CAUSAL INFERENCE IN HIV/AIDS RESEARCH: GENERALIZABILITY AND APPLICATIONS

Ashley L. Buchanan

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics.

Chapel Hill
2014

Approved by:

Michael G. Hudgens

Adaora A. Adimora

Shrikant I. Bangdiwala

Stephen R. Cole

Paul W. Stewart

ABSTRACT

ASHLEY L. BUCHANAN: Causal Inference in HIV/AIDS Research: Generalizability and Applications
(Under the direction of Michael G. Hudgens)

In this research, we develop and apply causal inference methods for the field of infectious diseases. In the first part of this research, we consider an inverse probability (IP) weighted Cox model to estimate the effect of a baseline exposure on a time-to-event outcome. IP weighting can be used to adjust for multiple measured confounders of a baseline exposure in order to estimate marginal effects, which compare the distribution of outcomes when the entire population is exposed versus the entire population is unexposed. IP weights can also be employed to adjust for selection bias due to loss to follow-up. This approach is illustrated using an example that estimates the effect of injection drug use on time until AIDS or death among HIV-infected women.

In the second part of this research, we develop and apply methods for generalizing trial results for continuous data. In a randomized trial, assuming participants are a random sample from the target population may be dubious. Lack of generalizability can arise when the distribution of treatment effect modifiers in trial participants is different from the distribution in the target. We consider an inverse probability of sampling weighted (IPSW) estimator for generalizing trial results to a user-specified target population. The IPSW estimator is shown to be consistent and asymptotically normal. Expressions for the asymptotic variance and a consistent sandwich-type estimator of the variance are derived. Simulation results comparing the IPSW estimator to a previously proposed stratified estimator are provided. The IPSW estimator is employed to generalize results from the AIDS Clinical Trials Group (ACTG) to all people currently living with HIV in the U.S.

In the third part of this research, we develop and apply methods for generalizing trial

results for right-censored data. The IPSW estimator is considered for right-censored data and is defined as an inverse weighted Kaplan-Meier (KM) estimator. Simulation results are provided to compare this estimator to an unweighted KM estimator and a stratified estimator. The average standard error is computed using a nonparametric bootstrap. The IPSW estimator is employed to generalize survival results from the ACTG to all people currently living with HIV in the U.S.

To my parents, who were always supportive,

and my dog, Jolie, who was always by my side.

ACKNOWLEDGMENTS

Many thanks to my committee members Dr. Adaora Adimora, Dr. Shrikant Bangdiwala, and Dr. Paul Stewart for their helpful comments. I would like to thank Dr. Michael Hudgens for guidance and inspiration that greatly strengthened this research. I would also like to thank Dr. Stephen Cole for providing me with the opportunity to apply causal inference methods and invaluable instruction on epidemiological research. I would also like to thank my friends in the Biostatistics and Epidemiology Departments, who made classes and research a joy: Amy Richardson, Tara Rao, Eric Jay Daza, Joe Rigdon, Catherine Lesko, and Katie Mollan. These findings are presented on behalf of the Women's Interagency HIV Study (WIHS), the Center for AIDS Research (CFAR) Network of Integrated Clinical Trials (CNICS), and the AIDS Clinical Trials Group (ACTG). We would like to thank all of the WIHS, CNICS, and ACTG investigators, data management teams, and participants who contributed to this project. Funding for this study was provided by National Institutes of Health (NIH) grants R01AI100654, R01AI085073, U01AI042590, U01AI069918, R56AI102622, 5 K24HD059358-04, 5 U01AI103390-02 (WIHS), R24AI067039 (CNICS), and P30AI50410 (CFAR). The views and opinions expressed in this research do not necessarily state or reflect those of the NIH.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In this research, we develop and apply causal inference methods for the field of infectious diseases. In the first part of this research, we consider inverse probability (IP) weighted Cox models as an alternative to the standard Cox model. Survival analysis can be used in infectious disease research to compare the time to occurrence of clinical events between treatment or exposure groups (Cole and Hudgens, 2010). Randomized trials are the gold standard to estimate exposure effects on survival time, but are not always ethical or feasible. Although observational studies may provide estimates of effects when trial data are unavailable, the estimates they yield are often riddled with confounding (Greenland and Morgenstern, 2001). Informally, confounding occurs when the exposure and outcome share a common cause. The Cox proportional hazards regression model (Cox, 1972), the standard approach in survival analysis, can account for multiple measured confounders. As an alternative to the standard Cox model, we present a method in Chapter 3 that uses IP weights to estimate the effect of a baseline exposure on survival time. Under certain assumptions, results from an IP-weighted Cox model of observational data can be interpreted similar to a randomized trial with no drop out (i.e., loss to follow-up).

In the second part of this research, we develop and apply methods for generalizing trial results for continuous data. Results obtained in randomized trials may not generalize to a target population. Ideally, trial participants are a random sample from a target population and the treatment assignment mechanism is known to the analyst. In a randomized trial, the treatment assignment mechanism is always known, but trial participants are often a non-random sample from a target population. Lack of generalizability can arise when the distribution of treatment effect modifiers in trial participants is different from the distribution in a target population. Following Cole and Stuart (2010) and Stuart et al. (2011), we consider

an inverse probability of sampling weighted (IPSW) estimator for generalizing trial results to a target population. In Chapter 4, the IPSW estimator is shown to be consistent and asymptotically normal. Expressions for the asymptotic variance and a consistent sandwich-type estimator of the variance are also derived.

In the third part of this research, we develop and apply methods for generalizing trial results for right-censored data. In Chapter 5, the IPSW estimator is considered for right-censored data and is defined as an inverse weighted Kaplan-Meier (KM) estimator. Simulation results are provided to compare this estimator to an unweighted KM estimator and a stratified estimator. The average standard error is computed using a nonparametric bootstrap and performance is evaluated empirically. The IPSW estimator is employed to generalize survival results from the acquired immunodeficiency syndrome (AIDS) Clinical Trials Group (ACTG) to all people currently living with human immunodeficiency virus (HIV) in the U.S.

CHAPTER 2: LITERATURE REVIEW

2.1 Classical Causal Inference

Public health researchers are often interested in estimating causal effects of treatment or exposures using data from randomized clinical trials (RCTs) or observational studies. Causal inference is a paradigm to estimate effects and is often framed using potential outcomes (Little and Rubin, 2000). A potential outcome is defined as the outcome that would have been observed had a participant (possibly contrary to fact) been exposed to a certain level of treatment or exposure. In general, Y^x is potential outcome under treatment $X = 1$ or lack of treatment $X = 0$. These potential outcomes are also referred to as factuals and counterfactuals in the literature (Morgan and Winship, 2007). This framework allows for extensions beyond the randomized trial to accommodate scenarios such as observational studies, noncompliance, or missing data (Holland, 1986; Robins and Finkelstein, 2000; Rubin, 1990). The notation of potential outcomes is historically related to Neyman's randomization-based inference (Neyman et al., 1992; Rubin, 1990; Robins, 1989). Once a participant is assigned to treatment, only one of the two potential outcomes is observed; thus, the problem of causal inference is akin to a missing data problem.

There are two randomization-based approaches to causal inference developed by Fisher (Fisher, 1973; Rubin, 1980) and Neyman et al. (1992). Using the Fisher approach, the sharp null hypothesis is evaluated, which states that the outcome is the same for both treatment groups for all participants (i.e., $Y_i^1 = Y_i^0$), where i indexes the study participants. Under the sharp null hypothesis, all information can be identified from the observed data (i.e., $Y_i^1 = Y_i^0 = Y_i$) and test statistics and P values can be computed. Little and Rubin (2000) argued that this approach is limited because the null hypothesis is restrictive and significant tests may not be clinically meaningful.

In Neyman's approach to causal inference, expectations of statistics are evaluated with respect to the distribution of the assignment mechanism and confidence intervals are calculated for the average causal effect (Little and Rubin, 2000). This idea is historically related to Neyman's randomization-based inference in surveys (Neyman et al., 1992). An unbiased estimator of the causal estimand and an unbiased (or upwardly biased) estimator of the variance is derived. The central limit theorem allows for the construction of confidence intervals. Let $\bar{Y}^1$ and $\bar{Y}^0$ be the averages of the potential outcomes in the population. Let $\bar{y}^1$ and $\bar{y}^0$ denote the sample means among those assigned to treatment and control, respectively. Let n_1 denote the number assigned to treatment, n_0 denote the number assigned to control, s_1^2 denote the variance among those assigned to treatment, and s_0^2 denote the variance among those assigned to control. The estimated variance is $\text{se}^2 = s_1^2/n_1 + s_0^2/n_0$. The 95% confidence interval for the average treatment difference $(\bar{Y}^1 - \bar{Y}^0)$ is $(\bar{y}^1 - \bar{y}^0) \pm 1.96 \times$ se. Neyman's approach to causal inference is commonly used in epidemiological studies and is employed throughout this research.

There are several assumptions necessary to estimate causal effects. First, there needs to be no interference between participants and treatment variation irrelevance needs to hold (i.e., stable unit treatment value assumption (SUTVA)) (Rubin, 1980). This means that the potential outcomes for any participant do not vary with the treatments assigned to other participants, and, for each participant, there are no different forms or versions of treatment which lead to different potential outcomes. This implies that the study design needs a well-defined treatment assignment mechanism (Robins et al., 2000), so there are not multiple versions of exposure, or if there are, they are unimportant (Cole and Frangakis, 2009; Pearl, 2010; VanderWeele, 2009a). In general, consistency always holds (i.e., $Y = Y^1X + Y^0(1-X)$). We must have measured enough variables so that we can effectively address confounding (Robins et al., 2000). Effectively addressing confounding can lead to exchangeability, which means the potential outcomes are independent of the exposure. In other words, the exposure assignment mechanism depends only on the data through the measured covariates and outcome (Little and Rubin, 2000). Confounders are variables associated with the exposure and independent risk factors for the outcome that are not affected by exposure (Greenland and Morgenstern,

2001). Informally, confounding occurs when exposure and outcome share a common cause. In a trial, exchangeability is gained through randomization. Lastly, the conditional probability of receiving every level of treatment given covariates must be greater than zero (i.e., positivity) (Cole and Hernan, 2008).

Randomization of exposure allows for straightforward estimation of causal effects. In a completely randomized design, exposure is not associated with any potential confounders. Estimation of causal effects is not possible when there is unmeasured confounding. In many public health studies, it is not ethical or feasible to randomize the treatment or exposure. Principles of causal inference can be applied to estimate effects in observational studies in the presence of (measured) confounding. In section 2.2, we present the literature for IP-weighted Cox models, which is one approach to applying causal inference methods to estimate effects with observational data.

Causal inference methods can also be employed to sharpen inference from randomized trials. Informative censoring and generalizability of results are two possible concerns in trials (Cole and Stuart, 2010; Stuart et al., 2011; Frangakis and Rubin, 1999). A recent paper by Hernan et al. (2013) discussed how randomized trials may be subject to post-randomization selection bias and confounding, particularly for studies with longer follow-up. The authors argued that intention-to-treat analyses may not always be estimating the effects of interest. The authors also proposed a set of g-methods, including inverse probability weighting methods. In Section 2.3, we present the literature for generalizing trial results, highlighting recently proposed methods using sampling scores.

2.2 Inverse Probability Weighted Cox Models

The Cox proportional hazards regression model (Cox, 1972) is the standard approach to account for multiple measured confounders in observational studies with separate curves presented for each baseline covariate group. As an alternative to the standard Cox model, IP weights can be utilized to estimate the effect of an exposure that is fixed at study entry (Cole and Hernan, 2004; Nieto and Coresh, 1996; Xie and Liu, 2005). IP weighting creates a pseudo-

population in which confounders are no longer associated with the exposure. Under certain assumptions, results from an IP-weighted Cox model of observational data can be interpreted similar to results from a randomized trial with no drop out (i.e., loss to follow-up). One curve represents the survival if everyone (possibly contrary to fact) had been exposed at baseline, while the other curve represents the survival if no one (possibly contrary to fact) had been exposed at baseline (Cole and Hernan, 2004; Xie and Liu, 2005). Herein, we refer to IP weighting as standardization, where the standardization is to the entire population under two different exposures (Cole and Hernan, 2004; Sato and Matsuyama, 2003).

An IP-weighted Cox model is fit by maximizing a weighted partial likelihood accounting for confounding and possibly informative drop out measured by covariates through the estimated IP weight $\hat{w}_i(t)$. The estimated IP weight $\hat{w}_i(t)$ is the product of an estimated time-fixed IP exposure weight $\hat{w}_{1i}(t)$ and an estimated time-varying IP drop out weight $\hat{w}_{2i}(t)$ for each participant i at each survival time t. If certain assumptions are met, IP weighting can account for confounding and selection bias due to drop out by multiple covariates using both exposure and drop out weights. Standardized survival curve estimates can be obtained by fitting an IP-weighted Cox model stratified by exposure with no covariates and then nonparametrically estimating the baseline survival functions for the two strata, which are (asymptotically) equivalent to Kaplan-Meier estimates in the absence of weighting (Cole and Hernan, 2004; Collett, 2003).

The standardized (i.e., IP weighted) method provides potential benefits that the covariate-adjusted method lacks. First, results from the standardized approach can be used to mimic a randomized trial when only observational data is available (under certain assumptions). In particular, the estimated hazard ratio using the standardized approach can be interpreted the same as the (marginal) hazard ratio one would obtain in a randomized experiment such as a clinical trial where there is no drop out. In contrast, a covariate-adjusted Cox model hazard ratio does not necessarily equal the marginal hazard ratio (even in the absence of confounding) because the Cox model is not collapsible for the hazard ratio parameter (Greenland, 1996). A regression model is said to be collapsible for a parameter (in this case, the hazard ratio) if

the covariate-adjusted parameter is the same as the unadjusted parameter (Greenland et al., 1999b).

Second, the IP weighting approach yields standardized survival curve estimates. Although the hazard ratio is a common summary parameter to compare survival distributions between exposure groups, there are drawbacks to focusing inference on hazard ratios. For instance, the hazard ratio can be difficult to interpret, especially when trying to summarize the effect of a treatment or exposure (Hernan, 2010). Presenting estimated survival curves is an alternative to reporting hazard ratios that may be more interpretable because survival curves summarize all information from baseline up to any time t. The IP-weighted approach leads to Kaplan-Meier type survival curve estimates that are standardized to the entire population under two different exposures at baseline while accounting for confounding by multiple covariates. A covariate-adjusted Cox model does not afford such survival curve estimates (Cole and Hernan, 2004; Xie and Liu, 2005).

Third, the IP-weighted approach with drop out weights requires a weaker assumption about censoring than the covariate-adjusted Cox model. Specifically, if there are measured time-varying covariates predictive of censoring and survival time, the IP-weighted approach will yield consistent estimates of the marginal hazard ratio, while the covariate-adjusted Cox model estimator will not be consistent for the marginal or conditional hazard ratio (Hernan et al., 2000; Robins et al., 2000; Robins and Finkelstein, 2000).

When interest focuses on exposures that change over time, methods must be adapted accordingly. When a time-varying confounder is a risk factor for the outcome, predicts later exposure, and is affected by prior exposure, standard statistical methods (e.g., Cox models with endogenous time-varying covariates) are biased and fail to provide consistent estimators of effects (Cole et al., 2003; Hernan et al., 2001, 2013; Robins et al., 2000). IP weighting can be generalized to account for time-varying confounders (Robins et al., 2000). For example, in HIV-infected individuals, CD4 count is a risk factor for death, predicts subsequent treatment with antiretroviral therapy, and is affected by prior treatment; thus, the IP-weighted Cox model is appropriate for studying the effect of time-varying antiretroviral therapy on overall

survival while adjusting for time-varying CD4 count.

There are several papers that provide theoretical justification and illustrative examples for IP-weighted Cox models. Robins *et al.*(Robins, 1998; Robins et al., 2000) demonstrated that the parameters of a marginal structural model can be consistently estimated using IP-weighted estimators. Hernan et al. (2000) provided an illustrative example using these methods to adjust for time-varying confounding to estimate the causal effect of zidovudine on the survival of HIV-infected men. A subsequent paper presents the use of these models to estimate the joint effect of two treatments (Hernan et al., 2001). Cole et al. (2003) employed IP-weighted Cox models to estimate the effect of highly active antiretroviral therapy on time to AIDS or death, which appropriately adjust for time-varying confounders affected by prior treatment or exposure. A subsequent paper provided additional guidelines on how to construct inverse probability weights (Cole and Hernan, 2008). Other efforts have been made to increase the understanding and utilization of causal inference methods among researchers (Petersen et al., 2006). In Chapter 3, we continue this effort by demonstrating how IP-weighted Cox models can be used to account for multiple measured confounders and selection bias due to drop out and providing graphical summaries of these effects. This approach is compared to a traditional Cox model and illustrated using an example that estimates the effect of injection drug use on AIDS-free survival among HIV-infected women.

2.3 Generalizability of Randomized Trials

Results obtained in RCTs may not generalize to target populations due to differences in characteristics between participants in the trial and target population, as well as the presence of effect modification by these same characteristics (Cole and Stuart, 2010; Stuart et al., 2011). Valid statistical inference depends both on the treatment allocation and the mechanism of trail participation. Ideally, both of these steps would be randomized; however, trials are often non-random samples from a target with the treatment assignment mechanism always known to the analyst.

Researchers are often interested in estimating a causal effect in a target population. In

simple settings, trial results can be mapped to a target population using nonparametric direct standardization (Rothman, 1986), which can accommodate only on a few categorical covariates; however, when there are many covariates or some covariates are continuous, direct standardization is no longer possible. The second part of this research addresses the situation often seen in a clinical trial: a known treatment assignment mechanism, but non-random selection of participants from a target population (Little and Rubin, 2000).

Generalizability is often a concern for clinical trials in public health research. One study highlighted the overrepresentation of African-American and Hispanic women among HIV cases in the U.S. and the limited clinical trial participation of members of these groups (Greenblatt, 2011). Another study reviewed eligibility criteria from 32 NIH-funded RCTs and applied those to data from the Women's Interagency HIV Study (WIHS). Of the 20 Adult Clinical Trials Group (ACTG) trials, 28% to 68% of the WIHS cohort would have been excluded (Gandhi et al., 2005). Historically, generalizability was assessed through comparisons of characteristics between the trial and target populations or comparisons of effects across various study samples (Weisberg et al., 2009). Recent developments in the literature have proposed novel quantitative approaches to evaluate generalizability of effects.

2.3.1 Definitions and Background

We define generalizability as the degree to which an internally valid measure of effect estimated in a sample from one population would change if the study had been conducted in a different target population. We view the terms "transportability" and "external validity" as synonymous with our definition of generalizability, although Pearl defines transportability specifically for the case where investigators would like to apply the results from a RCT to a population in which only an observational study is feasible or ethical (Bareinboim and Pearl, 2013).

In public health research, investigators would ideally like to estimate a treatment effect in a target population. In practice, a study sample is typically obtained from a source population that is likely different from the target and that information is used to estimate

effects (Rothman, 1986). The source population is often chosen instead of a target population due to such reasons as financial or time constraints and ethical considerations. The target population may be defined prior to designing or implementing a specific study, or researchers may be interested in drawing inference from a published study to a different target population. In any evaluation of generalizability, it is necessary to clearly define the target population. The sample average treatment effect (SATE) is the average treatment effect in the source population and the population average treatment effect (PATE) is average treatment effect in the target population (Stuart et al., 2011, 2014).

In an ideal randomized trial (i.e., assuming no confounding, full adherence to treatment, perfect blinding, no loss to follow-up), or in an observational study where the estimate of effect is identifiable (i.e., assuming exchangeability between the exposed and unexposed conditional on measured covariates, consistency within the study population, and a positive probability of exposure within each strata of covariates), the estimator in the study sample will be an unbiased estimate of the SATE, which we define as internal validity. Evaluation of generalizability should not be entertained unless the results are internally valid. Even under these ideal circumstances when our estimator is internally valid, the SATE may not be equal to the PATE (i.e., a lack of external validity). Results obtained in one study may not generalize to target populations due to 1) differences in the distribution of effect modifiers in the study population and target population; 2) the presence of interference; or 3) the existence of multiple versions of treatment (Bareinboim and Pearl, 2013; Hernan and VanderWeele, 2011; Stuart et al., 2011). In this research, we assume there is no interference and only one version of treatment, so we can focus on the scenario where the distribution of effect modifiers in the source population differs from the target population.

Most discussions of generalizability of trial results limit themselves to considering whether the study sample was representative of the target population. However, we suggest that a simple comparison of the distribution of characteristics between the study sample and target population is insufficient. An understanding of the characteristics that influence trial participation and modify the effect of interest is essential. Rothman et al. (2013) argues that

a representative sample is not necessary for generalizing estimates of causal effects if effects are homogeneous. Further, "it is not representativeness of the study subjects that enhances generalization, [but rather] it is knowledge of specific conditions and an understanding of mechanism that makes for a proper generalization" (Rothman et al., 2013). While the easiest and most efficient way to ensure generalizability to a specified target population is to ensure that a trial is a representative sample of the target (Stuart et al., 2011), without an understanding of factors that modify the effect of interest, the effect estimate from that trial may not be generalizable to target populations of interest.

