 (95% confidence limits: 0.56, 3.15), 1.43 (95% confidence limits: 0.59, 3.44), and 2.32 (95% confidence limits: 0.92, 5.82), respectively. However, neither male sex (hazard ratio = 0.98, 95% confidence limits: 0.55, 1.73) nor race was

Table 2. Odds Ratios and 95% Confidence Limits for Selection Into the AIDS Clinical Trial Group 320 Study in 1996–1997 From the Estimated US Population Infected With HIV in 2006

Characteristic[a]	Odds Ratio	95% CL
Age group, years[b]		
13–29	1	
30–39	4.93	4.02, 6.12
40–49	4.75	3.82, 5.89
≥50	4.29	3.34, 5.52
Male sex	1.51	1.29, 1.77
Race		
White, non-Hispanic	1	
Black, non-Hispanic	0.51	0.45, 0.59
Hispanic	1.53	1.28, 1.83

Abbreviations: AIDS, acquired immunodeficiency syndrome; CL, confidence limit; HIV, human immunodeficiency virus.
[a] Odds ratios were adjusted for the variables listed in the table.
[b] Youngest and oldest patients in the trial were aged 16 years and 75 years, respectively.

Table 3. Hazard Ratios and 95% Confidence Limits for Incident AIDS or Death Within 1 Year for Patients in the AIDS Clinical Trial Group 320 Study in 1996–1997 and for the Population of Individuals Infected With HIV in 2006, United States

	Hazard Ratio	95% CL	CL Ratio
Trial results			
Intent-to-treat[a]	0.51	0.33, 0.77	2.33
Age-group stratified, years[b,c]			
13–29	1.87	0.34, 10.2	
30–39	0.21	0.09, 0.48	
40–49	0.84	0.41, 1.70	
≥50	0.59	0.24, 1.45	
Sex stratified[d]			
Male	0.47	0.29, 0.74	
Female	0.76	0.28, 2.10	
Race stratified[e]			
White, non-Hispanic	0.59	0.34, 1.01	
Black, non-Hispanic	0.30	0.11, 0.83	
Hispanic	0.54	0.22, 1.36	
Population results			
Age weighted	0.68	0.39, 1.17	3.00
Sex weighted	0.53	0.34, 0.82	2.41
Race weighted	0.46	0.29, 0.72	2.48
Age-sex-race weighted	0.57	0.33, 1.00	3.03

Abbreviations: AIDS, acquired immunodeficiency syndrome; CL, confidence limit; HIV, human immunodeficiency virus.
[a] The age-group-adjusted hazard ratio was 0.50 (95% CL: 0.33, 0.77).
[b] P for homogeneity = 0.0348.
[c] Youngest and oldest patients in the trial were aged 16 years and 75 years, respectively.
[d] P for homogeneity = 0.3930.
[e] P for homogeneity = 0.5396.

strongly associated with incident AIDS or death (compared with the hazard ratio for whites, the hazard ratio for black non-Hispanics was 0.77 (95% confidence limits: 0.46, 1.28) and for Hispanics was 1.19 (95% confidence limits: 0.71, 2.00)).

Table 2 presents adjusted odds ratios for selection into the trial from the estimated US population infected in 2006. For presentation in Table 2, we omitted product terms between components of **Z**, but such terms are included for construction of W, as noted above. Males, those of white race or Hispanic ethnicity (compared with black race), or those older than age 30 years were more likely to be selected into the trial.