Lack of generalizability assumes that there are two true estimates of effect (SATE and PATE) that may be different, even if the estimator of the SATE is unbiased (Stuart et al., 2014). In order for the SATE to be equal to the PATE, the following assumptions must hold: no effect modification by the characteristic related to trial participation, no interference, and treatment variation irrelevance. The first assumption means that the treatment effects are homogeneous or covariates related to trial participation are distinct from the treatment effect modifiers (Tipton, 2013). The last two assumptions mean that there is no interference of participants within the trial (i.e., the potential outcomes of one participant are assumed to be unaffected by the treatment assignment of other participants) and that there is only one version of treatment, or if there are multiple versions, they are irrelevant for the outcome (Rubin, 1980).

Several additional assumptions are needed to generalize trial results to target populations. Once in the trial, participants are randomly assigned to a study arm, so that participants from either treatment group are balanced in that covariate distribution is the same regardless of treatment assignment conditional on trial participation. This is the ignorable treatment assignment mechanism. We also assume an ignorable trial participation mechanism conditional on covariates. This means that participants in the trial are no different from nonparticipants in regards to the treatment-outcome relationship conditional on covariates. There are not multiple versions of treatment, or if there are, they do not affect the outcome (Cole and Frangakis, 2009; Pearl, 2010; VanderWeele, 2009a). Trial participation and treatment positivity

assumptions are also necessary (Cole and Hernan, 2008).

The easiest way to allow for generalizability is to ensure that a trial is a representative sample of the target in regards to the treatment effect modifiers (Stuart et al., 2011). However, this is not always feasible or appropriate. New methods make it possible to use less representative sampling for maximal internal validity, and reweight the effect estimate for improved external validity, assuming that all treatment effect modifiers are known and the distribution of those same characteristics is available in the target population.

2.3.2 Sampling Score Methods to Generalize Trial Results

There are several sampling-score methods to generalize an estimate from a trial to a target population. The sampling score is defined as the probability of inclusion in the trial given some function of covariates ($\mathbf{Z}$) that are both effect modifiers and associated with trial participation ($S = 1$). The effect in the target population can be estimated using sampling scores, including matching (Stuart et al., 2011), stratification (Tipton, 2013) and an inverse probability of sampling weighted (IPSW) estimator (Cole and Stuart, 2010; Stuart et al., 2011).

Cole and Stuart (2010) proposed a method to standardize trial results to a target population using an IP-weighted Cox model with sampling score weights. In the illustrative example, results from ACTG 320 were generalized to all people living with HIV in the U.S. in 2006. Characteristics of the target population were ascertained using estimates from the Centers for Disease Control (CDC). The proposed inverse probability of sampling weights were defined as $P(S = 1)/P(S = 1|\mathbf{Z})$, where the sampling scores were estimated using logistic regression. The Cox proportional hazard model was inverse weighted by these sampling scores to obtain a hazard ratio and estimates of the marginal survival curves in the target population. A robust estimate of the variance was employed (Robins, 1998); however, no closed-form expression for the variance was provided. Using the IP-weighted Cox model with sampling weights, the trial results applied to the target population, but were attenuated towards the null. In simulations with a heterogeneous treatment effect, an intent-to-treat estimator was biased and the

corresponding confidence interval had poor coverage; whereas, the proposed estimator was unbiased and its corresponding confidence interval had appropriate coverage. The authors provided an expression for the bias of the intention-to-treat estimator when the parameter of interest is a difference in means and a proof that an inverse probability weighted estimator is unbiased for the mean of the potential outcomes in the target population; however, consistency and asymptotic normality of the proposed estimator were not formally shown.

In a subsequent paper by Stuart et al. (2011), sampling-score methods were used to quantify the similarity between trial participants and those in a target population. These methods were also used to match, weight, and subclassify the outcomes among the controls to a target population. The assumptions for this method were discussed, including positivity, no unmeasured confounders, and random treatment assignment. Assuming an infinite target population, the authors provided a proof that an inverse probability weighted estimator among control participants is unbiased for the mean of the potential outcome under control in the target population $E(Y^0)$. However, large sample properties (i.e., consistency and asymptotic normality) of the proposed estimator were not derived. The Positive Behavioral Interventions Support (PBIS) study was generalized to all elementary schools in Maryland. The sampling scores $P(S = 1|\mathbf{Z})$ were estimated using logistic regression. The sampling score difference between the trial and a target population was defined. The sampling scores were used to inverse weight the naive estimator, produce a stratified estimator, and perform full matching. For inverse probability weighting, each control subject was given their own weight defined as the inverse of the sampling score. For stratification, the target population was divided into strata according to the distribution of the sampling score (i.e., quintiles) in the target and the weights were defined as the proportion in each stratum. The authors did not provide expressions of the variance for either the inverse probability weighted estimator or the stratified estimator. Full matching was performed ensuring that each subclass had at least one member of the sample and at least one member of the target population. For the PBIS example, the control group in the trial was comparable to the state level and the weighted means in the trial control participants were reasonable estimators of the true means at the state level. In Stuart et al. (2014), differences between external and interval validity were

elucidated and an overview of existing methods for generalizability was provided, including an illustrative example of generalizing results from the PBIS study.

Tipton (2013) proposed a stratified sampling score estimator, including a discussion of the necessary assumptions. This estimator was computed in the following steps. The sampling scores were used to create strata in the target population (e.g., defined by quintiles). The weight was defined as the proportion within each stratum in the population. The difference between the treated and untreated within each stratum was weighted and the weighted differences were summed across strata. Tipton also developed expressions for bias reduction and variance inflation as compared to the intention-to-treat estimator in the trial (Tipton, 2013). Tipton extended this method by proposing a stratified sample recruitment approach based on the sampling scores, which requires identification of the target population and trial eligibility criteria (Tipton et al., 2014). A related paper discussed the application of this estimator to inform future trials, as well as the use of this estimator in trial design (O'Muircheartaigh and Hedges, 2013).

The limitations of the stratified sampling score estimator compared to the IP-weighted estimator include that it is coarser (i.e., limited by the number of stratum defined by the sampling score) and it does not have the interpretation of creating a pseudo-population. The stratified estimator does not always immediately allow for estimation of marginal effects, which is the average effect comparing the target population where everyone (possibly contrary to fact) was exposed to the target population where no one (possibly contrary to fact) was exposed (Kaufman, 2010; Lunceford and Davidian, 2004). The sampling score stratified estimator may be biased when there is residual confounding within strata (e.g., $\mathbf{Z}$ is continuous), as the estimator based on stratification is not consistent for the PATE (Lunceford and Davidian, 2004). On the other hand, the IP-weighted estimator is sensitive to model specifications (i.e., assumes the sampling score model is correctly specified) and has been shown to perform poorly when sampling probabilities are small (Kang and Schafer, 2007).

Results in Chapter 4 continue this effort by considering an IP-weighted estimator for a difference of means in the target population, where the weights are defined as the inverse

of the sampling score. This estimator is shown to be consistent and asymptotically normal. Expressions for the asymptotic variance and a consistent sandwich-type estimator of the variance are also derived. The performance of the IP-weighted estimator is compared to a previously proposed stratified estimator in a simulation study.

2.3.3 Other Methods to Generalize Trial Results

Greenhouse et al. (2008) described and illustrated an approach for determining if results from a randomized trial are generalizable. They outlined a methodological approach using four steps: identify data sources, subset the data on demographic variables to allow comparisons to a target population, measure the outcome, and perform sensitivity analysis. The authors provided an illustrative example assessing the risk of suicidality among pediatric antidepressant users. A meta-analysis of RCTs demonstrated an increased risk and observational studies did not confirm that result (i.e., showed a decreased risk). One explanation is that the two studies were sampling from different populations (i.e., trial exclusion criteria could limit the representativeness of the trial sample). There could be treatment heterogeneity based on variables related to trial participation. In that case, generalizability of the results without adjustment may not be appropriate. In the illustrative example, the goal was to assess the representativeness of the trial participants. The rate of suicidal ideation and suicidal behaviors in the depressed adolescents who participated in the RCTs was approximately one-half the adjusted rate among depressed adolescents in the United States. The authors posit that exclusion of those at high risk of the event could lead to an upwardly biased rate ratio (Greenhouse et al., 2008). Although Greenhouse et al. (2008) highlighted this issue, their approach only allows for determining if results are generalizable or not, and does not posit a solution for the latter.

Some initial approaches to this problem have been suggested; however, they are limited in the scope of their application. Weisberg et al. (2009) suggested using four stratum of outcome and treatment types (doomed, immune, causal and preventive), then defined an expression for the bias using these stratum. A doomed individual will experience the event of interest

regardless of which treatment is received; an immune individual will be spared in either instance; a causal individual experiences the event only if assigned to the treatment group; and a preventive individual, only if assigned to the control. Each stratum had a population proportion (P) and a selection probability (S). The risk ratio in the target population was defined as $RR_{tar} = (P_1 + P_2)/(P_1 + P_3)$. The observed risk ratio in the RCT was defined as $RR_{obs} = (S_1P_1 + S_2P_2)/(S_1P_1 + S_3P_3)$. If the the trial selection mechanism is not ignorable (i.e. selection is related to an individual's risk of an event), the value of RR_{tar} will be different from RR_{obs}. The authors proposed that when participants at high risk are more likely to be excluded, there may be an upwardly biased relative risk estimate for the true relative risk in the target population; whereas, if those at a lower risk are more likely to be excluded, there may be a downwardly biased estimate of the relative risk (Weisberg et al., 2009). The framework suggested by Weisberg et al. (2009) is useful, but limited to binary data. Extensions are necessary to accommodate other types of data, such as right-censoring, commonly seen in RCTs.

Frangakis suggested a principal stratification approach to adjust for differences in post-treatment effects between the participants in a RCT and participants in an observational study (Frangakis, 2009). Principal stratification was initially used to compare treatments adjusting for post-randomization variables to account for selection bias (Frangakis and Rubin, 2002). Using this approach, causal effect are estimated within strata defined by cross-classification of participants by the joint potential outcomes of the post-randomization variable under treatment and control. It is appropriate to condition on the principal strata because they are not affected by treatment (Hudgens et al., 2003; Hudgens and Halloran, 2006). Principal stratification could be applied to address generalizability when a target population has a different distribution of principal strata (defined by intermediates on the causal path between exposure and outcome that contain information on both the treatment and individual differences) than the source population. An illustrative example of a trial with different treatment compliance rates from a target population is provided. A limitation of this approach is that the distribution of the outcome within principal strata must be known in the source and target populations (Frangakis, 2009).

Another related concept in the literature is transportability. A recent paper examined the transportability of effects of compound treatments (Hernan and VanderWeele, 2011). Compound treatments are defined as treatments with multiple versions (i.e., multiple realizations of the treatment can be mapped onto one value). Transportability is a question of estimating the average causal effect in a target population that is different from the source population. Transportability of compound treatment effects depends on effect modification, interference, and versions of compound treatment. The authors provided an expression for the counterfactual mean when versions of compound treatment are known for those in the source population, which requires all information on versions of compound treatment in the target population for estimation. Determining the versions of compound treatment in the source population and target population can be complicated. Versions of treatment are necessary to evaluate exchangeability and positivity to allow for transport of effects to other populations.

Bareinboim and Pearl (2013) provided a graphical condition for evaluating transportability and an algorithm for transportability of causal effects, which produces a transport formula whenever those results are in fact transportable . The authors defined transportability as "a license to transfer information learned in experimental studies to a different population, on which only observational studies can be conducted" (Bareinboim and Pearl, 2013). The authors provided a graphical condition for determining the transportability of causal effects.

2.4 Methods for Generalizing Right-Censored Data

Kaplan-Meier estimators are used to quantify survival distributions and are a commonly used nonparametric estimator in survival analysis (Kalbfleisch and Prentice, 2002). If the trial is random sample from the target population, standard methods, such as the Kaplan-Meier estimator, are appropriate and comparisons between groups can be made with a log rank test. However, when the trial is possibly a non-random sample from the target population, properly weighting the observed trial data to estimate the Kaplan-Meier may be necessary. The Kaplan-Meier estimator is also appealing because it does not require a proportional hazards assumption or possible complications faced by hazard ratios estimated using a Cox

proportional hazards model. The hazard ratio can be difficult to interpret, especially when trying to summarize the effect of a treatment or exposure (Hernan, 2010).

Presenting estimated survival curves is an alternative to reporting hazard ratios that may be more interpretable because survival curves summarize all information from baseline up to any time t (Cole and Hernan, 2004). Robins and Finkelstein (2000) proposed an adjusted Kaplan-Meier estimator using inverse probability (IP) of censoring weights to estimate effects in the presence of selection bias using in randomized trial data. Xie and Liu (2005) developed an adjusted Kaplan-Meier estimator using inverse probability of treatment weights to estimate effects in the presence of (measured) confounding in an observational study. They presented a method for estimating marginal survival curves, including the development of a weighted log rank test.

Cole and Stuart (2010) proposed a method to standardize trial results to a target population for right-censored data using an IP-weighted Cox model with sampling weights. They reported the hazard ratios and displayed estimated survival curves from this model. A robust estimator of the variance of the hazard ratio was employed, but the performance of the nonparametric bootstrap standard error was not evaluated. In the IP-weighted model, the results seen in the trial applied to the target population, but were attenuated to the null. The authors provide simulations demonstrating that the proposed estimator of the hazard ratio was (asymptotically) unbiased and its corresponding confidence interval had coverage around the nominal level. Research in Chapter 5 will continue this effort by considering an IP-weighted Kaplan-Meier estimator, comparing this estimator to a stratified estimator, and empirically evaluating the performance of the nonparametric bootstrap standard error.

2.5 Summary

In summary, a large body of literature has been developed to address concerns of (measured) confounding in observational studies. Statistical properties of the IP-weighted estimators were demonstrated (Robins, 1998; Robins et al., 2000). There are several epidemiological papers that clarified technical details and provided guidance on implementation of these mod-

els (Cole et al., 2003; Cole and Hernan, 2008; Hernan et al., 2000, 2001). Other efforts have been made to increase the understanding and utilization of causal inference methods among researchers (Petersen et al., 2006). In Chapter 3, we continue that effort by summarizing the literature for IP-weighted Cox models through a comparison to the traditional Cox model and an illustrative example in HIV/AIDS research.

There is a growing literature on methods for generalizing trial results; however, these approaches do not develop the statistical properties or provide closed-form expressions for the estimators of the variance. Following Cole and Stuart (2010) and Stuart et al. (2011), we consider an inverse probability of sampling weighted (IPSW) estimator for generalizing trial results to a target population in Chapter 4, where the parameter of interest is a difference in average potential outcomes in a target population. We show that the IPSW estimator is consistent and asymptotically normal, and provide a closed-form expression for a consistent estimator of the variance. The IPSW estimator is employed to generalize results from the ACTG to all people currently living with HIV in the U.S. Following Cole and Stuart (2010) and Buchanan, et al. (In preparation), we consider an IPSW estimator for right-censored outcomes, which is defined as an inverse weighted Kaplan-Meier estimator. In Chapter 5, the IPSW estimator is employed to generalize survival effects from the ACTG to all people currently living with HIV in the U.S.

CHAPTER 3: WORTH THE WEIGHT: USING INVERSE PROBABILITY WEIGHTED COX MODELS IN AIDS RESEARCH

3.1 Introduction

Survival analysis is often used in infectious disease research to compare the time to occurrence of clinical events between treatment or exposure groups (Cole and Hudgens, 2010). Randomized trials are the gold standard to estimate exposure effects on survival time, but are not always ethical or feasible. Although observational studies may provide estimates of effects when trial data are unavailable, the estimates they yield are often riddled with confounding (Greenland and Morgenstern, 2001). Informally, confounding occurs when the exposure and outcome share a common cause. The standard approach in survival analysis to account for multiple measured confounders is the Cox proportional hazards regression model (Cox, 1972).

As an alternative to the standard Cox model, we present a method in this paper that uses inverse probability (IP) weights to estimate the effect of a baseline exposure on survival time. Under certain assumptions, results from an IP-weighted Cox model of observational data can be interpreted similar to a randomized trial with no drop out (i.e., loss to follow-up). In particular, unlike the standard Cox model, this approach allows for estimation of marginal effects which compare the distribution of outcomes when the entire population is exposed versus when the entire population is unexposed (Kaufman, 2010). For example, this IP-weighted approach yields marginal Kaplan-Meier (Kaplan and Meier, 1958) type survival curve estimates that account for confounding by measured covariates (Xie and Liu, 2005; Cole and Hernan, 2004). Informally, each participant is weighted to create a pseudopopulation where (i) exposure is not associated with covariates such that (measured) confounding is eliminated, and (ii) drop out is not associated with exposure or covariates such that selection bias due to drop out is eliminated (Hernan et al., 2000). This approach is akin to

survey sampling weighting used to estimate a quantity in the population (Thompson, 2012; Horvitz and Thompson, 1952). Herein, we refer to IP weighting as standardization, where the standardization is to the entire population under two different exposures (Cole and Hernan, 2004; Sato and Matsuyama, 2003). We illustrate this standardization method through an example that estimates the effect of injection drug use (IDU) on AIDS-free survival among HIV-infected women.

3.2 Motivating Example: AIDS-Free Survival Among Injection Drug Users

The Women's Interagency HIV Study (WIHS) is a prospective, observational, multicenter study of women living with HIV and women at risk for HIV infection in the U.S. (Bacon et al., 2005). A total of 4,129 women (1,065 HIV-uninfected) were enrolled between October 1994 and December 2012 at six U.S. sites. An institutional review board at each site approved study procedures and all study participants provided written informed consent. We were interested in determining if AIDS-free survival among HIV-infected women differed by IDU, accounting for possible confounding by factors measured at baseline and selection bias due to drop out by factors measured during study follow-up. We estimated the hazard ratio and the absolute risk difference at ten years to quantify this effect.

The study sample consisted of 1,164 women enrolled in WIHS who were alive, HIV-infected, and free of AIDS on 6 December 1995 (Lau et al., 2009). The endpoint was either death or a diagnosis of AIDS. Women who did not reach this endpoint by 6 December 2005 were censored at that time or at their last visit where they were known to be alive and AIDS-free, whichever came first. A history of IDU at WIHS enrollment is denoted as $X = 1$ ($X = 0$ otherwise). The baseline covariates African American race, age, and nadir CD4 count (in cells/uL) measured from WIHS enrollment to baseline (i.e., 6 December 1995) are denoted by the vector $\mathbf{Z}$. The time-varying covariate antiretroviral (ART) initiation during study follow-up is denoted by $Z(t)$, where $Z(t) = 1$ if an individual starts ART before time t since baseline and $Z(t) = 0$ otherwise.

3.3 Inverse Probability Weighted Cox Models

Researchers are often interested in estimating effects of an exposure fixed at study entry. IP-weighted Cox models are a method to compare the timing of clinical events under two different exposures. An appealing feature of the IP-weighted Cox model is that the results from this method can be interpreted similar to results from randomized trials with no drop out. An IP-weighted Cox model is fit by maximizing a weighted partial likelihood, where participant i who died or was diagnosed with AIDS at time t from baseline contributes the term

$$\{\exp(\hat{\beta}X_i)/\sum_{j\in R(t)}\left[\hat{w}_j(t)\exp(\hat{\beta}X_j)\right]\}^{\hat{w}_i(t)} \tag{3.1}$$

where $R(t)$ is the risk set at time t and $\exp(\beta)$ is the marginal hazard ratio for a unit difference in exposure X accounting for confounding and selection bias measured by covariates through the estimated IP weight $\hat{w}_i(t)$ (discussed below) (Robins et al., 2000). When the estimated IP weight $\hat{w}_j(t) = 1$ for all $j \in R(t)$, equation (3.1) is the usual contribution to the partial likelihood for the standard (i.e., unweighted) Cox model (See Appendix A). Slight modification of the likelihood is needed in the presence of tied survival times. The robust variance estimator (Lin and Wei, 1989) can be employed to account for the fact that the IP weights are estimated (Cole and Hernan, 2008). See Appendix A for a review of inference for the standard (i.e., unweighted) Cox proportional hazards model.

The estimated IP weight $\hat{w}_i(t)$ is the product of an estimated time-fixed IP exposure weight $\hat{w}_{1i}$ and an estimated time-varying IP drop out weight $\hat{w}_{2i}(t)$ for each participant i at each survival time t. The time-fixed IP exposure weights are constructed to account for confounding by covariates measured at baseline. The IP exposure weights essentially create a pseudopopulation where exposure is not associated with covariates, thus eliminating (measured) confounding. For example, if non-African Americans are more likely to report IDU than African Americans, then an African American in the study who reports IDU will be upweighted because she is representing more participants. Different versions of these

weights have been proposed. It is generally recommended to use the (estimated) stabilized IP exposure weight $\hat{w}_{1i}$ defined as the ratio of the estimated marginal probability of having the exposure that participant i had, formally $P(X_i = x_i)$, to the estimated covariate-conditional probability of having the exposure that participant i had, formally $P(X_i = x_i|\mathbf{Z}_i)$, where $\mathbf{Z}_i$ are the measured covariates for participant i assumed sufficient to adjust for confounding. Details on estimating the IP exposure weights using the observed data are provided in the next section tailored to the example.

The time-varying IP drop out weights $\hat{w}_{2i}(t)$ are constructed to account for possible selection bias due to drop out (Robins et al., 2000). The IP drop out weights essentially create a pseudopopulation as if no participants had dropped out. Participants last observed alive and AIDS-free more than one year prior to 6 December 2005 were considered drop outs. Participants receive a time-varying weight that corresponds to their probability of remaining free from drop out. This stabilized IP weight $\hat{w}_{2i}(t)$ is defined as the ratio of the estimated marginal probability of remaining free of drop out, formally $P(D_i > t|X_i)$, where D_i is the time from baseline to drop out for participant i, to the estimated covariate-conditional probability of remaining free of drop out, formally $P(D_i > t|\mathbf{Z}_i, Z_i(t), X_i)$, where $\mathbf{Z}_i$ and $Z_i(t)$ are the measured common causes of drop out and the study outcome for participant i up to time t. (Note the covariates in the drop out model can be different than the covariates in the exposure weight model). Details on estimating the IP drop out weights using the observed data are provided in the next section tailored to the example.

Standardized survival curve estimates can be obtained by fitting an IP-weighted Cox model stratified by exposure with no covariates and then nonparametrically estimating the baseline survival functions for the two strata (Cole and Hernan, 2004). In the absence of weighting, these survival curve estimates will be (asymptotically) equivalent to Kaplan-Meier estimates obtained separately for each of the exposure stratum (Collett, 2003).

For all Cox models presented below, we employed Efron's method to account for events that occurred on the same date (Efron, 1977). We obtained confidence intervals for the risk difference at 10 years using a nonparametric bootstrap with 200 random samples with

replacement (Efron and Tibshriani, 1994). The data analysis for this paper was conducted using SAS software version 9.3 (SAS Institute Inc., Cary, NC). SAS code for analyses in the present paper is provided in the Supplemental Materials.

3.4 Illustrative Example

The 1,164 women were 58% African American, median age was 36 years, and median nadir CD4 count was 349 cells/μL at baseline (Table 3.1). At enrollment, 38% of women reported a history of IDU. During follow-up, 664 (57%) of women initiated ARTs. Women were followed for up to 10 years with a total of 7,090 person-years during which 579 (50%) developed AIDS or died, and 117 (10%) dropped out of the study. In analyses that did not account for covariates, women with a history of IDU had notably worse AIDS-free survival than women without a history of IDU (Figure 3.1). The estimated hazard ratio from the unadjusted Cox model was 1.72 (95% confidence interval (CI): 1.46, 2.03; Wald P value $<$ 0.001), suggesting that the hazard of AIDS or death for those with a history of IDU was almost twice the hazard of those without a history of IDU (Table 3.2). We assessed the proportional hazards assumption graphically by examining whether the log cumulative hazard function estimates (See Figure 3.2) were approximately parallel. We also assessed this assumption statistically by inclusion of a product term between history of IDU and time in the Cox model, for which the Wald P value was 0.40. Neither graphical nor statistical assessment suggested a meaningful departure from proportional hazards.

We then obtained a standardized hazard ratio estimate from the IP-weighted Cox model, which involved two steps. In the first step, using separate logistic regression models, weights were estimated for the probability of exposure (i.e., history of IDU) and for the probability of not dropping out. For the exposure weights, we fit logistic regression models for both the numerator and denominator. The exposure model for the numerator had no covariates, while the exposure model for the denominator included age at baseline, race, and nadir CD4 count, as well as all pairwise interactions. Age and nadir CD4 were included as continuous variables using restricted quadratic splines with four knots placed at 5^{th}, 35^{th}, 65^{th}, and 95^{th} percentiles

(Howe et al., 2011b). For the drop out weights, time was coarsened into months since baseline (Hernan et al., 2001). Then, using pooled logistic regression (D'Agostino et al., 1990), the drop out model for the numerator included only exposure (i.e., history of IDU) and time (using restricted quadratic splines), while the drop out model for the denominator included exposure, time (spline), age (spline), race, nadir CD4 count (spline), and ART initiation (time-varying), as well as all pairwise interactions. In the pooled logistic regression model, each person contributed up to 120 records and the weights were cumulatively multiplied for each person. The estimated weights $\hat{w}_i(t)$ had a mean of 1.01 (with a standard deviation of 0.76), and ranged from 0.43 to 12.43 (see Table 3.3). In the second step, the IP-weighted Cox model was fit by weighting participants by their estimated weights, with outcome time to AIDS or death, and history of IDU as the sole covariate.

We obtained the estimated survival functions from an IP-weighted Cox model with no covariates stratified by history of IDU. After standardization for confounding and drop out by IP weighting, survival curves showed an attenuated difference in AIDS-free survival compared to the survival curves without accounting for any covariates (Figure 3.1). Under certain assumptions discussed below, the dashed black curve can be interpreted as an estimate of the AIDS-free survival if (contrary to fact) everyone had a history of IDU at enrollment and did not drop out, while the solid black curve can be interpreted as an estimate of the AIDS-free survival if (contrary to fact) no one had a history of IDU at enrollment and everyone did not drop out (6, 7). The standardized hazard ratio from the IP-weighted Cox model was 1.53 (95% CI: 1.26, 1.85; Wald P value < 0.001) (Table 3.2). We again assessed the proportional hazards assumption graphically by examining whether the IP-weighted log cumulative hazard function estimates (See Figure 3.3) were approximately parallel. We also assessed this assumption statistically by inclusion of a product term between history of IDU and time, for which the Wald P value was 0.18. Neither graphical nor statistical assessment suggested a meaningful departure from proportional hazards. From the standardized survival curves, the ten-year risk of AIDS or death was 0.59 if (contrary to fact) everyone had a history of IDU at enrollment and 0.46 if (contrary to fact) no one had a history of IDU at enrollment. The 10-year risk difference was 0.14 (bootstrap 95% CI: 0.06, 0.22). For comparison, we also estimated a

covariate-adjusted hazard ratio by including history of IDU, age (spline), race, and nadir CD4 count (spline) directly in an unweighted Cox model. The covariate-adjusted hazard ratio estimate was 1.62 (95% CI: 1.35, 1.95; Wald P value < 0.001).

3.5 Discussion

IP-weighted Cox models and standardized survival curves were presented as methods to compare the timing of clinical events for two different exposure conditions under certain assumptions. We compare this method to the traditional Cox model and discuss assumptions and caveats below.

Although hazard ratio estimates from the IP-weighted and covariate-adjusted Cox model were comparable in the WIHS example above, the standardized (i.e., IP-weighted) method provides several potential benefits over the covariate-adjusted Cox model. First, the results from the standardized approach may be interpreted similar to results from a randomized trial with no drop out when only observational data is available (under certain assumptions discussed below). In particular, the estimated hazard ratio using the standardized approach can be interpreted the same as the (marginal) hazard ratio one would obtain in a randomized experiment such as a clinical trial where there is no confounding and no drop out. In contrast, a covariate-adjusted Cox model hazard ratio does not necessarily equal the marginal hazard ratio because (even in the absence of unmeasured confounding) the Cox model is not collapsible for the hazard ratio parameter (Greenland, 1996). A regression model is said to be collapsible for a parameter (in this case, the hazard ratio) if the covariate-adjusted parameter is the same as the unadjusted parameter (Greenland et al., 1999b).

Second, the IP weighting approach yields standardized survival curve estimates. Although the hazard ratio is a common summary parameter to compare survival distributions between exposure groups, there are drawbacks to focusing inference on hazard ratios. For instance, the hazard ratio can be difficult to interpret, especially when trying to summarize the effect of a treatment or exposure (Hernan, 2010). Presenting estimated survival curves is an alternative to reporting hazard ratios that may be more interpretable because survival curves summarize

all information from baseline up to any time t. The IP-weighted approach leads to Kaplan-Meier type survival curve estimates that are standardized to the entire population under two different exposures at baseline while accounting for confounding by multiple covariates. A covariate-adjusted Cox model does not afford such survival curve estimates (Kaufman, 2010; Xie and Liu, 2005).

Third, the IP-weighted approach with drop out weights requires a weaker assumption about censoring than the covariate-adjusted Cox model (Hernan et al., 2000; Howe et al., 2011a; Kalbfleisch and Prentice, 2002). The adjusted Cox model assumes that the censoring hazard is independent of survival time conditional on being at risk, exposure, and baseline covariates, whereas the IP-weighted Cox model makes the weaker assumption that censoring is independent conditional on being at risk, exposure, baseline covariates, and time-varying covariates (Kalbfleisch and Prentice, 2002; Robins and Finkelstein, 2000). Specifically, if there are measured time-varying covariates predictive of censoring and survival time, the IP-weighted approach will yield consistent estimates of the marginal hazard ratio, while the covariate-adjusted Cox model estimator will not be consistent for the marginal or conditional hazard ratio (Hernan et al., 2000; Robins et al., 2000; Robins and Finkelstein, 2000).

Results using standardization by IP weights also have, in general, a different interpretation than results from an unadjusted Cox model. In particular, when exposure is confounded, the parameter of an unadjusted Cox model is a measure of association and will generally differ from the parameter of an IP-weighted Cox model (i.e., the marginal hazard ratio), which is a measure of effect (Robins et al., 2000). On the other hand, when exposure is unconfounded (e.g., as in randomized trials), the target parameter of both models is the marginal effect. In this case, drop out weights might still be employed to account for selection bias due to loss to follow-up (Hernan et al., 2013). Moreover, the use of IP drop out weights yields estimators that are more efficient (i.e., less variable) than those from an unadjusted Cox model even when there is no selection bias (Cole and Hernan, 2004; Robins and Finkelstein, 2000).

Estimation of the hazard ratio and survival curves using standardization by IP weights requires certain assumptions to yield valid inference about the exposure effect. In particular,

this approach assumes positivity, well-defined exposures, correctly specified models, and no unmeasured confounding or selection bias. For each level defined by the covariates, positivity means that there is a positive probability of each level of exposure (Cole and Hernan, 2008). For example, positivity assumes African American women could possibly have either a history of IDU or no history of IDU (and similarly for non-African American women). On the other hand, if African American women could never have a history of IDU, the positivity assumption would be violated. Well-defined exposures imply that there are not multiple versions of exposure, or if there are, that they are unimportant (Cole and Frangakis, 2009; Pearl, 2010; VanderWeele, 2009a). For instance, the duration of exposure to IDU in the example is assumed to be irrelevant in the sense that an individual's time until AIDS or death is assumed to be the same regardless of exposure duration. Alternatively, the marginal effects being estimated can be viewed as average effects over the distribution of IDU exposure. The standardized hazard ratio estimator and survival curves require correctly specified IP weights (i.e., correct covariate functional forms). It is also assumed that sufficient sets of covariates have been measured to effectively address confounding (i.e., no unmeasured confounding) (Hernan et al., 2000; Robins et al., 2000) and selection bias due to drop out (Robins and Finkelstein, 2000). In the example, age, race and nadir CD4 were assumed to be sufficient to account for confounding and these baseline covariates, time-varying ART initiation, and exposure were assumed to be sufficient to account for selection bias due to drop out.

Typically, when assessing the effect of a baseline exposure, one would not adjust for post-baseline covariates in order to avoid potential selection bias (Cole and Hernan, 2002; Pearl, 2001). For example, post-baseline covariates might be on the causal pathway from the exposure to the outcome and adjusting for such covariates may lead to attenuated estimates of the total effect of the exposure (Kalbfleisch and Prentice, 2002). In the example, the time-varying covariate ART initiation was not included in the covariate-adjusted Cox model. On the other hand, time-varying ART initiation may be predictive of both drop out and the survival time, so excluding that variable from the Cox model has the potential to introduce selection bias. In contrast, the use of IP drop out weights provides a valid approach to adjusting for a time-varying covariate associated with drop out and survival (Hernan et al.,

2000, 2001).

We only discussed exposure groups defined at baseline. When interest focuses on exposures that change over time, methods must be adapted accordingly. When a time-varying covariate is a risk factor for the outcome, predicts later exposure, and is affected by prior exposure, standard statistical methods (e.g., Cox models with time-varying covariates) are biased and fail to provide consistent estimators of effects (Hernan et al., 2001; Cole et al., 2003; Robins, 2000). IP weighting can be used to fit marginal structural Cox models of time-varying exposures in the presence of such time-varying confounders (Robins et al., 2000). For example, in HIV-infected individuals, CD4 count is a risk factor for death, predicts subsequent treatment with ART, and is affected by prior treatment; thus, the marginal structural Cox model is appropriate for assessing the effect of time-varying ART on overall survival while adjusting for time-varying CD4 count.

In the illustrative example, we estimated the total effect of IDU history on time to AIDS or death, which included the indirect effect mediated through ART and the direct effect not mediated through ART. Estimating the direct and indirect effects of IDU separately may be of interest and can be obtained by fitting marginal structural models using IP weights as long as all relevant data is available for these models (VanderWeele, 2009b).

We suggest using expert knowledge to determine which covariates to adjust for prior to model fitting. Many epidemiologists would retain a possible confounder if its inclusion changes the estimate of association by more than 10% or 20% and a great deal of precision is not sacrificed (Mickey and Greenland, 1989). Other approaches for determining which covariates to adjust for in a model include conditioning on (i) all causes of the exposure or outcome (VanderWeele and Shpitser, 2011) or (ii) a sufficient set of covariates based on a causal directed acyclic graph (Greenland et al., 1999a) informed by a priori beliefs or knowledge (Brookhart et al., 2006). For the weight models, inclusion of covariates that are unrelated to the exposure but related to the outcome may yield effect estimates with smaller variance and no increase in bias, so they should be included in the model; however, inclusion of covariates that are related to the exposure but not to the outcome may lead to effect

estimates with larger variance and no reduction in bias, so they should be excluded from the model (Brookhart et al., 2006). Machine learning techniques (Lee et al., 2010, 2011) can be used as an alternative approach to logistic regression for estimating weights.

Although the IP-weighted method used to analyze the WIHS data attempts to adjust for confounding and selection bias, the conclusions from the analysis are still subject to the following considerations. Comparisons of groups from observational studies may be susceptible to unmeasured confounding bias, as the assumption of no unmeasured confounding is untestable. Similarly, the IP-weighted method assumes drop out is independent of the survival time conditional on observed baseline and time-varying covariates. The absence of unmeasured covariates predictive of both censoring and survival times is also an untestable assumption. Even in the absence of unmeasured covariates, IP drop out weights could fail to correct for selection bias if there are not a sufficient number of participants during follow-up (Howe et al., 2011a). The models for the IP weights need to be correctly specified and sensitivity analysis should be performed to assess robustness of the effect estimates to model misspecification (Cole and Hernan, 2008). When there are longer follow-up periods (specifically, a large number of participant assessments) or near positivity violations, weights can become large, leading to imprecise effect estimates. Truncating estimated weights offers some solution to this problem, although results can be sensitive to choice of truncation cut-off points (Cole and Hernan, 2008; Kish, 1992). Finally, as with all methods, error in the measurement of exposure, covariates, or the event status or times could bias the results (Hernan and Cole, 2009).

In conclusion, we have presented an example of survival data pertinent to infectious disease research and illustrated how to compare groups of study participants using the IP-weighted Cox proportional hazards model. The methods presented here have broad applicability in infectious disease research. Careful use of this and other methods for survival analysis will continue to enrich the evidence base in the field of infectious diseases by providing answers to questions that are difficult or impossible to answer well without explicitly accounting for time. Inverse probability weighted Cox models provide a method to estimate covariate-standardized

hazard ratios and survival curves in observational studies, and obtain information about effects of treatments or exposures to prevent infectious diseases or their sequela.

Acknowledgements

These findings are presented on behalf of the Women's Interagency HIV Study (WIHS). We would like to thank all of the WIHS investigators, data management teams, and participants who contributed to this project. Funding for this study was provided by the following National Institutes of Health (NIH) Grants: R01AI100654, R01AI085073, U01AI069918, R56AI102622, 5 K24HD059358-04, 5 U01AI103390-02 (WIHS), and P30AI50410 (CFAR). Data in this manuscript were collected by the WIHS. The contents of this publication are solely the responsibility of the authors and do not represent the official views of the NIH.

B.L. acquired the WIHS data used in the example. A.L.B., M.G.H., and S.R.C. wrote the initial draft of the manuscript. A.L.B. performed the analyses under the guidance of M.G.H. and S.R.C. A.L.B., M.G.H., S.R.C., B.L., and A.A.A. participated in discussions on technical points. A.A.A. provided guidance on the data example and ensured that the level of technicality was appropriate for the readership. All authors were involved in the review and editing process of the final manuscript. The authors gratefully acknowledge the comments of the two anonymous reviewers, which greatly improved the article.

Table 3.1: Characteristics of 1,164 HIV-infected women in the Women's Interagency HIV Study December 6, 1995 through December 6, 2005

Characteristics[a]	**History of Injection Drug Use (IDU)** $n = 439$	**No History of Injection Drug Use (IDU)** $n = 725$	**Overall** $n = 1,164$
Age (years)	40 (35, 44)	33 (29, 39)	36 (31, 41)
African American race	273 (62%)	399 (55%)	672 (58%)
Nadir CD4+ count (cells/uL)	352 (208, 522)	348 (216, 505)	349 (213, 517)
Initiated antiretrovirals (ARTs)[b]	208 (47%)	456 (63%)	664 (57%)

[a] Median (interquartile range) or number (percent)

[b] During follow-up

Table 3.2: Association of history of injection drug use with time to AIDS or death for 1,164 HIV-infected women in the Women's Interagency HIV Study December 6, 1995 through December 6, 2005

	History of Injection Drug Use (IDU) $n = 439$	**No History of Injection Drug Use (IDU)** $n = 725$	**Overall** $n = 1,164$
Unadjusted			
AIDS cases and deaths	272 (62%)	307 (42%)	579 (50%)
Person-years	2,368	4,721	7,090
Hazard ratio (95% CI)	1.72 (1.46, 2.03)	1	–
10-year risk (95% CI)	0.64 (0.59, 0.68)	0.46 (0.42, 0.49)	0.53 (0.50, 0.56)
10-year risk difference (95% CI)	0.18 (0.13, 0.24)	0	–
Standardized[a]			
AIDS cases and deaths	248.49 (58%)	308.18 (43%)	556.67 (48%)
Person-years	3,730.97	7,582.69	11,313.66
Hazard ratio (95% CI)	1.53 (1.26, 1.85)	1	–
10-year risk (95% CI)	0.59 (0.54, 0.64)	0.46 (0.42, 0.49)	0.51 (0.47, 0.54)
10-year risk difference (95% CI)	0.14 (0.06, 0.22)	0	–

[a] IP weighted to account for confounding of exposure due to baseline covariates (age (spline), race, and nadir CD4 (spline)) and selection bias due to loss to follow-up (covariates included exposure, time-varying ART initiation, and baseline covariates).