Table 3 presents the hazard ratios and 95% confidence limits applicable to the trial patients as well as for the target population. As expected, based on the results in Table 2 and the age-stratified trial results in Table 3, when accounting only for the difference in age between the trial sample and target population, the hazard ratio was markedly muted from 0.51 to 0.68 because the trial selected for older population members, for whom the treatment effect appeared stronger. Similar results can be obtained by use of direct standardization when the dimension of **Z** is low. For instance, if the age-stratified hazard ratios (i.e., 1.87, 0.21, 0.84, and 0.59; Table 3) are log-transformed and combined by using the target population frequency distribution (i.e., 0.34, 0.31, 0.25, and 0.10; Table 1), the antilog of the direct standardized estimate is 0.69, which is similar to our model-based standardized estimate of 0.68. Furthermore, when we accounted only for the difference in sex between the trial sample and target population, the hazard ratio was slightly weaker because (as shown in Tables 2 and 3) the trial selected for males and the treatment effect appeared stronger in males. Finally, when we accounted only for the difference in race/ethnicity between the trial sample and target population, the hazard ratio was stronger because the trial selected against blacks and the treatment effect appeared stronger in blacks.

All weighted estimates have wider confidence intervals than those in the trial, expressed in Table 3 as confidence limit ratios. The wider interval widths reflect the difference between the trial sample and target population. In fact, for the 3 single-attribute-weighted estimates in Table 3, the ranking of the confidence limit ratios accords with the distance between the hazard ratio for the trial sample and the hazard ratio for the target population.

When we simultaneously accounted for differences in age, sex, and race/ethnicity between the trial sample and target population, the hazard ratio was weakened from 0.51 to 0.57. This somewhat muted effect is apparent in Figure 1, which presents the complement of the Kaplan-Meier survival curves for the trial and the analogous curves (41) for the target population. Moreover, precision is somewhat decreased when inference is generalized to the target population, as is evident by the confidence limit ratios in Table 3.

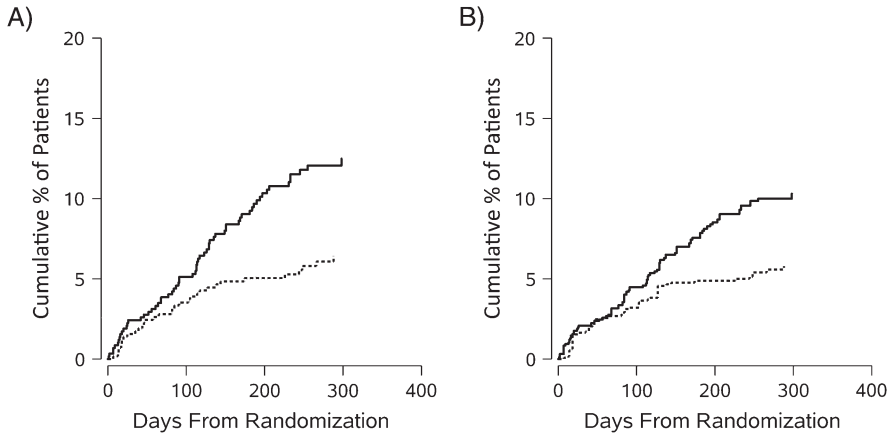

Figure 1. Complement of the Kaplan-Meier survival curves, acquired immunodeficiency syndrome (AIDS) Clinical Trial Group 320 Study, 1996–1997, United States. A) intent-to-treat; B) selection probability weighted. Solid lines represent patients randomly assigned to the control group; dashed lines represent patients randomly assigned to the treatment group.

In the next section, we describe a simulation experiment. Our goal was to assess some finite-sample properties of the proposed method.

SIMULATIONS

Simulation design

We compare the proposed method with conventional intent-to-treat estimates of the hazard ratio in a setting that mimics the ACTG 320 trial. To compare the approaches, we calculated bias, computed as the estimated log hazard ratio minus the true log hazard ratio (described below); standard error, computed as the average of the estimated standard errors; Monte Carlo standard error, computed as the standard deviation of the estimated log hazard ratios; root mean squared error, computed as the square root of the squared bias plus the squared Monte Carlo standard error; and confidence limit coverage, computed as the proportion of times the confidence limit contains the true hazard ratio (estimated with the standard error, not the Monte Carlo standard error). Simulation results are subject to Monte Carlo error; on the basis of the 10,000 simulations, the 95% confidence limit coverage estimates have a simulation standard error of about 0.2%.

A simulated data record comprises a value for Z, X, T, S; we drew $i = 1$ to 10,000 simulated population member records for each of 10,000 simulation data sets. First, a Bernoulli random variable was generated with marginal probability of 0.5 for a single demographic characteristic Z. Second, a Bernoulli random variable was generated with marginal probability 0.5 for treatment X. Third, a lognormal random variable was generated conditional on the realized value of Z and X with density

$$f = \exp\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{[\log(t) - \alpha_0 - \alpha_1 X_i - \alpha_2 X_i Z_i]^2}{2\sigma^2}\right\}\right).$$

The parameters $\{\alpha_0, \alpha_1, \alpha_2, \sigma\}$ were chosen to represent 3 scenarios. To inform these simulations, the parameters of a lognormal model (which was the best-fitting parametric model of those explored for the control group in the ACTG 320 trial) were $\alpha_0 = 8.585$ (standard error, 0.4340) and $\sigma = 2.709$ (standard error, 0.2885). First, we chose $\alpha_0 = 8.585$, $\alpha_1 = 0$, $\alpha_2 = 0$, and $\sigma = 2.709$ such that the expectation of the true hazard ratio was 1, and we term this scenario "no effect." Second, we chose $\alpha_0 = 8.585$, $\alpha_1 = 1.23$, $\alpha_2 = 0$, and $\sigma = 2.709$ such that the expectation of the true hazard ratio was approximately 0.5, as in the ACTG 320 trial; we term this scenario "homogeneous effect." Third, we chose $\alpha_0 = 8.585$, $\alpha_1 = 0.75$, $\alpha_2 = 2.10$, and $\sigma = 2.709$ such that the expectation of the true hazard ratio was again approximately 0.5, as in ACTG 320, and we term this scenario "heterogeneous effect."

The lognormal-distributed times were administratively censored at a fixed time such that, in expectation, we observed approximately 100 events per study, as in the ACTG 320 trial. In all cases, the reference for calculation of bias and confidence limit coverage was the true hazard ratio obtained in the complete target population. Last, a Bernoulli random variable was generated for selection into the trial S, conditional on the realized value of the demographic characteristic Z, as $1/\{1 + \exp[-\beta_0 - \beta_1 Z_i]\}$, with β_1 set at log(4) to reflect the size of selection effects observed in the ACTG 320 trial for age groups and β_0 chosen to maintain a marginal probability of 0.1. We calculated both naïve and robust (40) variance estimates for the weighted models.

Simulation results

Across all simulations, the estimated stabilized weights (in the selected samples) had a mean of 1.00 (standard deviation, ≅0.66) with minimum and maximum values of about 0.6 and 2.75, respectively.

In Table 4, for the no-effect and homogeneous-effect scenarios, both the conventional intent-to-treat estimate of the hazard ratio and the weighted estimate of the hazard ratio provide unbiased estimates with appropriate confidence limit coverage. In such cases, the conventional hazard ratio

Table 4. Simulation Results for 10,000 Samples per Scenario Each of Size 10,000 With 1,000 Patients per Trial

	Bias	Average SE	Monte Carlo SE	Root MSE	95% CL Coverage
Intent-to-treat					
No effect	−0.003	0.188	0.189	0.189	0.955
Homogeneous effect	−0.014	0.232	0.235	0.235	0.945
Heterogeneous effect	−0.842	0.229	0.230	0.810	0.070
Selection probability weighted					
No effect	−0.003	0.225	0.227	0.227	0.948
Homogeneous effect	−0.020	0.275	0.280	0.279	0.948
Heterogeneous effect	0.000	0.246	0.246	0.246	0.955

Abbreviations: CL, confidence limit; MSE, mean squared error; SE, standard error.

is more precise than the weighted hazard ratio, as evidenced by a 1.2-fold relative root mean squared error (0.227/0.189, Table 4).