Table 3.3: Example individual-level estimated exposure weights, drop out weights, and combined weights for 1,164 HIV-infected women in the Women's Interagency HIV Study December 6, 1995 through December 6, 2005

ID	Time in	Time out	History of Injection Drug Use (IDU)	Event	Drop out	Exposure weight	Drop out weight	Combined weight
34	0.00	0.08	Yes	No	No	0.499	1.000	0.499
34	0.08	0.17	Yes	No	No	0.499	1.001	0.499
34	0.17	0.21	Yes	No	Yes	0.499	1.001	0.500
36	0.00	0.08	No	No	No	1.135	1.000	1.135
36	0.08	0.17	No	No	No	1.135	1.001	1.136
36	0.17	0.22	No	No	Yes	1.135	1.001	1.136
37	0.00	0.08	Yes	No	No	1.061	1.000	1.061
37	0.08	0.17	Yes	No	No	1.061	1.000	1.060
37	0.17	0.23	Yes	Yes	No	1.061	1.000	1.060
38	0.00	0.08	No	No	No	0.961	1.000	0.961
38	0.08	0.17	No	No	No	0.961	0.999	0.961
38	0.17	0.25	No	No	No	0.961	0.999	0.960
38	0.25	0.33	No	No	No	0.961	0.998	0.959
38	0.33	0.42	No	No	No	0.961	0.997	0.958
38	0.42	0.43	No	Yes	No	0.961	0.996	0.957
66	0.00	0.08	Yes	No	No	0.558	1.000	0.558
66	0.08	0.17	Yes	No	No	0.558	0.999	0.558
66	0.17	0.25	Yes	No	No	0.558	0.999	0.558
66	0.25	0.33	Yes	No	No	0.558	0.998	0.557
66	0.33	0.38	Yes	No	Yes	0.558	0.998	0.557

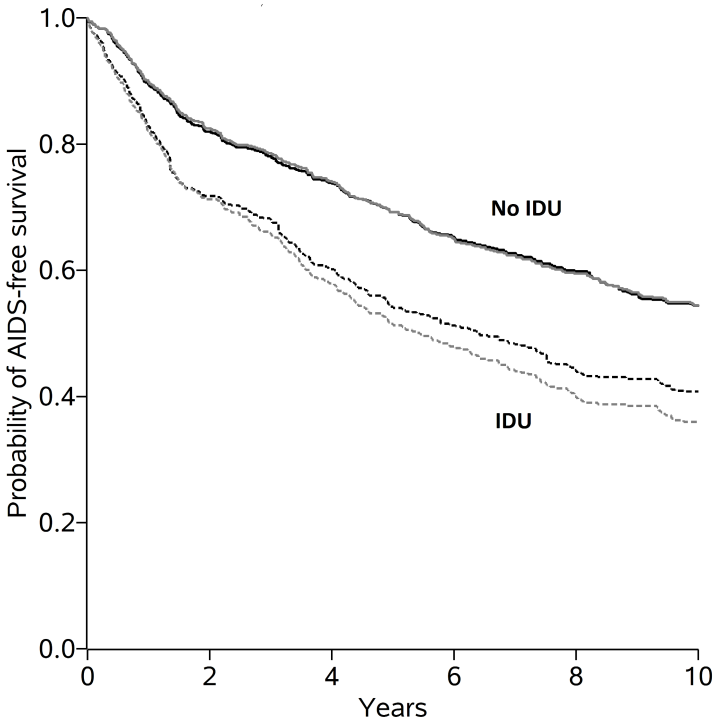


Figure 3.1: Kaplan-Meier estimated AIDS-free survival curves without accounting for any covariates (gray curves) and standardized estimated AIDS-free survival curves (accounting for age, race, nadir CD4, and antiretroviral therapy (ART) initiation) (black curves) for 1,164 HIV-infected women with and without a history of injection drug use (IDU) in the Women's Interagency HIV Study December 6, 1995 through December 6, 2005

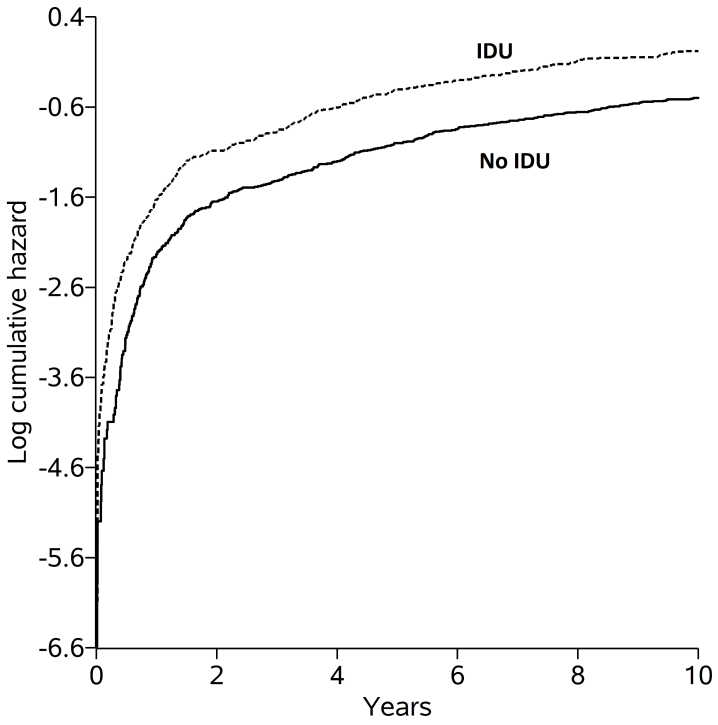


Figure 3.2: Estimated log cumulative hazard curves without accounting for any covariates calculated for 1,164 HIV-infected women with and without a history of injection drug use (IDU) in the Women's Interagency HIV Study December 6, 1995 through December 6, 2005

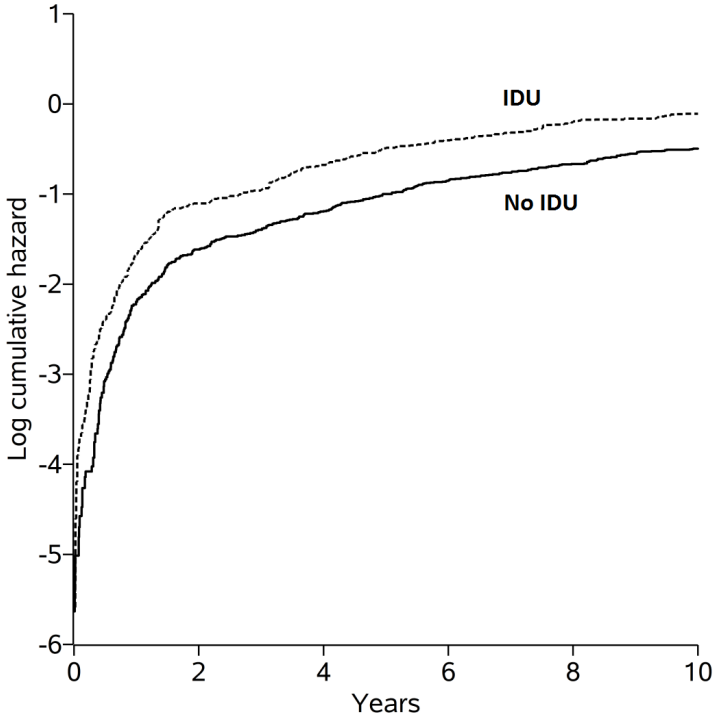


Figure 3.3: Standardized estimates of the log cumulative hazard curves (accounting for age, race, nadir CD4, and antiretroviral therapy (ART) initiation) calculated for 1,164 HIV-infected women with and without a history of injection drug use (IDU) in the Women's Interagency HIV Study December 6, 1995 through December 6, 2005

CHAPTER 4: GENERALIZING EVIDENCE FROM RANDOMZIED TRIALS USING INVERSE PROBABILITY OF SAMPLING WEIGHTS

4.1 Introduction

Generalizability is a concern for many studies in public health. Generalizability is defined as the degree to which an internally valid measure of effect estimated in a sample from one population would change if the study had been conducted in a different target population. For example, in trials of treatment for HIV-infected individuals, there is often concern that trial participants are not representative of the larger population of HIV-positive individuals. One study highlighted the overrepresentation of African American and Hispanic women among HIV cases in the U.S. and the limited clinical trial participation of members of these groups (Greenblatt, 2011). Another study reviewed eligibility criteria of 20 AIDS Clinical Trial Group (ACTG) studies and found that 28% to 68% of the Women's Interagency HIV Study (WIHS) cohort would have been excluded (Gandhi et al., 2005).

There are several quantitative methods that employ sampling scores to assess generalizability. The sampling score is defined as the probability of participation in the trial conditional on covariates. These approaches are akin to methods that use treatment propensity scores to adjust for (measured) confounding (Rubin, 1980) and include the use of inverse probability of sampling weights and stratification based on sampling scores. In Cole and Stuart (2010), sampling scores were estimated using logistic regression. An inverse-probability-of-sampling-weighted Cox proportional hazards model was fit to obtain a hazard ratio and estimated survival curves. A robust estimate of the variance was employed (Robins, 1998); however, no closed-form expression for the variance was provided. To date, there is no formal justification of the large sample statistical properties of these estimators (i.e., statistical consistency and asymptotic normality). As an alternative, a sampling score stratified estimator was proposed

to generalize trial results (Tipton, 2013; O'Muircheartaigh and Hedges, 2013; Tipton et al., 2014).

Following Cole and Stuart (2010) and Stuart et al. (2011), we consider an inverse weighting approach based on sampling scores to generalize trial effects for continuous outcomes to a target population and comparisons are made to the stratified estimator. In Section 4.2, the assumptions and notation for this method are discussed. The inverse probability of sampling weighted (IPSW) estimator and the stratified estimator are described in Section 4.3. In Section 4.4, large sample properties of the IPSW estimator are derived, including a closed form expression for the asymptotic variance and a consistent sandwich-type estimator of the variance. The finite sample performance of the IPSW and stratified estimators are compared using simulations in Section 4.5. In the Section 4.6, the IPSW estimator is applied to generalize results from the ACTG to all people currently living with HIV in the U.S. Section 7 concludes with a discussion.

4.2 Notation and Assumptions

Consider a setting where two data sources are available. A random sample (e.g., cohort study) of size m is drawn from the near infinite target population and assumed to be representative. A second sample of n individuals participate in a randomized trial, and the treatment assignment mechanism is known to the analyst. The trial is possibly a non-random (i.e., biased) sample from the near infinite target population. In addition, the treatment effect (measured as a difference in means) is possibly modified by the same covariates that differ between the trial and the near infinite target. A covariate is an effect modifier when the average causal effect of the treatment on the outcome varies across levels of the covariate. In general, let upper case letters denote random variables and lower case letters denote realizations of those random variables. Define $\mathbf{Z}_{(m+n)\times p}$ as a vector of fixed characteristics and assume that information on $\mathbf{Z}$ is available for those in the trial and those in the cohort. Let $S = 1$ denote trial participation. For those in the trial, define X as the treatment indicator, where $X = 1$ if assigned to treatment. Let $i = 1, \ldots, n + m$ index the trial and cohort participants.

Each individual has a vector (Y^0, Y^1) of potential outcomes in the target population. Y^0 is the value of the response that would have been seen if (possibly contrary to fact) the participant were randomized to control, and Y^1 is the value of the response that would have been seen if (possibly contrary to fact) the participant were randomized to treatment. It is assumed throughout that the stable unit treatment value assumption (SUTVA) (Rubin, 1980; Tipton, 2013) holds, i.e., there are no variations of treatment and there is no interference between participants. The observed response Y is the response that would have been seen under the treatment actually assigned in the trial (i.e., one of the two potential outcomes), defined as $Y = Y^1X + Y^0(1 - X)$. Assume $(S, \mathbf{Z})$ are observed for cohort participants and $(S, \mathbf{Z}, X, Y)$ are observed for trial participants. Let $\mu_1 = E\left(Y^1\right)$ and $\mu_0 = E\left(Y^0\right)$, where the expectations are with respect to the potential outcomes in the near infinite target population. The population average treatment effect (PATE) is $\Delta = \mu_1 - \mu_0$.

Once in the trial, participants are randomly assigned to a treatment group, so that individuals from either group are balanced in that the distribution of $\mathbf{Z}$ is the same regardless of treatment assignment conditional on trial participation. This is ignorable treatment assignment mechanism gained through randomization $P(X = x|S = 1, \mathbf{Z}, Y^0, Y^1) = P(X = x|S = 1)$. Assume an ignorable trial participation mechanism conditional on $\mathbf{Z}$, so $P(S = s|\mathbf{Z}, Y^0, Y^1) = P(S = s|\mathbf{Z})$. In other words, participants in the trial are no different from nonparticipants in regards to the treatment-outcome relationship conditional on $\mathbf{Z}$. Measurement of all treatment effect modifiers associated with trial participation will be sufficient to assume an ignorable trial participation mechanism. Trial participation and treatment positivity are also assumed, so $P(X = x|\mathbf{Z}, S = 1) > 0$ and $P(S = s|\mathbf{Z}) > 0$ for all $\mathbf{Z} = \mathbf{z}$. Assume that participants in the trial are adherent to their treatment assignment (i.e., ignoring noncompliance issues) and the model for the sampling scores is correctly specified (i.e., correct covariate functional forms).

4.3 Estimators of the Population Average Treatment Effect

A traditional (i.e., unweighted) approach to estimating treatment effects is a difference in means. The within-trial intention-to-treat (ITT) estimator is defined as

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

where here and in the sequel $\sum_i = \sum_{i=1}^{n+m}$.

Two estimators that employ sampling scores are considered. In practice, the sampling scores are likely unknown and can be estimated using a parametric model. Following Cole and Stuart (2010), the sampling scores $P(S = 1|\mathbf{Z} = \mathbf{z})$ are estimated using logistic regression. Let $w(\mathbf{Z}, \boldsymbol{\beta}) = w = P(S = 1|\mathbf{Z})$, $w_i = w(\mathbf{Z_i}, \boldsymbol{\beta})$, and $\hat{w}_i = w(\mathbf{Z_i}, \hat{\boldsymbol{\beta}})$. Let $\boldsymbol{\beta}_0$ be the vector of true values of $\boldsymbol{\beta}_{1\times p}$ and $\hat{\boldsymbol{\beta}}_{1\times p}$ be the vector of estimators of $\boldsymbol{\beta}_{1\times p}$. To account for the random sampling of the cohort from the near infinite target population when estimating $\boldsymbol{\beta}$, each participant in the cohort is inverse weighted by the sampling fraction $r_i = m/(N - n)$, where N is the size of the near infinite target population with $N >> n$ and $N >> m$. When estimating $\boldsymbol{\beta}$, each trial participant is given a weight of $r_i = 1$.

The IPSW estimator of the PATE is

$$\hat{\Delta}_{IPW} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{\sum_i S_i Y_i X_i / \hat{w}_i}{\sum_i S_i X_i / \hat{w}_i} - \frac{\sum_i S_i Y_i (1 - X_i) / \hat{w}_i}{\sum_i S_i (1 - X_i) / \hat{w}_i} \tag{4.1}$$

An alternative approach for estimating the PATE uses stratification based on the sampling scores (Tipton, 2013; O'Muircheartaigh and Hedges, 2013; Tipton et al., 2014). This estimator is computed in the following steps. First, $\boldsymbol{\beta}_0$ is estimated using a (weighted) logistic regression model and the sampling scores $\hat{w}_i$ are computed. These sampling scores are used to form L strata according to the quintiles of the distribution in the target population. Because we assume the trial and cohort both arise from the same near infinite target population, the distribution of sampling scores in the combined trial and cohort are used to estimate the quintiles (Tipton, 2013). The difference of sample means within each stratum is computed

among those in the trial. Lastly, the PATE is estimated as a weighted sum of the differences of sample means across strata, where the weight $\hat{w}_{p_l}$ is the proportion of observations in stratum l in the target population. Let S_{il} denote trial participation for participant i in stratum l for $i = 1, \ldots, (n+m)$ and $l = 1, \ldots, L$ (and $S_{il} = 0$ otherwise). Let X_{il} and Y_{il} denote treatment assignment and outcome in the trial, respectively, for participant i in stratum l for $i = 1, \ldots, (n+m)$ and $l = 1, \ldots, L$ (and $X_{il} = 0$, $Y_{il} = 0$ otherwise). The sampling score stratified estimator is defined as

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

where the L stratum are defined by the distribution of the sampling scores in the near infinite target population, $l = 1, \ldots, L$ and $i = 1, \ldots, (n+m)$ and $\hat{w}_{p_l} = N_l/N$ with N_l as number in stratum l in the target and N is the size of the near infinite target.

4.4 Large Sample Properties of the Inverse Probability of Sampling Weighted Estimator

Let Δ_0 be the true value of Δ. Let $w_0 = w(\mathbf{Z}_i, \boldsymbol{\beta}_0)$ be the true weight. Using the fact that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ and $w(\mathbf{Z}_i, \hat{\boldsymbol{\beta}}) \xrightarrow{p} w(\mathbf{Z}_i, \boldsymbol{\beta}_0)$ as $n, m \to \infty$ with $n < m$ and $n/(n+m) \to c$ with $0 < c \leq 1$,

$$\frac{\sum_i S_i Y_i X_i / \hat{w}_i}{\sum_i S_i X_i / \hat{w}_i} = \frac{n^{-1} \sum_i Y_i X_i / \hat{w}_i}{n^{-1} \sum_i X_i / \hat{w}_i} \xrightarrow{p} \frac{E\left(YX/w_0\right)}{E\left(X/w_0\right)} = E\left(Y|X=1\right) = E\left(Y^1\right)$$

where the last step follows from (counterfactual) consistency. Similarly,

$$\frac{\sum_i S_i Y_i (1 - X_i) / \hat{w}_i}{\sum_i S_i (1 - X_i) / \hat{w}_i} \xrightarrow{p} E\left(Y^0\right)$$

Thus, $\hat{\Delta}_{IPW}$ is a consistent estimator of Δ_0.

The distribution of $\mathbf{Z}$ can differ between the trial and cohort participants. The observed data $(S_i, \mathbf{Z}_i)$ for $i = 1, \ldots, n+m$ is an independent, but not necessarily identically dis-

tributed sample. We express the IPSW estimator in terms of estimating equations and appeal to the theory of M-estimation in the Appendix A.6 of Carroll et al. (2010) to demonstrate that the estimator is asymptotically normal and provide a consistent sandwich-type estimator of the variance (Stefanski and Boos, 2002). The theory of M-estimation implies that $(n+m)^{1/2}(\hat{\Delta}_{IPW} - \Delta_0)$ converges in distribution to $N(0, \Sigma_{IPW})$ (Carroll et al., 2010). Additionally, the sandwich-type estimator of the variance $\hat{\Sigma}_{IPW}$ is consistent for Σ_{IPW}, under the suitable regularity conditions as $n, m \to \infty$ with $n < m$ and $n/(n+m) \to c$ with $0 < c \leq 1$.

First, consider the case when $\boldsymbol{\beta}_{1\times p}$ is known, so the solution does not require a score equation for the sampling score model. Let $\hat{\boldsymbol{\theta}}^* = (\hat{\mu}_1, \hat{\mu}_0)$ and $\boldsymbol{\theta}_0^* = (\mu_1, \mu_0)$. The estimating equations for $\hat{\boldsymbol{\theta}}^*$ are

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

By equation (3) of Stefanski and Boos (2002), since the expectation of $\hat{\Delta}_{IPW}$ is zero at the true value Δ_0, the estimator converges in probability to the true value. Thus, $\hat{\Delta}_{IPW}$ is a consistent estimator of Δ_0, which was also demonstrated above. Define the following matrices: $\mathbf{A}(\boldsymbol{\theta}_0^*) = (n+m)^{-1} \sum_i E\left[\frac{\partial}{\partial \boldsymbol{\theta}_0^*} \Psi_\Delta^*(Y_i, \mathbf{Z}_i, X_i, S_i, \theta_0^*)\right]$ and $\mathbf{B}(\boldsymbol{\theta}_0^*) = (n+m)^{-1} \sum_i E\{\mathrm{cov}\left[\Psi_\Delta^*(Y_i, \mathbf{Z}_i, X_i, S_i, \theta_0^*)\right]\}$. $\hat{\boldsymbol{\theta}}^*$ is asymptotically normally distributed with mean $\boldsymbol{\theta}_0^*$ and covariance matrix $\Sigma_\theta^* = (n+m)^{-1} \mathbf{A}^{-1}(\boldsymbol{\theta}_0^*) \mathbf{B}(\boldsymbol{\theta}_0^*) \mathbf{A}^{-T}(\boldsymbol{\theta}_0^*)$. When $\boldsymbol{\beta}_{1\times p}$ is known, the large sample variance of $\hat{\Delta}_{IPW}$ is

$$\Sigma_{IPW}^* = \lim_{(n+m)\to\infty} \left(\Sigma_\theta^{*(11)} + \Sigma_\theta^{*(22)} - 2 \times \Sigma_\theta^{*(12)} \right) \tag{4.2}$$

In the more likely case that $\boldsymbol{\beta}_{1\times p}$ is not known, an additional estimating equation for each element of $\boldsymbol{\beta}$ is needed. Using M-estimation, this suggests that the estimating equation based on the score function of the logistic regression model can be used to obtain the consistent sandwich-type estimator of the variance (Carroll et al., 2010; Stefanski and Boos, 2002). The

vector of parameters $\boldsymbol{\beta}_{1\times p}$ can be consistently estimated by solving the estimating equations

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

(Manski and Lerman, 1977; Scott and Wild, 1986, 2002). The estimating equations for μ_1, μ_0, and $\boldsymbol{\beta}$ are

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

Let $\hat{\boldsymbol{\theta}} = (\hat{\mu}_1, \hat{\mu}_0, \hat{\boldsymbol{\beta}})$ and $\boldsymbol{\theta}_0 = (\mu_1, \mu_0, \boldsymbol{\beta}_0)$. Define the following matrices:
$\mathbf{A}(\boldsymbol{\theta}_0) = (n+m)^{-1} \sum_i E\left[\frac{\partial}{\partial \boldsymbol{\theta}_0} \Psi_\Delta(Y_i, \mathbf{Z}_i, X_i, S_i, \Delta)\right]$ and
$\mathbf{B}(\boldsymbol{\theta}_0) = (n+m)^{-1} \sum_i E\{\mathrm{cov}[\Psi_\Delta(Y_i, \mathbf{Z}_i, X_i, S_i, \Delta)]\}$. $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed with mean $\boldsymbol{\theta}_0$ and covariance matrix $\Sigma_\theta = (n+m)^{-1} \mathbf{A}^{-1}(\boldsymbol{\theta}_0) \mathbf{B}(\boldsymbol{\theta}_0) \mathbf{A}^{-T}(\boldsymbol{\theta}_0)$. When $\boldsymbol{\beta}$ is not known, the large sample variance of $\hat{\Delta}_{IPW}$ is

$$\Sigma_{IPW} = \lim_{(n+m)\to\infty} \left(\Sigma_\theta^{(11)} + \Sigma_\theta^{(22)} - 2 \times \Sigma_\theta^{(12)}\right) \tag{4.3}$$

By comparison of equations (4.2) and (4.3), it can be shown that the variance is smaller when the sampling scores are estimated because $\Sigma_\theta^{(12)}$ is positive definite and larger than $\Sigma_\theta^{*(12)}$. This is analogous to a well-known result for inverse probability of treatment weighted estimators (Hirano et al., 2003; Robins et al., 1992; Wooldridge, 2007). Even if the correct sampling scores are known, estimation of the sampling scores is preferable due to improved efficiency. It is common practice to compute the variance using standard software assuming the weights are known. This leads to valid, but conservative confidence intervals. The consistent sandwich-type estimators of the variance of $\hat{\Delta}_{IPW}$ are provided in Appendix B. In the Supplemental Materials, an R function is provided to compute the IPSW estimator and its corresponding sandwich estimator of the variance.