For the heterogeneous-effect scenario in Table 4, the conventional estimate is severely biased, leading to abysmal confidence limit coverage. However, the weighted estimate is largely unbiased with appropriate confidence limit coverage. Use of naïve standard errors for the weighted estimate led to undercoverage of the confidence limits (coverage of 0.902, 0.906, and 0.871 for the 3 scenarios, respectively), but the use of robust standard errors raised the coverage to nominal levels, as shown in Table 4. Similar simulations were conducted for the odds ratio, with equally supportive results (data not shown).

DISCUSSION

We illustrated a method to generalize inferences from a randomized clinical trial to a specified target population using inverse probability-of-selection weights. In the ACTG 320 trial, the method demonstrated that inferences apply, albeit muted by 12%, to estimates of the US population infected in 2006, under the assumption that we have measured and correctly modeled the determinants of selection that reflect heterogeneity in the etiologic effect. The proposed method is supported by a limited Monte Carlo experiment.

The approach proposed here is one of model-based standardization (16, 42–44). Perhaps Lane and Nelder stated the idea most succinctly, in its simpler one-sample form, ''consider a survey of the incidence of disease in cattle, where proportions affected are recorded for different age groups in different regions. After the selection and fitting of a suitable model, the fitted proportions can be combined with population frequencies (assuming these to be known) and summed over regions, to give a prediction of the total incidence for the whole country'' (42, p. 614). Next, we provide some important caveats.

First, akin to the assumptions of no unmeasured confounders and no unmeasured informative censoring, in practice we will only at best be able to identify and measure a subset of the characteristics that lead to effect heterogeneity. Therefore, the proposed estimator will only approximate the etiologic effect in a defined target population to the extent that we capture said characteristics and specify them appropriately in the selection model. The proposed approach, or any that we can imagine, will need to have information on the joint distribution of the modifying characteristics in the target population and have these characteristics measured in the trial. In some settings, this may be difficult. In our case, we did not have individual-level data on the target population and instead used the summary statistics in the target population to construct a pseudopopulation with the appropriate joint distribution of sex, race, and age groups. The method could easily be extended to incorporate individual-level data on the population if it were available, as illustrated by Stuart et al. (45).

Furthermore, in our example, the CD4 cell count differs between the target population and the trial sample. Based on an understanding of the natural history of HIV infection, the target population of recently infected people has relatively normal immune function. However, the trial sample is immune suppressed, as shown in Table 1. If the treatment effect is heterogeneous with respect to immune function, then we would be missing an important characteristic. Indeed, the related issue of when to start HIV therapies with respect to CD4 cell count is of prime clinical concern (46–48) but beyond our current scope. Moreover, even with a correct set of measured characteristics, one must correctly specify the selection model to maintain valid inferences; this requirement is relaxed if the selection model is saturated, as in our example. By ''saturated'' we mean that there is a parameter for every cell in the cross-classified data table such that data are not smoothed (of course, data reflecting continuous factors, such as age, are categorized). Exploration of this central assumption of measuring the relevant heterogeneity factors requires more attention. However, it seems that even a partially corrected mapping of the trial result to a target population is a step forward.

A second caveat is that we concentrated on mapping an observed intent-to-treat result estimated in a trial to a specified target population defined by baseline characteristics that are measured in the trial and are known in the target population. Here, we did not attempt to account for post-randomization variables (12), but such steps may be important especially when the intent-to-treat estimate is subject to nontrivial bias due to noncompliance (49, 50).

Third, we ignored uncertainty in the distribution of characteristics in the target population. Information for the target population used here was based on a large, nationally representative sample. However, in settings where the distribution of characteristics in the target population is subject to a large amount of random error, the proposed method should be extended to account for this uncertainty. This is a topic for future research.

Fourth, the information observed for the trial patients is optimal for standardization to the same population structure, so the reduction in precision observed when generalizing from a trial to a target population will be an increasing function of the ''distance'' between the trial patients and target population. At one extreme, where the trial patients and target population are completely divergent, there is no information for estimating the effect in the target population, and one must rely on extrapolation.

Last, we illustrated methods here using the Cox proportional hazards model (27) because the hazard ratio is a central measure of association in randomized clinical trials. However, the hazard ratio is intrinsically susceptible to selection bias, and perhaps survival curves or relative survival times provide more clear measures of causal effects (51–53).

In conclusion, the proposed method standardizes observed trial results to a specified target population. Therefore, the proposed method provides direct information regarding generalizability of the trial results to the specified target population. The approach may prove useful in projecting the effects of interventions in populations that may differ in composition from those studied in randomized clinical trials. Moreover, the present approach can itself be generalized to nonrandomized settings, which is a topic we intend to discuss in future work.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Gillings School of Global Public Health and Center for AIDS Research, University of North Carolina, Chapel Hill, North Carolina (Stephen R. Cole); Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (Elizabeth A. Stuart); and Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (Elizabeth A. Stuart).

Dr. Cole was supported in part through National Institutes of Health (NIH) grants R03-AI-071763, R01-AA-01759, and P30-AI-50410. Dr. Stuart was supported in part through NIH grant K25-MH083946.

The authors thank Drs. Sander Greenland and Tyler VanderWeele for their expert advice.

Conflict of interest: none declared.

REFERENCES

1. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006; 60(7):578–586.
2. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology.* 2009;20(1):3–5.
3. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology.* 2009;20(6):880–883.
4. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656–664.
5. Robins JM, Morgenstern H. The foundations of confounding in epidemiology. *Comput Math Appl.* 1987;14:869–916.
6. Greenland S. Randomization, statistics, and causal inference [see comments]. *Epidemiology.* 1990;1(6):421–429.
7. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology.* 2004;15(5): 615–625.
8. Szklo M. Population-based cohort studies. *Epidemiol Rev.* 1998;20(1):81–90.
9. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med.* 2001;134(8):663–694.
10. Greenhouse JB, Kaizar EE, Kelleher K, et al. Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. *Stat Med*. 2008;27(11): 1801–1813.
11. Weisberg HI, Hayden VC, Pontes VP. Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality? *Clin Trials*. 2009;6(2):109–118.
12. Frangakis C. The calibration of treatment effects from clinical trials to target populations. *Clin Trials*. 2009;6(2): 136–140.
13. Gandhi M, Ameli N, Bacchetti P, et al. Eligibility criteria for HIV clinical trials and generalizability of results: the gap between published reports and study protocols. *AIDS*. 2005; 19(16):1885–1896.
14. Barkan SE, Melnick SL, Preston-Martin S, et al. The Women's Interagency HIV Study. WIHS Collaborative Study Group. *Epidemiology.* 1998;9(2):117–125.
15. Miettinen OS. Standardization of risk ratios. *Am J Epidemiol*. 1972;96(6):383–388.
16. Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology.* 3rd ed. New York, NY: Lippincott-Raven; 2008.
17. Hammer SM, Squires KE, Hughes MD, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team. *N Engl J Med*. 1997;337(11): 725–733.
18. Sommer A, Zeger SL. On estimating efficacy from clinical trials [published erratum appears in *Stat Med*. 1994;13(18): 1897]. *Stat Med*. 1991;10(1):45–52.
19. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*. 2000;56(3):779–788.
20. Cain LE, Cole SR. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. *Stat Med.* 2009;28(12):1725–1738.
21. Hall HI, Song R, Rhodes P, et al. Estimation of HIV incidence in the United States. *JAMA*. 2008;300(5):520–529.
22. Subpopulation estimates from the HIV incidence surveillance system—United States, 2006. *MMWR Morb Mortal Wkly Rep*. 2008;57(36):985–989.
23. Brookmeyer R, Quinn TC. Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. *Am J Epidemiol*. 1995; 141(2):166–172.
24. Janssen RS, Satten GA, Stramer SL, et al. New testing strategy to detect early HIV-1 infection for use in incidence estimates