In practice, it is routine to approximate the sampling variance of $\hat{\Delta}_S$ by treating the esti-

mator as the average of L independent, within-stratum, treatment effect estimators (Tipton, 2013; Lunceford and Davidian, 2004). Define the quintiles of $\hat{w}_i$, where the l^{th} sample quintile is $\hat{q}_l$, $l = 1, \ldots, L$, such that the proportion of $\hat{w}_i \leq \hat{q}_l$ is roughly l/L in the near infinite target. Since we assume the trial and cohort both arise from the same target, the distribution of sampling scores in the combined trial and target are used to estimate the quintiles (Tipton, 2013). In practice, the cohort data will need to be weighted to get the correct distribution of the sampling scores in the target. Let $\hat{q}_0 = 0$ and $\hat{q}_L = 1$. Define $\hat{Q}_l = (\hat{q}_{l-1}, \hat{q}_l)$. Let $N_l = \sum_{i=1}^{N} I(\hat{w}_i \in \hat{Q}_l)$ be the number of individuals in stratum l in the target. Let $n_l = \sum_{i=1}^{n+m} S_i I(\hat{w}_i \in \hat{Q}_l)$ be the number of individuals in stratum l who are selected into the trial. Let $n_{1l} = \sum_{i=1}^{n+m} S_i X_i I(\hat{w}_i \in \hat{Q}_l)$ be the number of individuals in stratum l who are selected into the trial and randomized to treatment. The (approximate) sampling variance of $\hat{\Delta}_S$ is

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

assuming an equal number of participants in each stratum, where $\hat{\sigma}_l^2 = n_{1l}^{-1} s_{1l}^2 + (n_l - n_{1l})^{-1} s_{0l}^2$, $s_{1l}^2 = n_{1l}^{-1} \sum_{i=1}^{n} I(\hat{w}_i \in \hat{Q}_l)(X_i Y_i - \bar{y}_{1l})^2$, $s_{0l}^2 = (n_l - n_{1l})^{-1} \sum_{i=1}^{n} I(\hat{w}_i \in \hat{Q}_l)((1 - X_i) Y_i - \bar{y}_{0l})^2$, $\bar{y}_{1l} = n_{1l}^{-1} \sum_{i=1}^{n} I(\hat{w}_i \in \hat{Q}_l) X_i Y_i$, and $\bar{y}_{0l} = (n_l - n_{1l})^{-1} \sum_{i=1}^{n} I(\hat{w}_i \in \hat{Q}_l)(1 - X_i) Y_i$.

4.5 Simulations

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

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

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

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

For those included in the randomized trial ($S_i = 1$), X_i was generated as Bern(0.5) and the response Y was generated according to $Y_i = \nu_0 + \nu_1 Z_{1i} + \xi X_i + \alpha Z_{1i} X_i + \epsilon_i, \epsilon_i \sim N(0, 1)$. For scenarios 1 to 4, $(\nu_0, \nu_1, \xi, \alpha) = (0, 1, 2, 1)$. For scenarios 5 to 6, $(\nu_0, \nu_1, \xi, \alpha) = (0, 1, 2, 2)$. Two sampling score models were considered (i.e., weak or moderate Z and S association):

Scenario 1, 3, and 5 set $\boldsymbol{\beta} = (-7, 0.4)$; Scenario 2, 4, and 6 set $\boldsymbol{\beta} = (-7, 0.6)$. The truth was calculated for each scenario using the distribution of Z in the target population. The truth was $\Delta_0 = 2.2$ for scenarios 1 and 2 and $\Delta_0 = 2$ for scenarios 3 through 6.

Comparisons between the IPSW and stratified estimator when the sampling score model is correctly specified are summarized in Table 4.1. The estimated sampling scores were computed using logistic regression with the covariate Z_{1i}. The ITT estimator was biased for all scenarios and had low coverage (results not shown). Depending on the scenario, the size of the trial ranged from $n = 987$ to $n = 1{,}091$ participants on average over the simulations for each scenario. For all scenarios, $\hat{\Delta}_{IPW}$ was unbiased. For scenarios 1 to 2, $\hat{\Delta}_S$ was unbiased and standard errors were comparable for the two estimators. For scenarios 3 to 6, $\hat{\Delta}_S$ was biased, possibly due to residual confounding from a continuous covariate in the sampling score model. For the IPSW estimator, the average of the estimated standard error was approximately equal to the Monte Carlo standard error, supporting the derivations of the sandwich-type estimator of the variance. Coverage was around 95% for Wald confidence interval of $\hat{\Delta}_{IPW}$ for all scenarios. With a continuous covariate, the Wald confidence interval of the stratified estimator had poor coverage, particularly in the presence of stronger effect modification (i.e., scenarios 5 and 6). Upon visual inspection, the IPSW estimator appeared to be normally distributed (Figure 4.1).

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

calculated for each scenario using the distribution of $\boldsymbol{Z}$ in the target population. The truth was $\Delta_0 = 2.8$ for scenarios 1 and 2 and $\Delta_0 = 2$ for scenarios 3 through 6.

When the sampling score model is misspecified, comparisons between the IPSW and stratified estimator are summarized in Table 4.2. The bias was reduced by approximately half when either the IPSW or the stratified estimator was employed, as compared to the naive within-trial estimator. The empirical sandwich estimator of the variance of the IPSW estimator performed reasonably well when the sampling score model was misspecified.

4.6 Applications

4.6.1 ACTG 320

The ACTG 320 trial examined the safety and efficacy of adding a protease inhibitor (PI) to an HIV treatment regimen with two nucleoside analogues. A total of 1,156 participants were enrolled in ACTG 320 between January 1996 and January 1997 and were recruited from 33 AIDS clinical trial units and 7 National Hemophilia Foundation sites in the U.S. and Puerto Rico (Hammer et al., 1997). 200 women were enrolled in ACTG 320 (Hammer et al., 1997). The baseline characteristics of these women and all participants are shown in Table 4.3 and Table 4.5, respectively.

WIHS and CNICS were considered to be representative samples of their respective target populations and this analysis only included participants who were HIV-positive, highly active antiretroviral therapy (HAART) naive, and had CD4 cell counts $\leq$ 200 cells/mm^3 at the previous visit ($m = 493$ and $m = 6{,}158$, respectively). Lab information (i.e., CD4 cell count) was carried forward for up to two years. The WIHS is a prospective, observational, multicenter study of women living with HIV and women at risk for HIV infection in the U.S. (Bacon et al., 2005). A total of 4,129 women (1,065 HIV-uninfected) were enrolled between October 1994 and December 2012 at six U.S. sites. Of the 493 women included in the WIHS sample, 82% were non-white, median age was 40 years, and 37% had a history of injection drug use (IDU). The median CD4 count was 108 cells/mm^3. Table 4.3 displays the characteristics of the

women in the WIHS sample.

The CNICS captures comprehensive and standardized clinical data from point-of-care electronic medical record systems for population-based HIV research (Kitahata et al., 2008). CNICS is considered to be representative of all people living with HIV and in clinical care in the U.S. The CNICS cohort includes over 27,000 HIV-infected adults (at least 18 years of age) engaged in clinical care since January 1, 1995 at eight CFAR sites in the U.S. Of the 6,158 participants included in the CNICS sample, 80% were male, 60% were non-white, median age was 41 years, and 20% had a history of IDU. The median CD4 count was 89 cells/mm^3. Table 4.5 displays the characteristics of participants in the CNICS sample.

The IPSW estimator was employed to assess the generalizability of the difference in the average change in CD4 from baseline to week 4 between treatment groups observed among women in the ACTG 320 to all women currently living with HIV in the U.S. and among all participants in the ACTG 320 to all people currently living with HIV in the U.S. Based on CDC estimates, the size of the first target population was assumed to be 280,000 women and the size of the second target population was assumed to be 1.1 million people (CDC, 2012).

First, the presence of conditions that could induce a lack of generalizability was assessed in the datasets. Namely, the variables associated with trial participation that are also treatment effect modifiers were identified. The distributions of baseline covariates differed between the women in the trial and WIHS cohort participants (Table 4.3). Age, history of injection drug use (IDU), race, and CD4 were associated with trial participation (P value < 0.001, P value $= 0.003$ and P value < 0.001, and P value $= 0.003$, respectively). Among women in the trial, baseline CD4 was associated with the outcome (P value $= 0.004$), but none of the other measured covariates were associated with the outcome. There was effect modification on the difference scale by CD4 at baseline (P value $= 0.003$), but not by any of the other (measured) covariates. There were differences in the point estimates of treatment effects across levels of all four covariates (Table 4.4).

The distributions of all covariates except sex differed between all participants in the trial and the CNICS cohort participants (Table 4.5). Age, race, and CD4 were associated with

trial participation (P value < 0.001 for each). In the trial, age, sex and baseline CD4 were associated with the outcome (P value $= 0.04$, P value $= 0.04$ and P value < 0.001, respectively). In the trial, there was effect modification on the difference scale by race and history of IDU (P value $= 0.001$ and P value $= 0.05$, respectively), but not any of the other (measured) covariates. There were differences in the point estimates of treatment effects across levels of all five covariates (Table 4.6).

Second, the within-trial treatment effects were computed separately among women only and all participants. This was an as-treated analysis and ignored treatment compliance issues. At week 4, women randomized to a regimen with a PI had an average change in CD4 cell count 24 cells/mm^3 higher than women randomized to a regimen without a PI (95% confidence interval (CI) $= (7, 41)$). The average change in CD4 cell count was 55 cells/mm^3 among those on a PI, compared to 31 cells/mm^3 among those not on a PI. At week 4 among all ACTG 320 participants, those randomized to the regimen with a PI had an average change in CD4 cell count 19 cells/mm^3 higher than those randomized to a regimen without a PI (95% CI $= (12, 25)$). Those on the regimen with a PI had an average change of 46 cells/mm^3, compared to an average of 27 cells/mm^3 among those on the regimen without a PI.

Third, the population average treatment effect was estimated used the IPSW estimator in equation (4.1). To estimate the sampling scores, the data from the ACTG trial and cohort (i.e., WIHS or CNICS) were analyzed together, with $S = 1$ for those in the ACTG trial and $S = 0$ for those in the cohort. A logistic regression model was fit on the combined trial and weighted cohort data. 116 (10%) of trial participants were missing CD4 count at week 4, so they were excluded. Cohort participants were inverse weighted by the size of the cohort divided by the size of the target (i.e., $\hat{r}_i = 493/(280{,}000-173)$ for WIHS and $\hat{r}_i = 6{,}158/(1{,}100{,}000\text{-}1{,}040)$ for CNICS) and trial participants were given a weight of $\hat{r}_i = 1$. The outcome was trial participation and the possible covariates were sex, race, age, history of IDU, and baseline CD4. Variables associated with trial participation, the outcome, or effect modifiers, as well as all pairwise interactions, were included in the sampling score model. Due to positivity issues, sex was excluded from the analysis generalizing the ACTG 320 results

among women to all women living with HIV in the U.S.

Table 4.11 displays the results for ACTG 320 generalized to both target populations. Among women, there was a significant difference in change in CD4 cell count between the two treatment groups. At week 4, women randomized to the regimen with a PI had an average change in CD4 cell count 46 cells/mm^3 higher than women randomized to regimen without a PI (95% CI = $(23, 70)$). Among all participants, there was a significant difference in average change in CD4 cell count between the two treatment groups. At week 4, those randomized to a regimen with a PI had a change in an average CD4 cell count 17 cells/mm^3 higher than those randomized to a regimen without a PI (95% CI = $(9, 25)$).

4.6.2 ACTG A5202

The ACTG A5202 trial examined equivalence of abacavir-lamivudine (ABC-3TC) or tenofovir disoproxil fumarate-emtricitabine (TDF-FTC) plus efavirenz or ritonavir-boosted atazanavir. A total of 1,857 participants were enrolled in ACTG A5202 between September 2005 and November 2007 and were recruited from 59 ACTG sites in the U.S. and Puerto Rico (Sax et al., 2009, 2011). 322 women were enrolled in ACTG A5202 (Sax et al., 2009, 2011). The baseline characteristics are shown in Table 4.7 among women and in Table 4.9 among all participants.

WIHS and CNICS were considered to be representative samples of their respective target populations and this analysis only included participants who were HIV-positive, antiretroviral (ART) naive, and had viral load > 1,000 copies/ml at the previous visit (m = 1,012 and m = 12,302, respectively). Lab information was carried forward for up to two years (i.e., CD4 and viral load). Of the $1,012$ women included in the WIHS sample, 83% were non-white, median age was 39 years, 38% had a history of IDU, 35% had hepatitis B/C, and 37% had an AIDS diagnosis. The median CD4 count was 290 cells/mm^3 and the median log10 viral load was 4.61 copies/ml. Table 4.7 displays the characteristics of the women in the WIHS sample. Of the 12,302 participants included in the CNICS sample, 82% were male, 55% were non-white, median age was 39 years, 17% had a history of IDU, 18% had hepatitis B/C, and 23% had

an AIDS diagnosis. The median CD4 count was 271 cells/mm^3 and the median log10 viral load was 4.64 copies/ml. Table 4.9 displays the characteristics of participants in the CNICS sample.

The IPSW estimator was employed to assess the generalizability of the difference in the average change in CD4 from baseline to week 48 between treatment groups observed among women in ACTG A5202 to all women currently living with HIV in the U.S. and among all participants in the ACTG A5202 to all people currently living with HIV in the U.S. Based on CDC estimates, the size of the first target population was assumed to be 280,000 women and the size of the second target population was assumed to be 1.1 million people (CDC, 2012). Because randomization was the same for both stratum, the screening viral load strata were ignored in the illustrative example. Only blinded follow-up was included in the analysis.

First, the presence of conditions that could induce a lack of generalizability was assessed in the datasets. Namely, the variables associated with trial participation that are also treatment effect modifiers were identified. The distributions of baseline CD4, history of IDU, hepatitis B/C, and AIDS diagnosis differed between the trial and WIHS cohort participants (Table 4.7). Age, AIDS diagnosis, history of IDU, baseline CD4, hepatitis, and viral load were associated with trial participation (P value < 0.001, P value < 0.001, P value < 0.001, P value < 0.001, P value $= 0.003$, and P value < 0.001, respectively). Age (P value $= 0.02$) and CD4 (P value $= 0.01$) were associated with the outcome. There was effect modification on the difference scale by age (P value $= 0.02$), history of IDU (P value $= 0.03$), and hepatitis B/C (P value $= 0.04$). There were differences in the point estimates of treatment effects across levels of all covariates (Table 4.8).

The distributions of baseline CD4, history of IDU, hepatitis, and AIDS diagnosis differed between the trial and CNICS cohort participants (Table 4.9). Race, AIDS diagnosis, hepatitis B/C, history of IDU, CD4, and log viral load were associated with trial participation (P value < 0.001 for each variable). Age, hepatitis B/C, viral load, and CD4 were associated with the outcome (P value < 0.001, P value < 0.001, P value < 0.001, and P value $= 0.03$, respectively). There was effect modification on the difference scale by history of IDU and baseline CD4 (P

value = 0.007 and P value = 0.05, respectively). There were differences in the point estimates of treatment effects across levels of all covariates, except AIDS diagnosis (Table 4.10).

Second, the within-trial treatment effects were computed separately among women only and all participants. This was an as-treated analysis and ignored treatment compliance issues. Among the 322 women in A5202 at week 48, those randomized to ABC-3TC had an average change in CD4 cell count 1 cell/mm^3 higher than those randomized to a regimen with TDF-FTC (95% CI = $(-35, 37)$). The average change in CD4 cell count was 194 cells/mm^3 among those on ABC-3TC, compared to 193 cells/mm^3 among those on TDF-FTC. Among the 1,857 participants in A5202, those randomized to ABC-3TC had an average change in CD4 cell count 6 cells/mm^3 higher than those randomized to a regimen with TDF-FTC (95% CI = $(-8, 20)$). The average change in CD4 cell count was 193 cells/mm^3 among those on ABC-3TC, compared to 187 cells/mm^3 among those on TDF-FTC.

Third, the population average treatment effect was estimated used the IPSW estimator in equation (4.1). To estimate the sampling scores, the data from the ACTG trial and cohort (i.e., WIHS or CNICS) were analyzed together, with $S = 1$ for those in the ACTG trial and $S = 0$ for those in the cohort. A logistic regression model was fit on the combined trial and weighted cohort data. 417 (22%) of trial participants were missing CD4 count at week 48, so they were excluded. Cohort participants were inverse weighted by the size of the cohort divided by the size of the target (i.e., $\hat{r}_i$ = 1,012/(280,000-255) for WIHS and $\hat{r}_i$ = 12,302/(1,100,000-1,440) for CNICS) and trial participants were given a weight of $\hat{r}_i = 1$. The outcome was trial participation and the possible covariates were sex, race, age, history of IDU, hepatitis B/C, AIDS diagnosis, baseline CD4 and baseline log10 viral load. Variables associated with trial participation, the outcome, or effect modifiers, as well as all pairwise interactions, were included in the sampling score model. Because hepatitis B/C and history of IDU were correlated ($r = 0.69$), history of IDU was excluded from the sampling score model. Due to positivity issues, sex was excluded from the analysis generalizing the ACTG A5202 results among women to all women living with HIV in the U.S.

Table 4.11 displays the results for ACTG A5202 generalized separately to both target

populations. Among women, the differences in the average change in CD4 cell count at week 48 between the regimens was computed. Women randomized to ABC-3TC had an average change in CD4 cell count 35 cells/mm^3 higher than women randomized to TDF-FTC (95% CI $= (-45, 115)$). Among all participants, the differences in the average change in CD4 cell count at week 48 between the regimens was computed. Those randomized to ABC-3TC had an average change in CD4 cell count 2 cells/mm^3 lower than those randomized to TDF-FTC (95% CI $= (-31, 28)$).

4.7 Discussion

Following Cole and Stuart (2010) and Stuart et al. (2011), we considered an estimator using inverse probability of sampling weights to generalize results from a randomized trial to a specific target population. The IPSW estimator compares the outcome in the target population if (possibly contrary to fact) everyone had been randomized to treatment with the outcome in the target population if (possibly contrary to fact) everyone had been randomized to control. The IPSW estimator was shown to be consistent and asymptotically normal and a consistent sandwich-type estimator of the variance was provided. In the following, we discuss some recent work addressing generalizability and explore caveats of this approach.

In the illustrative example, the IPSW estimator was employed to generalize results from the ACTG to all people currently living with HIV in the U.S. For ACTG 320, the effect estimated with the ITT was comparable to the effect estimated with the IPSW, so the results appear to be generalizable to all people living with HIV in the U.S. For the A5202 results among women, the difference in the effect estimates is primarily due to hepatitis, which was associated with participation in the trial and a treatment effect modifier. Results were not sensitive to the specification of the size of the target population; however, some results were sensitive to the specification of the sampling score model.

In a previous paper by Cole and Stuart (2010), the ACTG 320 results were generalized to a target population of all people infected with HIV in the U.S. Consistency and asymptotic normality of the proposed estimator were not formally shown. The results herein complete that

effort. In Cole and Stuart (2010), information on the target was obtained using survey data (i.e., CDC estimates). The approach presented in herein uses richer data from representative cohorts.

When applying this method, the analysis is subject to the following considerations. The absence of unmeasured covariates associated with the trial participation mechanism and treatment effect modifiers is an untestable assumption. Treatment compliance issues were ignored in this method; however, this issue should be considered in analyses. The sampling score model was assumed to be correct (i.e., correct covariate functional forms); however, this is not guaranteed in practice. The stratified estimator (Tipton et al., 2014; O'Muircheartaigh and Hedges, 2013) requires that individuals sharing the same sampling score can be identified, which may be difficult in practice. This estimator may be biased when there is residual confounding within strata and, therefore, is not a consistent estimator of the PATE in some cases (e.g., a continuous covariate in the sampling score model) (Lunceford and Davidian, 2004).

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

Acknowledgments

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

Table 4.1: Summary of Monte Carlo results for estimators of the population average treatment effect when the sampling score model was correctly specified with a continuous outcome for 5,000 samples with $m = 4{,}000$ and $n \approx 1{,}000$. Scenarios are described in the text. $\Delta_0 = 2.2$ for scenarios 1 and 2 and $\Delta_0 = 2.0$ for scenarios 3 to 6 (ITT = intention-to-treat; S = stratified; IPSW = inverse probability of sampling weighted; ESE = Empirical standard error (×100); ASE = Average standard error (×100); ECP = Empirical coverage probability)

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

Table 4.2: Summary of Monte Carlo results for estimators of the population average treatment effect when the sampling score model was misspecified with a continuous outcome for 5,000 samples with $m = 4{,}000$ and $n \approx 1{,}000$. Scenarios are described in the text. $\Delta_0 = 2.8$ for scenarios 1 and 2 and $\Delta_0 = 2.0$ for scenarios 3 to 6 (ITT = intention-to-treat; S = stratified; IPSW = inverse probability of sampling weighted; ESE = Empirical standard error (×100); ASE = Average standard error (×100); ECP = Empirical coverage probability)

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

Table 4.3: Characteristics of 493 women in the Women's Interagency HIV Study (WIHS) who were HIV-positive, HAART naive, and had CD4 cell count $\leq$ 200 cells/mm^3 at the previous visit and 200 women at baseline in AIDS Clinical Trials Group (ACTG) 320 by treatment group (with and without a protease inhibitor (PI))

Variable	**WIHS** (m = 493)	**ACTG 320** (n = 200)	**ACTG 320 Protease Inhibitor (PI)** (n_1 = 106)	**ACTG 320 No Protease Inhibitor (PI)** (n_0 = 94)
Race or ethnic group - no. (%)				
White, non-Hispanic	87 (18)	61 (31)	26 (25)	35 (37)
Black, non-Hispanic	272 (55)	95 (48)	54 (51)	41 (44)
Hispanic	124 (25)	42 (21)	25 (24)	17 (18)
Asian/Other	10 (2)	2 (1)	1 (1)	1 (1)
Median age - yr (Q1-Q3)	40 (35-45)	36 (30-42)	37 (31-42)	36 (30-43)
Age group - no. (%)				
16-<30 yr	35 (7)	46 (23)	22 (21)	24 (26)
30-<40 yr	211 (43)	88 (44)	48 (45)	40 (43)
40-<50 yr	196 (40)	53 (27)	27 (26)	26 (28)
$\geq$50 yr	51 (10)	13 (7)	9 (9)	4 (4)
Injection drug use - no. (%)	180 (37)	36 (18)	24 (23)	12 (13)
Median CD4 count (Q1-Q3)	108 (41-172)	82 (26-139)	93 (29-139)	70 (23-138)
Baseline CD4 count - no. (%)				
<50 cells/mm^3	148 (30)	72 (36)	35 (33)	37 (39)
50-<100 cells/mm^3	83 (17)	43 (22)	22 (21)	21 (22)
100-<200 cells/mm^3	182 (37)	73 (37)	44 (42)	29 (31)
$\geq$200 cells/mm^3	80 (16)	12 (6)	5 (5)	7 (7)

Table 4.4: Difference in the average change in CD4 from baseline to week 4 between treatment groups (protease inhibitor (PI) vs. no PI) for each level of the covariates among 173 women in AIDS Clinical Trials Group 320 with a corresponding 95% confidence interval (CI)

Variable	**Difference in Change in CD4 at Week 4** Mean (95 % CI)
Race or ethnic group	
White, non-Hispanic	21 (-11, 53)
Black, non-Hispanic	19 (-8, 45)
Hispanic/Asian/Other	35 (-2, 72)
Age group	
18-<30 yr	-3 (-47, 41)
30-<40 yr	33 (7, 57)
40-<50 yr	26 (-7, 60)
$\geq$50 yr	16 (-43, 75)
Injection drug use	
Yes	43 (-5, 90)
No	23 (4, 42)
Baseline CD4 count	
<50 cells/mm^3	14 (-13, 41)
50-<100 cells/mm^3	57 (22, 92)
$\geq$100 cells/mm^3	14 (-14, 41)

Table 4.5: Characteristics of 6,158 participants in the CFAR Network of Integrated Clinical Systems (CNICS) who were HIV-positive, HAART naive, and had CD4 cell count ≤ 200 cells/mm^3 at the previous visit and 1,156 participants at baseline in AIDS Clinical Trials Group (ACTG) 320 by treatment group (with and without a protease inhibitor (PI))

Variable	**CNICS** (m = 6,158)	**ACTG 320** (n = 1,156)	**ACTG 320 Protease Inhibitor (PI)** ($n_1 = 577$)	**ACTG 320 No Protease Inhibitor (PI)** ($n_0 = 579$)
Male sex - no. (%)	4,909 (80)	956 (83)	471 (82)	485 (84)
Race or ethnic group - no. (%)[a]				
White, non-Hispanic	2,436 (40)	598 (52)	303 (53)	295 (51)
Black, non-Hispanic	2,690 (44)	328 (28)	163 (28)	165 (29)
Hispanic	734 (12)	205 (18)	99 (17)	106 (18)
Asian/Other	298 (5)	25 (2)	12 (2)	13 (2)
Median age - yr (Q1-Q3)	41 (34-47)	38 (33-44)	38 (33-44)	38 (33-44)
Age group - no. (%)				
16-<30 yr	714 (12)	142 (12)	69 (12)	73 (13)
30-<40 yr	2,108 (34)	536 (47)	272 (47)	264 (46)
40-<50 yr	2,315 (38)	350 (30)	169 (29)	181 (31)
≥50 yr	1,021 (17)	128 (11)	67 (12)	61 (11)
Injection drug use - no. (%)	1,241 (20)	184 (16)	91 (16)	93 (16)
Median CD4 count (Q1-Q3)	89 (27-172)	75 (23-137)	80 (24-138)	70 (23-135)
Baseline CD4 count - no. (%)				
<50 cells/mm^3	2,237 (36)	453 (39)	219 (38)	234 (41)
50-<100 cells/mm^3	1,047 (17)	248 (22)	118 (20)	130 (23)
100-<200 cells/mm^3	1,818 (30)	372 (32)	200 (35)	172 (30)
≥200 cells/mm^3	1,056 (17)	82 (7)	40 (7)	42 (7)

[a]One A5202 participant missing baseline CD4 cell count.

Table 4.6: Difference in the average change in CD4 from baseline to week 4 between treatment groups (protease inhibitor (PI) vs. no PI) for each level of the covariates among 1,040 participants in AIDS Clinical Trials Group (ACTG) 320 with a corresponding 95% confidence interval (CI)

Variable	**Difference in Change in CD4 at Week 4** Mean (95 % CI)
Sex	
Male	17 (10, 25)
Female	24 (8, 40)
Race or ethnic group	
White, non-Hispanic	14 (4, 23)
Black, non-Hispanic	11 (-2, 23)
Hispanic	48 (32, 64)
Asian/Other	9 (-35, 53)
Age group	
18-<30 yr	14 (-9, 37)
30-<40 yr	23 (13, 33)
40-<50 yr	13 (2, 25)
$\geq$50 yr	20 (1, 38)
Injection drug use	
Yes	3 (-14, 20)
No	21 (14, 29)
Baseline CD4 count	
<50 cells/mm^3	24 (14, 35)
50-<100 cells/mm^3	26 (12, 41)
100-<200 cells/mm^3	9 (-3, 21)
$\geq$200 cells/mm^3	-3 (-29, 23)

Table 4.7: Characteristics of 1,012 women in the Women's Interagency HIV Study (WIHS) who were HIV-positive, ART naive, and had viral load > 1000 copies/ml at the previous visit and 322 women at baseline in AIDS Clinical Trials Group (ACTG) A5202 by treatment group (abacavir-lamivudine (ABC-3TC) vs. tenofovir disoproxil fumarate-emtricitabine (TDF-FTC))

	WIHS	**ACTG A5202**	**ACTG A5202 ABC-3TC**	**ACTG A5202 TDF FTC**
Variable	(m = 1,012)	(n = 322)	(n_1 = 173)	(n_0 = 149)
Race or ethnic group - no. (%)				
White, non-Hispanic	171 (17)	57 (18)	30 (17)	27 (18)
Black, non-Hispanic	586 (58)	172 (53)	94 (54)	78 (52)
Hispanic	222 (22)	82 (26)	42 (24)	40 (27)
Asian/Other	33 (3)	11 (3)	7 (4)	4 (3)
Median age - yr (Q1-Q3)	39 (33-44)	39 (31-46)	39 (31-46)	39 (31-46)
Age group - no. (%)				
16-<30 yr	123 (12)	57 (18)	30 (17)	27 (18)
30-<40 yr	435 (43)	110 (34)	62 (36)	48 (32)
40-<50 yr	345 (34)	107 (33)	54 (31)	53 (36)
≥50 yr	109 (11)	48 (15)	27 (16)	21 (14)
Injection drug use - no. (%)	388 (38)	18 (6)	9 (5)	9 (6)
Hepatitis B/C - no. (%)	356 (35)	25 (8)	14 (8)	11 (7)
AIDS diagnosis - no. (%)	373 (37)	62 (19)	39 (23)	23 (15)
CD4 count - no. (%)				
<50 cells/mm^3	102 (10)	61 (19)	38 (22)	23 (15)
50-<100 cells/mm^3	61 (6)	24 (7)	15 (9)	9 (6)
100-<200 cells/mm^3	162 (16)	55 (17)	28 (16)	27 (18)
200-<350 cells/mm^3	295 (29)	130 (40)	65 (38)	65 (44)
≥350 cells/mm^3	392 (39)	52 (16)	27 (16)	25 (17)
Median CD4 count (Q1-Q3)	290 (162-423)	226 (87-313)	209 (60-308)	249 (129-316)
Viral load - no. (%)				
<50,000 cp/ml	552 (55)	187 (58)	93 (54)	94 (63)
50,000-<100,000 cp/ml	144 (14)	62 (19)	33 (19)	29 (20)
100,000-<300,000 cp/ml	193 (19)	38 (12)	24 (14)	14 (9)
300,000-<500,000 cp/ml	55 (5)	9 (3)	6 (3)	3 (2)
≥500,000 cp/ml	68 (7)	26 (8)	17 (10)	9 (6)
Median log10 viral load (Q1-Q3)	4.61 (4.04-5.11)	4.58 (4.07-4.93)	4.62 (4.10-5.09)	4.55 (4.04-4.86)

Table 4.8: Difference in the average change in CD4 from baseline to week 48 between treatment groups (abacavir-lamivudine (ABC-3TC) vs. tenofovir disoproxil fumarate-emtricitabine (TDF-FTC)) for each level of the covariates among 255 women in AIDS Clinical Trials Group (ACTG) A5202 with a corresponding 95% confidence interval (CI)

Variable	**Difference in the Average Change in CD4 at Week 48** Mean (95% CI)
Race or ethnic group	
White, non-Hispanic	-58 (-144, 28)
Black, non-Hispanic	21 (-27, 70)
Hispanic	18 (-55, 91)
Asian/Other	-90 (-301, 120)
Age group	
18-<30 yr	108 (24, 191)
30-<40 yr	-20 (-80, 40)
40-<50 yr	-39 (-102, 24)
≥50 yr	-15 (-110, 79)
Injection drug use	
Yes	-7 (-44, 30)
No	-103 (-202, -5)
Hepatitis B/C	
Yes	130 (4, 257)
No	-10 (-48, 27)
AIDS diagnosis	
Yes	-27 (-116, 62)
No	5 (-35, 45)
CD4 count	
<50 cells/mm^3	-13 (-103, 77)
50-<100 cells/mm^3	49 (-94, 192)
100-<200 cells/mm^3	2 (-87, 90)
200-<350 cells/mm^3	34 (-21, 89)
≥350 cells/mm^3	-83 (-168, 3)
Viral load	
<50,000 cp/ml	-1 (-46, 45)
50,000-<100,000 cp/ml	-16 (-99, 68)
100,000-<300,000 cp/ml	57 (-56, 171)
300,000-<500,000 cp/ml	-48 (-312, 216)
≥500,000 cp/ml	-40 (-196, 116)

Table 4.9: Characteristics of 12,302 participants in the CFAR Network of Integrated Clinical Systems (CNICS) who were HIV-positive, ART naive, and had viral load > 1000 copies/ml at the previous visit and 1,857 participants at baseline in AIDS Clinical Trials Group (ACTG) A5202 by treatment group (abacavir-lamivudine (ABC-3TC) vs. tenofovir disoproxil fumarate-emtricitabine (TDF-FTC))

	CNICS	**ACTG A5202**	**ACTG A5202 ABC-3TC**	**ACTG A5202 TDF FTC**
Variable	(m = 12,302)	(n =1,857)	(n_1 = 928)	(n_0 = 929)
Male sex - no. (%)	10,063 (82)	1,535 (83)	755 (81)	780 (84)
Race or ethnic group[a] - no. (%)				
White, non-Hispanic	5,567 (45)	746 (46)	363 (39)	383 (41)
Black, non-Hispanic	4,682 (38)	615 (33)	317 (34)	298 (32)
Hispanic	1,420 (12)	429 (23)	214 (23)	215 (23)
Asian/Other	633 (5)	62 (3)	31 (3)	31 (3)
Median age - yr (Q1-Q3)	39 (31-46)	38 (31-45)	38 (30-45)	39 (31-45)
Age group - no. (%)				
16-<30 yr	2,454 (20)	404 (22)	201 (22)	203 (22)
30-<40 yr	4,225 (34)	625 (34)	335 (36)	290 (31)
40-<50 yr	3,896 (32)	573 (31)	273 (29)	300 (32)
≥50 yr	1,727 (14)	255 (14)	119 (13)	136 (15)
Injection drug use - no. (%)	2,042 (17)	162 (9)	77 (8)	85 (9)
Hepatitis B/C - no. (%)	2,245 (18)	165 (9)	75 (8)	90 (10)
AIDS diagnosis - no. (%)	2,834 (23)	312 (17)	172 (19)	140 (15)
CD4 count[b] - no. (%)				
<50 cells/mm^3	2,000 (16)	339 (18)	176 (19)	163 (18)
50-<100 cells/mm^3	920 (7)	150 (8)	74 (8)	76 (8)
100-<200 cells/mm^3	1,692 (14)	311 (17)	159 (17)	152 (16)
200-<350 cells/mm^3	3,262 (27)	656 (35)	312 (34)	344 (37)
≥350 cells/mm^3	4,428 (36)	400 (22)	207 (22)	193 (21)
Median CD4 count (Q1-Q3)	271 (109-427)	230 (90-334)	229 (84-338)	230 (96-330)
Viral load - no. (%)				
<50,000 cp/ml	6,450 (52)	1,000 (54)	492 (53)	508 (55)
50,000-<100,000 cp/ml	1,861 (15)	391 (21)	196 (21)	195 (21)
100,000-<300,000 cp/ml	2,232 (18)	203 (11)	106 (11)	97 (10)
300,000-<500,000 cp/ml	744 (6)	72 (4)	38 (4)	34 (4)
≥500,000 cp/ml	1,015 (8)	191 (10)	96 (10)	95 (10)
Median log10 viral load (Q1-Q3)	4.64 (3.95-5.18)	4.66 (4.33-5.01)	4.66 (4.31-5.06)	4.65 (4.34-4.96)

[a]Five A5202 participants were missing race.
[b]One A5202 participant was missing CD4 cell count.

Table 4.10: Difference in the average change in CD4 from baseline to week 48 between treatment groups (abacavir-lamivudine (ABC-3TC) vs. tenofovir disoproxil fumarate-emtricitabine (TDF-FTC)) for each level of the covariates among 1,440 participants in ACTG A5202 with a corresponding 95% confidence interval (CI)

Variable	**Difference in the Average Change in CD4 at Week 48** Mean (95% CI)
Sex	
Male	22 (-5, 48)
Female	3 (-13, 18)
Race or ethnic group	
White, non-Hispanic	-4 (-27, 18)
Black, non-Hispanic	14 (-11, 39)
Hispanic	11 (-19, 41)
Asian/Other	19 (-56, 94)
Age group	
18-<30 yr	10 (-22, 41)
30-<40 yr	5 (-20, 29)
40-<50 yr	10 (-15, 36)
≥50 yr	-13 (-51, 26)
Injection drug use	
Yes	-0.03 (-15, 15)
No	-38 (-72, -3)
Hepatitis B/C	
Yes	49 (1, 99)
No	1 (-14, 16)
AIDS diagnosis	
Yes	6 (-31, 43)
No	5 (-11, 21)
CD4 count	
<50 cells/mm^3	14 (-21, 51)
50-<100 cells/mm^3	8 (-44, 60)
100-<200 cells/mm^3	13 (-22, 49)
200-<350 cells/mm^3	16 (-8, 39)
≥350 cells/mm^3	-21 (-51, 8)
Viral load	
<50,000 cp/ml	-3 (-21, 15)
50,000-<100,000 cp/ml	38 (7, 69)
100,000-<300,000 cp/ml	20 (-27, 68)
300,000-<500,000 cp/ml	-46 (-125, 32)
≥500,000 cp/ml	-24 (-74, 26)

Table 4.11: Results for continuous outcomes in two AIDS Clinical Trials Group (ACTG) trials where the sampling score model included all variables associated with trial participation, the outcome, or effect modifiers (with a linear term for continuous variables) and all pairwise interactions (ITT = intention-to-treat; IPSW = inverse probability of sampling weighted; S = stratified). Difference in means with a 95% confidence interval displayed below.

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

[a]For ACTG 320, the treatment contrast was PI vs. no PI.
[b]For A5202, the treatment contrast was ABC-3TC vs. TDF-FTC.

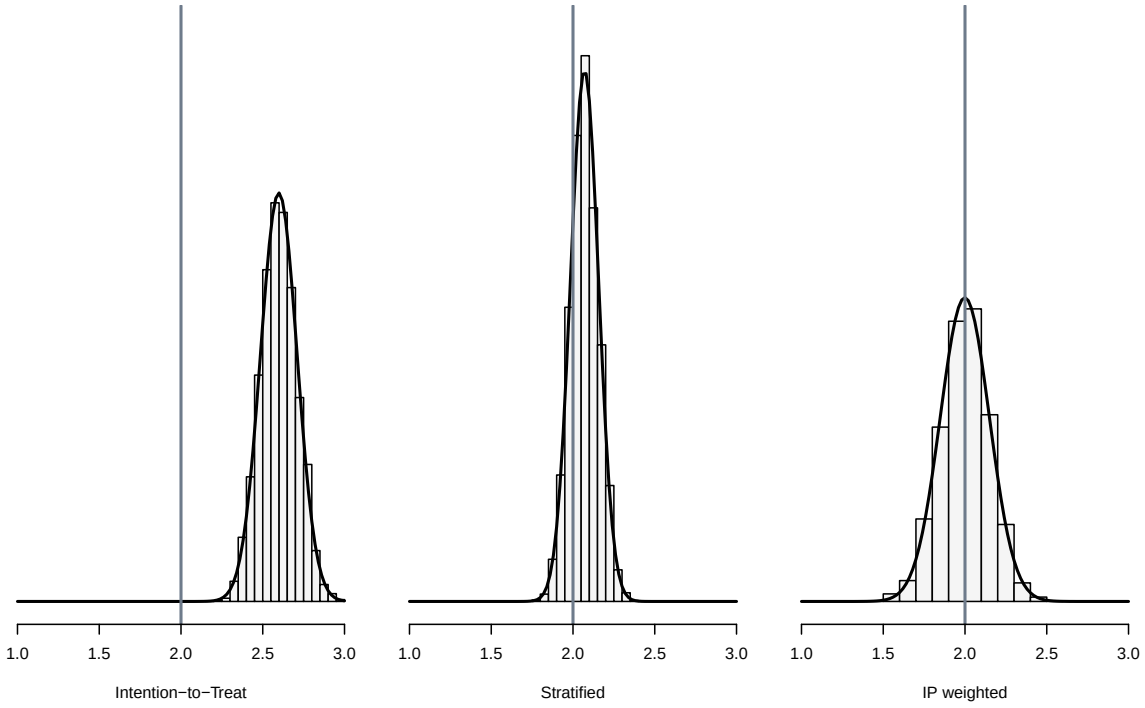


Figure 4.1: Comparison of the distributions of intention-to-treat estimator $\hat{\Delta}_{ITT}$, inverse probability of sampling weighted estimator $\hat{\Delta}_{IPSW}$, and stratified estimator $\hat{\Delta}_S$ based on 5,000 simulated datasets where the sampling score model is correctly specified and $\Delta_0 = 2.0$ with one continuous covariate, $\boldsymbol{\beta} = (-7, 0.6)$, and $\alpha = 1$

CHAPTER 5: GENEARLIZING TRIAL RESULTS FOR RIGHT-CENSORED DATA USING INVERSE PROBABILITY OF SAMPLING WEIGHTS

5.1 Introduction

Time-to-event endpoints are often of interest in clinical trials. Kaplan-Meier estimators are used to quantify survival distributions and are a commonly used nonparametric estimator in survival analysis (Kalbfleisch and Prentice, 2002). If the trial is random sample from the target population, standard methods, such as the Kaplan-Meier estimator, are appropriate and comparisons between groups can be made with a log rank test. However, when the trial participation mechanism is possibly non-random sample, properly weighting the observed trial data to estimate the survival distribution in the target population may be necessary.

The Kaplan-Meier estimator is also appealing because it does not require a proportional hazards assumption or possible complications faced by hazard ratios estimated using a Cox proportional hazards model. Presenting estimated survival curves is an alternative to reporting hazard ratios that may be more interpretable because survival curves summarize all information from baseline up to any time t (Cole and Hernan, 2004; Hernan, 2010). Robins and Finkelstein (2000) proposed an adjusted Kaplan-Meier estimator using inverse probability (IP) of censoring weights to estimate effects in the presence of selection bias using randomized trial data. Xie and Liu (2005) developed an adjusted Kaplan-Meier estimator using inverse probability (IP) of treatment weights to estimate effects in the presence of confounding using observational data. They presented estimators for the marginal survival curves, including the development of a weighted log rank test.

Cole and Stuart (2010) proposed a method for generalizing trial results to a target population for right-censored data using an IP-weighted Cox model with sampling weights. A robust estimator of the variance of the hazard ratio was employed (Lin and Wei, 1989). The authors provided an expression for the bias of the intention-to-treat estimator when the parameter of interest is a difference in means in the target and a proof that an inverse probability weighted estimator is unbiased for the mean of the potential outcomes in the target population; however, the performance of the bootstrap standard error

was not evaluated. The authors demonstrated this method by generalizing results from the acquired immunodeficiency syndrome (AIDS) Clinical Trials Group (ACTG) to all people currently living with human immunodeficiency virus (HIV) in the U.S.