and for clinical and prevention purposes. *JAMA*. 1998;280(1): 42–48.
25. Cole SR, Chu H, Brookmeyer R. Confidence intervals for biomarker-based human immunodeficiency virus incidence estimates and differences using prevalent data. *Am J Epidemiol*. 2007;165(1):94–100.
26. Karon JM, Song R, Brookmeyer R, et al. Estimating HIV incidence in the United States from HIV/AIDS surveillance data and biomarker HIV test results. *Stat Med*. 2008;27(23):4617–4633.
27. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc (B)*. 1972;34:187–220.
28. Cox DR. Partial likelihood. *Biometrika*. 1975;62(2):269–276.
29. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc.* 1952;47(260):663–685.
30. Kish L. Weighting for unequal P_i. *J Off Stat*. 1992;8(2): 183–200.
31. Binder DA. Fitting Cox's proportional hazards models from survey data. *Biometrika*. 1992;79(1):139–147.
32. Lin DY. On fitting Cox's proportional hazards model to survey data. *Biometrika*. 2000;87(1):37–47.
33. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550–560.
34. Haneuse S, Schildcrout J, Crane P, et al. Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology.* 2009;32(3): 229–239.
35. Pan Q, Schaubel DE. Evaluating bias correction in weighted proportional hazards regression. *Lifetime Data Anal*. 2009; 15(1):120–146.
36. Hoggatt KJ, Greenland S, Ritz BR. Adjustment for response bias via two-phase analysis: an application. *Epidemiology.* 2009;20(6):872–879.
37. White HA. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982;50(1):1–25.
38. Hanley JA, Negassa A, Edwardes MD, et al. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol*. 2003;157(4): 364–375.
39. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Stat Assoc.* 1989;84(408):1065–1078.
40. Robins JM. Marginal structural models. *1997 Proceedings of the Section on Bayesian Statistical Science*. Alexandria, VA: American Statistical Association; 1998:1–10.
41. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. *Comput Methods Programs Biomed*. 2004;75(1):45–49.
42. Lane PW, Nelder JA. Analysis of covariance and standardization as instances of prediction. *Biometrics*. 1982;38(3): 613–621.
43. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc.* 1987;82(398):387–394.
44. Greenland S. Estimating standardized parameters from generalized linear models. *Stat Med.* 1991;10(7):1069–1074.
45. Stuart EA, Cole SR, Bradshaw CP, et al. The use of propensity scores to assess the generalizability of results from randomized trials. Baltimore, MD: Department of Biostatistics, Johns Hopkins University; 2010. (Working paper 210). (http://www.bepress.com/jhubiostat/paper210).
46. Kitahata MM, Gange SJ, Abraham AG, et al. Effect of early versus deferred antiretroviral therapy for HIV on survival. *N Engl J Med*. 2009;360(18):1815–1826.
47. When to Start Consortium. Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet*. 2009; 373(9672):1352–1363.
48. Cain LE, Hernán MA. Dynamic marginal structural models to find optimal treatment regimes. *Am J Epidemiol*. 2009; 169(suppl 11):S41.
49. Cole SR, Chu H. Effect of acyclovir on herpetic ocular recurrence using a structural nested model. *Contemp Clin Trials*. 2005;26(3):300–310.
50. Greenland S, Lanes S, Jara M. Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. *Clin Trials*. 2008;5(1):5–13.
51. Flanders WD, Klein M. Properties of 2 counterfactual effect definitions of a point exposure. *Epidemiology.* 2007;18(4): 453–460.
52. Cole SR, Chu H, Nie L. Nonparametric estimator of relative time with application to the Acyclovir Prevention Trial. *Clin Trials*. 2009;6(4):320–328.
53. Hernán MA. The hazards of hazard ratios. *Epidemiology.* 2010;21(1):13–15.
54. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937–2960.

APPENDIX

In this Appendix, we provide a proof of the asymptotic consistency of the proposed method for the case of the mean or proportion; results extend to the hazard but are not proven here. As a preliminary step, we derive the bias in the conventional intent-to-treat estimator.