Following Cole and Stuart (2010) and Buchanan et al. (In preparation), an inverse probability of sampling weighted (IPSW) estimator is considered for right-censored data using a weighted Kaplan-Meier (KM) estimator and comparisons are made to a proposed stratified estimator. The outline of the remainder of this paper is as follows. In Section 5.2, the assumptions and notation for this method are discussed. Expressions for the IPSW estimator and the stratified estimator are provided in Section 5.3. The large sample performance of the IPSW and stratified estimators are compared using simulations in Section 5.4. In the penultimate section, the IPSW estimator is applied to generalize results from two ACTG trials to all people currently living with HIV in the U.S. Related work and caveats of this method are discussed in the last section.

5.2 Assumptions and Notation

Consider a setting where two data sources are available. A random sample (e.g., cohort study) of size m is drawn from the near infinite target population and assumed to be representative. A second sample of n individuals participate in a randomized trial, and the treatment assignment mechanism is known to the analyst. The trial is possibly a non-random (i.e., biased) sample from the near infinite target population. In addition, the treatment effect is possibly modified by the same covariates that differ between the trial and the near infinite target. Note that the population may remain unenumerated. Define $\mathbf{Z}_{(n+m)\times p}$ as a vector of fixed characteristics and assume that information on $\mathbf{Z}$ is available for those in the trial and cohort. Let $S = 1$ denote trial participation. For those in the trial, define X as the treatment indicator, where $X = 1$ if assigned to treatment. Let (T, C, δ, X) denote the right-censored trial data, where T is the event time, C is the censoring time, and δ is the event indicator with $\delta = 0$ if $T > C$ (and $\delta = 1$ if $T \leq C$). Define $T^* = \min(T, C)$ as the observed time in the trial. Let $i = 1, \ldots, n + m$ index trial and cohort participants.

Define T^1 as the event time that would have been seen if (possibly contrary to fact) the participant were to receive treatment at the time of randomization (i.e., $t = 0$) (and followed until they experienced the event) and T^0 as the event time that would have been seen if (possibly contrary to fact) the participant were to receive control at the time of randomization (and followed until they experienced the event). It is assumed throughout that the stable unit treatment value assumption (SUTVA)

(Rubin, 1980; Tipton, 2013) holds, i.e., there are no variations of treatment and there is no interference between participants. Let $\mathfrak{S}^1(t) = P(T^1 > t)$ and $\mathfrak{S}^0(t) = P(T^0 > t)$ denote the survival functions for the potential outcomes. Define $F^1(t) = P(T^1 \leq t)$ and $F^0(t) = P(T^0 \leq t)$ as the cumulative distribution functions for the potential outcomes. Consistency also holds, so $\mathfrak{S}(t) = \mathfrak{S}^1(t)X + \mathfrak{S}^0(t)(1 - X)$. The following additional conditions are also assumed: ignorable treatment assignment mechanism (no unmeasured confounders) gained through randomization $P(X = x|S = s, \mathbf{Z}, T^0, T^1) = P(X = x|S = s)$, ignorable trial participation mechanism conditional on $\mathbf{Z}$, so $P(S = s|\mathbf{Z}, T^0, T^1) = P(S = s|\mathbf{Z})$, and trial participation and treatment positivity $P(X = x|\mathbf{Z}, S = 1) > 0$ and $P(S = s|\mathbf{Z}) > 0$ for all $\mathbf{Z} = \mathbf{z}$. Measurement of all treatment effect modifiers associated with trial participation (for both those in the trial and cohort) will be sufficient to assume an ignorable trial participation mechanism. Under these assumptions, one can successfully map from the observed data to the counterfactual data needed to make inference and estimators can be expected to be consistent.

5.3 Estimators of the Marginal Survival Functions in the Target Population

The parameter of interest is the marginal survival function in the near infinite target population at a particular time t_j for each treatment group: $\mathfrak{S}^1(t_j) = P(T^1 > t_j)$ and $\mathfrak{S}^0(t_j) = P(T^0 > t_j)$. A traditional (i.e., unweighted) approach to estimating treatment effects is a difference in KM estimators. Suppose the events in the trial occur at D distinct times $t_1 < t_2 < \ldots < t_D$ for $j = 1, \ldots, D$ and ties are allowed. Let k index treatment with $k = 1$ for those randomized to treatment and $k = 0$ for those randomized to control. At time t_j, there are N_{j0} events out of Y_{j0} individuals at risk in the control group and N_{j1} events out of Y_{j1} individuals at risk in the treatment group. Define $\hat{N}_{j0} = \sum_{i:T_i=t_j} S_i(1 - X_i)\delta_i$ and $\hat{Y}_{j0} = \sum_{i:T_i \geq t_j} S_i(1 - X_i)$ and $\hat{N}_{j1} = \sum_{i:T_i=t_j} S_i X_i \delta_i$ and $\hat{Y}_{j1} = \sum_{i:T_i \geq t_j} S_i X_i$. For treatment group k, the standard (i.e., unweighted) KM estimator is defined as

$$\hat{\mathfrak{S}}^k_{KM}(t) = \prod_{t_j \leq t} \left[1 - \frac{\hat{N}_{jk}}{\hat{Y}_{jk}}\right] \tag{5.1}$$

if $\hat{Y}_{jk} > 0$ and $t_1 \leq t$. Otherwise, $\hat{S}^k_{KM}(t) = 1$ if $t < t_1$.

Two estimators that employ sampling scores are considered. The sampling scores are time-fixed and $\mathbf{Z}$ is defined using participant characteristics at the baseline visit of the trial and those same characteristics among participants in the cohort. As in Buchanan et al. (In preparation), a (weighted)

logistic regression is used to estimate the sampling scores $P(S = 1|\mathbf{Z} = \mathbf{z})$. To account for the random sampling of the cohort from the near infinite target, each participant in the cohort is inverse weighted by the sampling fraction $r_i = m/(N - n)$, where N is the size of the near infinite target population with $N >> n$ and $N >> m$. Each trial participant is given a weight of $r_i = 1$. Let $w(\mathbf{Z}, \boldsymbol{\beta}) = w = P(S = 1|\mathbf{Z})$, $w_i = w(\mathbf{Z_i}, \boldsymbol{\beta})$, and $\hat{w}_i = w(\mathbf{Z_i}, \hat{\boldsymbol{\beta}})$. Let $\boldsymbol{\beta}_0$ be the vector of true values of $\boldsymbol{\beta}_{1\times p}$ and $\hat{\boldsymbol{\beta}}_{1\times p}$ be the vector of estimators of $\boldsymbol{\beta}_{1\times p}$. The vector of parameters $\boldsymbol{\beta}_{1\times p}$ can be consistently estimated by solving the estimating equations

$$\sum_{i=1}^{n+m} \psi_\beta(S_i, \mathbf{Z}_i, \boldsymbol{\beta}) = \sum_{i=1}^{n+m} r_i^{-1} \frac{S_i - w_i}{w_i(1 - w_i)} \frac{\partial}{\partial \boldsymbol{\beta}} w_i \tag{5.2}$$

(Manski and Lerman, 1977; Scott and Wild, 1986, 2002).

Among those randomized to treatment, $\hat{N}^w_{j1} = \sum_{i:T_i=t_j} \frac{S_i X_i \delta_i}{\hat{w}_i}$ and $\hat{Y}^w_{j1} = \sum_{i:T_i \geq t_j} \frac{S_i X_i}{\hat{w}_i}$. Among those randomized to control, $\hat{N}^w_{j0} = \sum_{i:T_i=t_j} \frac{S_i(1-X_i)\delta_i}{\hat{w}_i}$ and $\hat{Y}^w_{j0} = \sum_{i:T_i \geq t_j} \frac{S_i(1-X_i)}{\hat{w}_i}$. For treatment group k, the IPSW estimator is defined as

$$\hat{\mathfrak{S}}^k(t) = \prod_{t_j \leq t} \left[1 - \frac{\hat{N}^w_{jk}}{\hat{Y}^w_{jk}}\right] \tag{5.3}$$

if $\hat{Y}^w_{jk} > 0$ and $t_1 \leq t$. Otherwise, $\hat{S}^k(t) = 1$ if $t < t_1$.

An alternative approach for estimating survival function in the target population uses stratification based on the estimated sampling scores (Tipton, 2013; O'Muircheartaigh and Hedges, 2013; Tipton et al., 2014). This proposed estimator is an extension of the estimator for continuous outcomes (Tipton, 2013) and is computed in the following steps. First, $\boldsymbol{\beta}_0$ is estimated using a (weighted) logistic regression model and the sampling scores $\hat{w}_i$ are computed. Sampling scores are used to form L strata according to the quintiles of the distribution in the target population. Since we assume the trial and cohort both arise from the same near infinite target, the distribution of sampling scores in the combined trial and cohort are used to estimate the quintiles (Tipton, 2013). The survival curve within each stratum is computed among those in the trial. Lastly, the survival curve in the target population is estimated as a weighted sum of the survival curve across strata, where the weight $\hat{w}_{pl}$ is the proportion of observations in stratum l in the target population. Let $i = 1, \ldots, (n + m)$ index participants and $l = 1, \ldots, L$ index stratum. For treatment group k, the stratified estimator of the survival function is defined as

$$\hat{\mathfrak{S}}^k_S(t) = \sum_{l=1}^{L} \hat{w}_{p_l} \left[\hat{\mathfrak{S}}^k_{KM,l}(t)\right]$$

where the L stratum are defined by the distribution of the sampling scores in the near infinite target population, k indexes treatment group and $\hat{w}_{p_l} = N_l/N$ with N_l as number in stratum l in the target and N is the size of the near infinite target.

In the illustrative examples, the average standard error of all three estimators was computed using a nonparametric bootstrap with 200 random samples with replacement of the trial and 200 random samples with replacement of the cohort data (Efron and Tibshriani, 1994). The risk difference (RD) at time $t = t_j$ was defined as the difference between the complement of the marginal survival curve for $k = 1$ (i.e., treated group) and the complement of the marginal survival curve for $k = 0$ (i.e., control group) at time $t = t_j$. The risk ratio (RR) at time $t = t_j$ was defined as the ratio of the complement of the marginal survival curve for $k = 1$ (i.e., treated group) over the complement of the marginal survival curve for $k = 0$ (i.e., control group) at time $t = t_j$.

5.4 Simulations

Simulations were performed to compare the IPSW estimator to the stratified and traditional (i.e., unweighted) KM estimators. A total of 5,000 datasets per scenario were simulated as follows. There were $N = 10^6$ observations in the near infinite target population and each had (Z_i, w_i), where $w_i = \{1 + \exp(-\beta_0 - \beta_1 Z_i)\}^{-1}$ was the true sampling score. One binary covariate $Z_i \sim \text{Bern}(0.2)$ and one continuous covariate $Z_i \sim N(0, 1)$, which were associated with trial participation and treatment effect modifiers, were considered in separate scenarios. 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 the trial. The parameters β_0 and β_1 were set to ensure that the probability of trial participation was a rare event (i.e., the size of the trial was approximately $n \approx 1,000$). The cohort was a random sample of size $m = 4,000$ from the target population (less those selected into the trial) and S_i was set to zero for those included in cohort. The trial was small compared to the size of the target, so the cohort was essentially a random sample from the target.

To estimate the weights, the combined trial ($S_i = 1$) and cohort data ($S_i = 0$) was used to fit a (weighted) logistic regression model with S_i as the outcome and the covariate Z_i. To account for the sampling of the cohort from the target, each participant in the cohort was inverse weighted by $\hat{r}_i = m/(N-n)$ in the logistic model for the sampling scores. Each trial participant was given a weight of $\hat{r}_i = 1$ (in the logistic model). This allowed for unbiased estimation of the parameters in the logistic regression model.

For those included in the randomized trial ($S_i = 1$), X_i was generated as Bern(0.5) and the lognormal survival time T was generated according to $T_i^* = \exp(\nu_0 + \nu_1 Z_i + \xi X_i + \alpha Z_i X_i + \epsilon_i), \epsilon_i \sim N(0,1)$. Note the survival times are lognormal and do not follow the proportional hazards assumption. Survival times greater than 10 years were administratively censored at that time. For scenarios 1, 2, 5, and 6, all participants were observed until the end of the study (i.e., if $T_i \leq 10$, then $T_i = T_i^*$). For the remaining scenarios, there was an independent censoring mechanism $C_i \sim \exp(0.5)$ and $T_i = \min(T_i^*, C_i)$. For scenarios 1 to 8, $\boldsymbol{\nu} = (\nu_0, \nu_1, \xi, \alpha) = (0, 1, \log(2), \log(4))$ and $\boldsymbol{\nu} = (0, 1, \log(2), \log(6))$ for scenarios 9 and 10. Various sampling score models were considered (i.e., weak or moderate Z and S association): Scenarios 1, 3, 5, 7, and 9 set $\boldsymbol{\beta} = (-7, 0.4)$; Scenario 2, 4, 6, 8, and 10 set $\boldsymbol{\beta} = (-7, 0.6)$. The true survival distributions (T^1, T^0) for each scenario were calculated using the distribution of Z and ϵ in the target population. When Z was binary, $S(t|X=1) = P(T > t|X=1, Z=1)P(Z=1) + P(T > t|X=1, Z=0)P(Z=0)$ and $S(t|X=0) = P(T > t|X=0, Z=1)P(Z=1) + P(T > t|X=0, Z=0)P(Z=0)$. When Z was continuous, $S(t|X=1) = \int_Z P(T > t|X=1, Z \leq z)P(Z \leq z)dz$ and $S(t|X=0) = \int_Z P(T > t|X=0, Z \leq z)P(Z \leq z)dz$

Using all 5,000 simulated datasets, the following quantities were computed for each estimator at time $t_j = 3$ years: the bias, which was the difference between the average estimated survival and the true survival, the average standard error, which was the average of the estimated bootstrap standard errors, the empirical standard error, which was the standard deviation of the estimated survival, and empirical coverage probability, which was the proportion of times the 95% confidence interval contained the true survival (Table 5.2 and 5.1). The average trial size ranged from 987 to 1,091. The average standard error of all three estimators was computed using a nonparametric bootstrap with 200 random samples with replacement of the trial and 200 random samples with replacement of the cohort data (Efron and Tibshriani, 1994). To obtain confidence intervals for the survival function in the proper range, confidence intervals were computed using the log-log approach. For scenarios 4 and 10, the estimated survival curves for each estimator were plotted for the first 100 datasets (Figure 5.1(a) to Figure 5.2(b)). For all scenarios, the Kaplan-Meier estimator was biased for both marginal curves. The IPSW estimator was unbiased for both marginal curves in the near infinite target population. The stratified estimator was unbiased for both marginal curves when there was a binary covariate in the sampling score model; however, for some scenarios with a continuous covariate, the stratified estimator was biased and its corresponding confidence intervals had coverage below the nominal level. The average of the bootstrap standard error was approximately equal to the Monte Carlo standard error, supporting the utilization of the bootstrap for estimating the variance of the IPSW estimator

and the stratified estimator.

5.5 Applications

Motivated by Gandhi et al. (2005), the IPSW estimator was employed to generalize results from the ACTG to all people currently living with HIV in the U.S. Although the trials had internal validity, these trial results may not be generalizable to the population of people living with HIV in the U.S. Information on the target population was ascertained from representative cohort studies. Two target populations were considered: all women currently living with HIV in the U.S. and all people currently living with HIV in the U.S. The IPSW estimator was employed to generalize the ACTG results among women using a cohort of women in the WIHS. In a separate analysis, the IPSW estimator was utilized to generalize the trial results using a cohort defined in the Center for AIDS Research (CFAR) Network of Integrated Clinical Systems (CNICS).

5.5.1 ACTG 320

The IPSW estimator was employed to generalize results from ACTG 320 to all people currently living with HIV in the U.S. The ACTG 320 trial examined the safety and efficacy of adding a protease inhibitor (PI) to an HIV treatment regimen with two nucleoside analogues. A total of 1,156 participants were enrolled in ACTG 320 between January 1996 and January 1997 and were recruited from 33 AIDS clinical trial units and 7 National Hemophilia Foundation sites in the U.S. and Puerto Rico (Hammer et al., 1997). 200 women were enrolled in ACTG 320 (Hammer et al., 1997). The baseline characteristics of these women and all participants are shown in Table 5.3 and in Table 5.5, respectively.

WIHS and CNICS were considered to be representative samples of their respective target populations and this analysis only included participants who were HIV-positive, highly active antiretroviral therapy (HAART) naive, and had CD4 cell counts ≤ 200 cells/mm^3 at the previous visit ($m = 493$ and $m = 6{,}158$, respectively). Lab values (i.e., CD4 cell count) were carried forward for up to two years. The WIHS is a prospective, observational, multicenter study of women living with HIV and women at risk for HIV infection in the U.S. (Bacon et al., 2005). A total of 4,129 women (1,065 HIV-uninfected) were enrolled between October 1994 and December 2012 at six U.S. sites. Of the 493 women included in the WIHS sample, 82% were non-white, median age was 40 years, and 37% had a history of injection drug use (IDU). The median CD4 count was 108 cells/mm^3. Table 5.3 displays

the characteristics of the women in the WIHS sample.

The CNICS captures comprehensive and standardized clinical data from point-of-care electronic medical record systems for population-based HIV research (Kitahata et al., 2008). CNICS is considered to be representative of all people living with HIV and in clinical care in the U.S. The CNICS cohort includes over 27,000 HIV-infected adults (at least 18 years of age) engaged in clinical care since January 1, 1995 at eight CFAR sites in the U.S. Of the 6,158 participants included in the CNICS sample, 80% were male, 60% were non-white, median age was 41 years, and 20% had a history of IDU. The median CD4 count was 89 cells/mm^3. Table 5.5 displays the characteristics of participants in the CNICS sample.

The IPSW Estimator using the WIHS Cohort

The IPSW estimator was employed to assess the generalizability of the marginal survival at one year observed among women in the ACTG 320 to all women currently living with HIV in the U.S. Based on Centers for Disease Control (CDC) estimates, the size of the target population was assumed to be 280,000 women (CDC, 2012). Among the 200 women in ACTG 320, 15 (8%) experienced AIDS or death by one year with 7 of those randomized to a regimen with PI and 8 of those randomized to a regimen without a PI. At one year, those randomized to the regimen with a PI had an estimated AIDS-free survival of 0.93 (95% CI = $(0.85, 0.97)$), compared to an estimated AIDS-free survival of 0.90 (95% confidence interval (CI) = $(0.79, 0.95)$) for those randomized to a regimen without a PI. Among women, there was not a statistically significant difference in the risk of AIDS or death at one year between the two treatment groups (RD= -0.03; 95% CI = (-0.11, 0.05) and RR = 0.70; 95% CI = (0.23, 2.09)).

The distributions of baseline age, history of IDU, and CD4 differed between the trial and cohort participants (Table 5.3). Age, race, history of IDU, and CD4 were associated with trial participation (P value < 0.001, P value $= 0.003$, P value < 0.001, and P value $= 0.003$, respectively). Among the 200 women in the trial, CD4 was associated with the time-to-event outcome (P value $= 0.005$), but no covariates were effect modifiers on the difference scale. The estimates of the risk difference varied across levels of all covariates, except history of IDU (Table 5.4).

To estimate the sampling scores, the data from the ACTG trial and WIHS were analyzed together, with $S = 1$ for those in the ACTG trial and $S = 0$ for those in the WIHS cohort. A logistic regression model was fit on the combined trial and weighted cohort data. Cohort participants were inverse

weighted by the size of the cohort divided by the size of the target (i.e., $\hat{r}_i = 493/(280{,}000-200)$) and trial participants were given a weight of $\hat{r}_i = 1$. The outcome was trial participation and the possible covariates were race, age, history of IDU, and baseline CD4. Age and baseline CD4 were modeled as continuous linear variables. Only variables associated with trial participation, the outcome, or treatment effect modifiers were included in the sampling score model, as well as all pairwise interactions. Using the IPSW estimator in equation (5.3), the estimated AIDS-free survival at one year was 0.97 (95% CI = $(0.86, 0.99)$) among those randomized to a regimen with a PI and 0.95 (95% CI = $(0.88, 0.98)$) among those randomized to the regimen without a PI. Among women, there was not a statistically significant difference in the risk of AIDS or death at one year between the treatment groups (RD = -0.02; 95% CI = (-0.08, 0.04) and RR = 0.58; 95% CI = (0.09, 3.84)). Figure 5.3 displays the complement of marginal survival curves estimated using the intention-to-treat (ITT) and IPSW estimators, respectively.

The IPSW Estimator using the CNICS Cohort

The IPSW estimator was employed to assess the generalizability of the marginal survival at one year observed in the ACTG 320 to all people currently living in the U.S. with HIV. Based on CDC estimates, the size of the target population was assumed to be 1.1 million people (CDC, 2012). Among the 1,156 participants in ACTG 320, 96 (8%) died or developed AIDS by one year with 33 of those randomized to a regimen with a PI and 63 of those randomized to a regimen without a PI. At one year, those randomized to the regimen with a PI had an estimated AIDS-free survival of 0.94 (95% CI = $(0.91, 0.95)$), compared to an estimated AIDS-free survival of 0.88 (95% CI = $(0.84, 0.90)$) for those randomized to a regimen without a PI. There was a statistically significant difference in the risk of AIDS or death at one year between the treatment groups (RD = -0.06; 95% CI = (-0.10, -0.02) and RR = 0.51; 95% CI = (0.34, 0.77)).