Let $\alpha = E(Y^1) - E(Y^0)$ and $\beta = E(Y^1 | S = 1) - E(Y^0 | S = 1)$, where $E(.|.)$ is the conditional expectation taken with respect to an enumerated target population indexed by $i = 1$ to n, Y_i^x is the potential outcome that would have occurred for person i under treatment $X = x$, and $S \in \{0, 1\}$ is an indicator of pretreatment selection into a sample of the target population. Although public health practitioners and clinicians may be interested in the population average treatment effect α, a conventional intent-to-treat comparison of groups randomized to treatment $X = x$ from the sample provides an estimate of β.

If $\alpha \neq E(Y^1 | Z = z) - E(Y^0 | Z = z)$ and $P(S = 1) \neq P(S = 1 | Z = z)$ for a pretreatment covariate Z, where $P(.|.)$ is a conditional probability taken with respect to the target population, then, except in circumstances of chance balancing cancellations, $\alpha \neq \beta$. Informally, the expectation of the difference in potential outcomes in the target population differs from the expectation of the difference in potential outcomes in an observed sample of the target population defined by $S = 1$ when there is heterogeneity in the causal effect of treatment X due to Z and the sample selection mechanism depends on Z.

If we limit our setting only in that $Z \in \{0, 1\}$, then the bias of β, as a measure of α, is $b_{xz} \times \left\{ \frac{P(Z = 1)}{P(S = 1)} \times \Big[P(S = 1 \mid Z = 1) - P(S = 1) \Big] \right\}$, where b_{xz} is a coefficient representing the heterogeneity in the X effect due to Z in the outcome model $E(Y_i) = b_0 + b_x X_i + b_{xz} X_i Z_i$. A main effect

for Z can be added to the outcome model without inducing any problem other than unnecessarily complicating the steps below. Therefore, the bias depends positively on the effect heterogeneity, the prevalence of the heterogeneity characteristic Z, the proportion of the target population not sampled, and the strength of the association between the heterogeneity characteristic and selection. In particular, there is no bias if there is no heterogeneity in the X effect due to Z, no one in the population has the heterogeneity characteristic (i.e., $P(Z=1)=0$), the "sample" consists of the entire population (i.e., $P(S=1)=1$), or sample selection does not depend on Z (i.e., $P(S=1|\,Z=z)=P(S=1)$). A derivation of the bias follows. First,

$$\begin{aligned}\alpha &= E\left(Y^1\right) - E\left(Y^0\right)\\ &= [b_0 + b_x + b_{xz}P(Z=1)] - b_0\\ &= b_x + b_{xz}P(Z=1),\end{aligned}$$

and

$$\begin{aligned}\beta &= E\left(Y^1|\,S=1\right) - E\left(Y^0|\,S=1\right)\\ &= \frac{1}{\sum_{i=1}^N S_iX_i}\sum_{i=1}^N S_iX_iY_i - \frac{1}{\sum_{i=1}^N S_i(1-X_i)}\sum_{i=1}^N S_i(1-X_i)Y_i\\ &= \frac{1}{\sum_{i=1}^N S_iX_i}\sum_{i=1}^N S_iX_i[b_0 + b_x + b_{xz}Z_i] -\\ &\quad \frac{1}{\sum_{i=1}^N S_i(1-X_i)}\sum_{i=1}^N S_i(1-X_i)(b_0)\\ &= \frac{1}{\sum_{i=1}^N S_iX_i}b_0\sum_{i=1}^N S_iX_i + b_x\sum_{i=1}^N S_iX_i + b_{xz}\sum_{i=1}^N S_iX_iZ_i - b_0\\ &= b_0 + b_x + b_{xz}\frac{1}{\sum_{i=1}^N S_iX_i}\sum_{i=1}^N S_iX_iZ_i - b_0\\ &= b_x + b_{xz}P(Z=1|\,S=1,X=1),\end{aligned}$$

where $Y_i = X_iY_i^1 + \left(1-X_i\right)Y_i^0$ by the consistency assumption (2). Then,

$$\begin{aligned}\beta-\alpha &= [b_x + b_{xz}P(Z=1|\,S=1,X=1)] - [b_x + b_{xz}P(Z=1)]\\ &= b_x + b_{xz}P(Z=1|\,S=1,X=1) - b_x - b_{xz}P(Z=1)\\ &= b_{xz}\times[P(Z=1|\,S=1,X=1) - P(Z=1)]\\ &= b_{xz}\times\left[\frac{P(S=1|Z=1)P(Z=1)}{P(S=1)} - P(Z=1)\right]\\ &= b_{xz}\times\left[\frac{P(S=1|Z=1)P(Z=1)}{P(S=1)} - \frac{P(Z=1)P(S=1)}{P(S=1)}\right]\\ &= b_{xz}\times\left[\frac{P(S=1|Z=1)P(Z=1) - P(Z=1)P(S=1)}{P(S=1)}\right]\\ &= b_{xz}\times\left[\frac{P(Z=1)}{P(S=1)}[P(S=1\mid Z=1) - P(S=1)]\right],\end{aligned}$$

where, in the fourth line of the above derivation, $P(Z=1\mid S=1,X=1) = \dfrac{P(S=1|\,Z=1)P(Z=1)}{P(S=1)}$ by Bayes' rule. We have also assumed the independence condition $E(X|\,S,Z,Y^x) = E(X)$. This condition says that treatment assignment is independent of sample selection, the covariate, and the potential outcomes. This ignorable treatment mechanism would be granted in expectation by random treatment assignment that does not depend on S or Z. This assumption may be relaxed for stratified randomization or nonrandomized studies (where $E(X)$ may depend on Z), but that leads to a more complex formula for the bias.

Assume further that $P(S=1|\,Z=z,Y^x) = P(S=1|\,Z=z)$ or, in words, that we have an ignorable sample selection mechanism, conditional on Z, and that $P(S=1|\,Z=z)>0$ for all individuals. Such a mechanism would be granted in expectation by Z-stratified random sampling from the target population, where every individual has a positive probability of being selected.

Let $\gamma = E\left[\dfrac{SXY}{w(Z)e(\emptyset)}\right] - E\left[\dfrac{S(1-X)Y}{w(Z)e(\emptyset)}\right]$, where $w(Z) = P(S=1|\,Z=z)$ and $e(\emptyset) = P(X=x)$. Then, γ is an inverse probability-of-selection-weighted expectation of the difference in potential outcomes in an observed sample of the target population, and $\gamma=\alpha$ under the above-stated assumptions. A proof builds on the work of Horvitz and Thompson (29) and extends that of Lunceford and Davidian (54) to allow for selection as

$$\begin{aligned}E\left[\frac{SXY}{w(Z)e(\emptyset)}\right] &= E\left\{E\left[\frac{I(S=1)I(X=1)Y}{w(Z)e(\emptyset)}\mid Y,Z\right]\right\}\\ &= E\left\{E\left[\frac{I\left(S=1\right)I\left(X=1\right)Y^1}{w(Z)e(\emptyset)}\mid Y^1,Z\right]\right\}\\ &= E\left\{\frac{Y^1}{w(Z)e(\emptyset)}E[I(S=1)I(X=1)\mid Y^1,Z]\right\}\\ &= E\left[\frac{Y^1}{w(Z)e(\emptyset)}P(S=1\mid Z=z)P(X=1)\right]\\ &= E\left(Y^1\right),\end{aligned}$$

where $I(.)$ is the indicator function. Steps are given by the law of conditional expectations, the consistency assumption, rearrangement, Z-conditional ignorable selection and ignorable treatment mechanisms, and the definitions of $w(Z)$ and $e(\emptyset)$, respectively. Similarly, $E\left[\dfrac{S(1-X)Y}{w(Z)e(\emptyset)}\right] = E(Y^0)$.