The distributions of baseline CD4 and IDU history differed between the trial and cohort participants (Table 5.5). Age, race, and CD4 were associated with trial participation (P value < 0.001 for each). Among the trial participants, baseline CD4 was associated with the time-to-event outcome (P value < 0.001) and there was effect modification on the difference scale by age (Comparing those ages 16 to <30 years to those ages 30 to <40 years (P value = 0.03)). Estimates of the risk difference varied across levels of all covariates, except history of IDU (Table 5.6).

To estimate the sampling scores, the data from the ACTG trial and CNICS were analyzed together,

with $S = 1$ for those in the ACTG trial and $S = 0$ for those in the CNICS cohort. A logistic regression model was fit on the combined trial and weighted cohort data. Cohort participants were inverse weighted by the size of the cohort divided by the size of the target (i.e., $\hat{r}_i = 6{,}158/(1{,}100{,}000 - 1{,}156)$) and trial participants were given a weight of $\hat{r}_i = 1$. The outcome was trial participation and the possible covariates were sex, race, age, history of IDU, and baseline CD4. Age and baseline CD4 were modeled as continuous linear variables. Only covariates related to trial participation, the outcome, or treatment effect modifiers were included in the sampling score model, as well as all pairwise interactions. Using the IPSW estimator in equation (5.3), the estimated AIDS-free survival at one year was 0.95 (95% CI = $(0.92, 0.97)$) among those randomized to a regimen with a PI and 0.89 (95% CI = $(0.87, 0.92)$) among those randomized to the regimen without a PI. There was a statistically significant difference in the risk of AIDS or death at one year between the treatment groups (RD = -0.05; 95% CI = (-0.09, -0.01) and RR = 0.52; 95% CI = (0.31, 0.86)). Figure 5.4 displays the complement of marginal survival curves estimated using the ITT and IPSW estimators, respectively.

5.5.2 ACTG A5202

The IPSW estimator was employed to generalize results from ACTG A5202 to all people currently living with HIV in the U.S. The ACTG A5202 trial examined equivalence of abacavir-lamivudine (ABC-3TC) or tenofovir disoproxil fumarate-emtricitabine (TDF-FTC) plus efavirenz or ritonavir-boosted atazanavir. A total of 1,857 participants were enrolled in ACTG A5202 between September 2005 and November 2007 and were recruited from 59 ACTG sites in the U.S. and Puerto Rico (Sax et al., 2009, 2011). 322 women were enrolled in ACTG A5202 (Sax et al., 2009, 2011). The baseline characteristics are shown in Table 5.7 among women and Table 5.9 among all participants.

WIHS and CNICS were considered to be representative samples of their respective target populations and this analysis only included participants who were HIV-positive, antiretroviral (ART) naive, and had viral load > 1,000 copies/ml at the previous visit ($m = 1{,}012$ and $m = 12{,}302$, respectively). Of the $1{,}012$ women included in the WIHS sample, 83% were non-white, median age was 39 years, 38% had a history of IDU, 35% had hepatitis B or C, and 37% were diagnosed with AIDS. The median CD4 count was 290 cells/mm^3 and the median log10 viral load was 4.61 copies/ml. Table 5.7 displays the characteristics of the women in the WIHS sample. Of the 12,302 participants included in the CNICS sample, 82% were male, 55% were non-white, median age was 39 years, 17% had a history of IDU, 18% had hepatitis B or C, and 23% were diagnosed with AIDS. The median CD4 count was 271 cells/mm^3 and the median log10 viral load was 4.64 copies/ml. Table 5.9 displays the characteristics

of participants in the CNICS sample.

The IPSW Estimator using the WIHS Cohort

The IPSW estimator was employed to assess the generalizability of the probability of virologic failure at week 48 for each treatment group (ABC-3TC vs. TDF-FTC) observed among women in the ACTG A5202 to all women currently living with HIV in the U.S. Based on CDC estimates, the size of the target population was assumed to be 280,000 women (CDC, 2012). The outcome analyzed was time to virologic failure (defined as confirmed HIV-1 RNA level $\geq$ 1000 copies per milliliter at or after 16 weeks and before 24 weeks, or $\geq$ 200 copies per milliliter at or after 24 weeks). Among the 322 women in A5202, 49 (15%) experienced virologic failure by week 48 with 26 of those randomized to ABC-3TC and 23 of those randomized to TDF-FTC. The estimated probability of remaining free of virologic failure beyond week 48 was 0.86 (95% CI = (0.80, 0.91)) among those on ABC-3TC, compared to 0.91 (95% CI = (0.84, 0.94)) among those on TDF-FTC. Among women, there was not a statistically significant difference between the treatment groups (RD = 0.04; 95% CI = (-0.03, 0.11) and RR = 1.43; 95% CI = (0.73, 2.78)).

The distributions of baseline CD4, history of IDU, hepatitis, and AIDS differed between the trial and cohort participants (Table 5.7). Age, history of IDU, hepatitis, AIDS diagnosis, baseline CD4, and viral load were associated with trial participation (P value = 0.004, P value $<$ 0.001, P value = 0.003, P value $<$ 0.001, P value $<$ 0.001, and P value $<$ 0.001, respectively). Among the trial participants, age, AIDS diagnosis, and baseline CD4 were associated with the time-to-event outcome (P value = 0.04, P value = 0.04, and P value = 0.03, respectively). There was no effect modification on the difference scale by any of the covariates. The estimates of the risk difference varied across levels of all covariates, except history of IDU (Table 5.8).

To estimate the sampling scores, the data from the ACTG trial and WIHS were analyzed together, with $S = 1$ for those in the ACTG trial and $S = 0$ for those in the WIHS cohort. A logistic regression model was fit on the combined trial and weighted cohort data. Cohort participants were inverse weighted by the size of the cohort divided by the size of the target (i.e., $\hat{r}_i = 1,012/(280{,}000-322)$) and trial participants were given a weight of $\hat{r}_i = 1$. The outcome was trial participation and the possible covariates were race, age, history of IDU, hepatitis B/C, AIDS diagnosis, and baseline CD4, and baseline log10 viral load. Only variables associated with either trial participation, the outcome or effect modifiers were included in the sampling score model, as well as all pairwise interactions. Age,

baseline CD4, and baseline log10 viral load were modeled as continuous linear variables. Because hepatitis B/C and history of IDU were correlated ($r = 0.69$), history of IDU was excluded from the sampling score model.

Using the IPSW estimator in equation (5.3), the marginal survival estimates of virologic failure at week 48 for each treatment group were computed. The estimated probability of remaining free of virologic failure beyond week 48 was 0.82 (95% CI = (0.58, 0.93)) among those on ABC-3TC, compared to 0.90 (95% CI = (0.79, 0.96)) among those on TDF-FTC. Among women, there was not a statistically significant difference between the treatment groups (RD = 0.08; 95% CI = (-0.12, 0.28) and RR = 1.80; 95% CI = (0.54, 6.05)). Figure 5.5 displays the complement of marginal survival curves estimated using the ITT and IPSW estimators, respectively.

The IPSW Estimator using the CNICS Cohort

The IPSW estimator was employed to assess the generalizability of the marginal probability of virologic failure at week 48 between the ABC-3TC and TDF-FTC treatment groups observed in the ACTG A5202 to all people currently living with HIV in the U.S. Based on CDC estimates, the size of the target population was assumed to be 1.1 million people (CDC, 2012). Among the 1,857 participants in A5202, 219 (12%) experienced virologic failure by week 48 with 131 of those randomized to ABC-3TC and 88 of those randomized to TDF-FTC. The estimated probability of remaining free of virologic failure beyond week 48 was 0.88 (95% CI = (0.85, 0.90)) among those on ABC-3TC, compared to 0.93 (95% CI = (0.92, 0.95)) among those on TDF-FTC. Among all participants, there was a statistically significant difference between the treatment groups (RD = 0.05; 95% CI = (0.03, 0.08) and RR = 1.83; 95% CI = (1.33, 2.52)).

The distributions of history of IDU, hepatitis, AIDS diagnosis, and baseline CD4 differed between the trial and cohort participants (Table 5.9). Race, AIDS diagnosis, hepatitis B/C, history of IDU, CD4, and viral load were associated with trial participation (P value < 0.001 for each). Among the trial participants, age, race, hepatitis B/C, AIDS diagnosis, baseline CD4, and viral load were associated with the outcome (P value < 0.001, P value < 0.001, P value $= 0.002$, P value < 0.001, P value < 0.001, and P value $= 0.001$, respectively). There was effect modification on the difference scale by CD4 (100 to 200 cells/mm^3 versus < 50 cells/mm^3 (P value $= 0.05$)) and viral load (100,000 to 300,000 copies/ml versus $< 50{,}000$ copies/ml (P value $= 0.05$) and $> 500,000$ copies/ml versus $< 50,000$ copies/ml (P value $= 0.04$)). The estimates of the risk difference varied across levels of all

covariates (Table 5.10).

To estimate the sampling scores, the data from the ACTG trial and CNICS were analyzed together, with $S = 1$ for those in the ACTG trial and $S = 0$ for those in the CNICS cohort. A logistic regression model was fit on the combined trial and weighted cohort data. Cohort participants were inverse weighted by the size of the cohort divided by the size of the target (i.e., $\hat{r}_i = 12,302/(280{,}000-1,857)$) and trial participants were given a weight of $\hat{r}_i = 1$. The outcome was trial participation and possible covariates were sex, race, age, history of IDU, hepatitis B/C, AIDS diagnosis, baseline CD4 and baseline log10 viral load. Age, baseline CD4, and baseline log10 viral load were modeled as continuous linear variables. Only covariates associated with trial participation, the outcome, or effect modifiers were included in the sampling score model, as well as all pairwise interactions.

Using the IPSW estimator in equation (5.3), the marginal survival estimates of virologic failure at week 48 for each regimen group were computed. The estimated probability of remaining free of virologic failure beyond week 48 was 0.87 (95% CI = (0.84, 0.90)) among those on ABC-3TC, compared to 0.93 (95% CI = (0.91, 0.95)) among those on TDF-FTC. Among all participants, there was a statistically significant difference between the treatment groups (RD = 0.06; 95% CI = (0.02, 0.10) and RR = 1.83; 95% CI = (1.23, 2.72)). Figure 5.6 displays the complement of marginal survival curves estimated using the ITT and IPSW estimators, respectively.

5.6 Discussion

Following Cole and Stuart (2010) and Buchanan et al. (In preparation), we considered an estimator using inverse probability of sampling weights to generalize results for right-censored data in a randomized trial to a specified target population. We use the term generalizability to describe the degree to which an internally valid measure of effect estimated in a sample from one population would change if the trial were conducted in a different target population. The IPSW estimator compares the outcome in the target population if (possibly contrary to fact) everyone had been randomized to treatment with the outcome in the target population if (possibly contrary to fact) everyone had been randomized to control. Simulation results were provided to compare this estimator to an unweighted KM estimator and a stratified estimator. The average standard error was computed using a non-parametric bootstrap. In the following, we discuss some recent work addressing generalizability and explore caveats of this approach.

In the illustrative example, the IPSW estimator was employed to generalize results for right-

censored data in the ACTG to all people currently living with HIV in the U.S. For both trials, the risk difference estimated with the ITT was comparable to the risk difference estimated with the IPSW. Thus, the ACTG 320 results appear to be generalizable to all people living with HIV in the U.S. Among women in A5202, the risk difference doubled when calculated with the IPSW estimator, as compared to the ITT estimator (RD = 0.08 vs. RD = 0.04, respectively). The marginal event rates typically had a larger change in magnitude than the relative measures (i.e., risk difference and risk ratio). Results were not sensitive to the specification of the size of the target population.

In a previous paper by Cole and Stuart (2010), the ACTG 320 results were generalized to a target population of all people infected with HIV in the U.S. in 2006, as estimated by the CDC. This paper continues that effort by empirically demonstrating the reasonable performance of the bootstrap standard error, as well using cohort data to obtain richer information on the target. This paper continues the effort in Tipton (2013) by extending their estimator for right-censored data and evaluating the performance of the nonparametric bootstrap standard error for this estimator.

When applying this method, the analysis is subject to the following considerations. The nonparametric bootstrap standard error was used to estimate the variance. Additional research to demonstrate the large sample properties of this estimator could allow for a closed-form expression for the variance. The absence of unmeasured covariates associated with the trial participation and treatment effect modifiers is an untestable assumption. Treatment compliance issues were ignored in this method; however, this issue could be considered in analyses. The sampling score model was assumed to be correct; however, this is not guaranteed in practice. Weighted logistic regression was used as an approach to consistently estimate the parameters of the logistic regression model (i.e., the intercept); however, other approaches may be possible. Future research could extend the IPSW estimator to a doubly-robust estimator (Bang and Robins, 2005). This method could also be extended to accommodate the presence of interference.

In conclusion, we considered an inverse probability of sampling weighted estimator for estimating marginal survival curves in the target population. The bootstrap standard error appears to be a reasonable estimator of the variance for the IPSW estimator. Quantitative methods for generalizability is a growing field of statistical research and methods for right-censored data are essential to address questions in infectious disease research. We hope that this paper will be useful for implementation of these methods and increasing interest in quantitative methods for generalizability of trial results with right-censored data.

Acknowledgments

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

Table 5.1: Summary of Monte Carlo results for estimators of the marginal survival curves in the target population with a right-censored outcome for $X = 1$ with $m = 4{,}000$ and $n \approx 1{,}000$ in 5,000 simulated datasets. Scenarios are described in the text. Bias, average standard error (ASE) ($\times$ 100), empirical standard error (ESE) ($\times$ 100), and 95% empirical coverage probability (ECP) at $t_j = 3$ are reported for each estimator. $\hat{\mathfrak{S}}^1_{IPSW}(t_j)$ is the inverse probability of sampling weighted estimator, $\hat{\mathfrak{S}}^1_S(t_j)$ is the stratified estimator, and $\hat{\mathfrak{S}}^1_{KM}(t_j)$ is the Kaplan-Meier estimator. For scenarios 1 to 4, $\mathfrak{S}^1(t_j) = 0.47$, for scenarios 5 to 8, $\mathfrak{S}^1(t_j) = 0.44$, and, for scenarios 9 and 10, $\mathfrak{S}^1(t_j) = 0.45$.

	Cov	(β,e^α)	Cens	$\hat{\mathfrak{S}}^1_{IPSW}(t_j)$				$\hat{\mathfrak{S}}^1_S(t_j)$				$\hat{\mathfrak{S}}^1_{KM}(t_j)$			
				Bias	ASE	ESE	ECP	Bias	ASE	ESE	ECP	Bias	ASE	ESE	ECP
1	Bin	(0.4,4)	Adm	9e-5	2.2	2.2	0.95	6e-5	2.1	2.1	0.95	0.05	2.2	2.3	0.49
2	Bin	(0.6,4)	Adm	-2e-4	2.2	2.2	0.94	-3e-4	2.0	2.1	0.94	0.07	2.2	2.2	0.10
3	Bin	(0.4,4)	Ind	-1e-3	3.6	3.6	0.95	-1e-3	3.3	3.3	0.95	0.04	3.5	3.5	0.77
4	Bin	(0.6,4)	Ind	1e-3	3.5	3.6	0.95	6e-4	3.3	3.4	0.95	0.07	3.4	3.4	0.47
5	Con	(0.4,4)	Adm	4e-4	2.1	2.2	0.95	5e-3	1.6	1.6	0.93	0.15	2.2	2.2	<0.01
6	Con	(0.6,4)	Adm	5e-4	2.3	2.3	0.95	8e-3	1.6	1.6	0.93	0.22	2.0	2.0	<0.01
7	Con	(0.4,4)	Ind	8e-4	3.1	3.1	0.95	6e-3	2.5	2.4	0.94	0.15	3.1	3.1	<0.01
8	Con	(0.6,4)	Ind	2e-3	3.2	3.2	0.95	0.01	2.5	2.5	0.93	0.22	2.9	2.9	<0.01
9	Con	(0.4,6)	Ind	1e-3	3.0	3.0	0.95	7e-3	2.3	2.3	0.95	0.15	3.0	3.0	<0.01
10	Con	(0.6,6)	Ind	2e-3	3.2	3.2	0.95	0.01	2.4	2.3	0.93	0.22	2.8	2.9	<0.01

Table 5.2: Summary of Monte Carlo results for estimators of the marginal survival curves in the target population with a right-censored outcome for $X = 0$ with $m = 4{,}000$ and $n \approx 1{,}000$ in 5,000 simulated datasets. Scenarios are described in the text. Bias, average standard error (ASE) ($\times$ 100), empirical standard error (ESE) ($\times$ 100), and 95% empirical coverage probability (ECP) at $t_j = 3$ are reported for each estimator. $\hat{\mathfrak{S}}^0_{IPSW}(t_j)$ is the inverse probability of sampling weighted estimator, $\hat{\mathfrak{S}}^0_S(t_j)$ is the stratified estimator, and $\hat{\mathfrak{S}}^0_{KM}(t_j)$ is the Kaplan-Meier estimator. For scenarios 1 to 4, $\mathfrak{S}^0(t_j) = 0.20$ and, for scenarios 5 to 10, $\mathfrak{S}^0(t_j) = 0.22$.

	Cov	(β,e^α)	Cens	$\hat{\mathfrak{S}}^0_{IPSW}(t_j)$				$\hat{\mathfrak{S}}^0_S(t_j)$				$\hat{\mathfrak{S}}^0_{KM}(t_j)$			
				Bias	ASE	ESE	ECP	Bias	ASE	ESE	ECP	Bias	ASE	ESE	ECP
1	Bin	(0.4,4)	Adm	-2e-4	1.7	1.7	0.95	-2e-4	1.7	1.7	0.95	0.02	1.9	1.9	0.76
2	Bin	(0.6,4)	Adm	3e-4	1.7	1.7	0.95	2e-4	1.7	1.7	0.95	0.04	1.9	1.8	0.47
3	Bin	(0.4,4)	Ind	3e-4	2.9	2.9	0.95	6e-5	2.8	2.8	0.96	0.02	3.0	3.0	0.87
4	Bin	(0.6,4)	Ind	6e-4	2.8	2.8	0.95	6e-4	2.7	2.7	0.95	0.04	3.0	3.0	0.76
5	Con	(0.4,4)	Adm	2e-4	1.7	1.7	0.95	9e-3	1.5	1.5	0.92	0.09	2.1	2.1	<0.01
6	Con	(0.6,4)	Adm	4e-4	1.7	1.7	0.95	0.01	1.5	1.5	0.86	0.14	2.1	2.1	<0.01
7	Con	(0.4,4)	Ind	6e-4	2.7	2.7	0.95	0.01	2.6	2.6	0.93	0.09	3.3	3.3	0.17
8	Con	(0.6,4)	Ind	2e-4	2.6	2.6	0.95	0.02	2.5	2.5	0.91	0.14	3.2	3.2	<0.01
9	Con	(0.4,6)	Ind	6e-4	2.7	2.7	0.95	0.01	2.6	2.6	0.93	0.09	3.3	3.3	0.17
10	Con	(0.6,6)	Ind	2e-4	2.6	2.6	0.95	0.02	2.5	2.5	0.91	0.14	3.2	3.2	<0.01

Table 5.3: Characteristics of 493 women in the Women's Interagency HIV Study (WIHS) who were HIV-positive, HAART naive, and had CD4 cell count $\leq$ 200 cells/mm^3 at the previous visit and 200 women at baseline in AIDS Clinical Trials Group (ACTG) 320 by treatment group (with and without a protease inhibitor (PI))

Variable	**WIHS** (m = 493)	**ACTG 320** (n = 200)	**ACTG 320 Protease Inhibitor (PI)** (n_1 = 106)	**ACTG 320 No Protease Inhibitor (PI)** (n_0 = 94)
Race or ethnic group - no. (%)				
White, non-Hispanic	87 (18)	61 (31)	26 (25)	35 (37)
Black, non-Hispanic	272 (55)	95 (48)	54 (51)	41 (44)
Hispanic	124 (25)	42 (21)	25 (24)	17 (18)
Asian/Other	10 (2)	2 (1)	1 (1)	1 (1)
Median age - yr (Q1-Q3)	40 (35-45)	36 (30-42)	37 (31-42)	36 (30-43)
Age group - no. (%)				
16-<30 yr	35 (7)	46 (23)	22 (21)	24 (26)
30-<40 yr	211 (43)	88 (44)	48 (45)	40 (43)
40-<50 yr	196 (40)	53 (27)	27 (26)	26 (28)
$\geq$50 yr	51 (10)	13 (7)	9 (9)	4 (4)
Injection drug use - no. (%)	180 (37)	36 (18)	24 (23)	12 (13)
Median CD4 count (Q1-Q3)	108 (41-172)	82 (26-139)	93 (29-139)	70 (23-138)
Baseline CD4 count - no. (%)				
<50 cells/mm^3	148 (30)	72 (36)	35 (33)	37 (39)
50-<100 cells/mm^3	83 (17)	43 (22)	22 (21)	21 (22)
100-<200 cells/mm^3	182 (37)	73 (37)	44 (42)	29 (31)
$\geq$200 cells/mm^3	80 (16)	12 (6)	5 (5)	7 (7)

Table 5.4: Difference in the estimated risk of AIDS or death at one year between treatment groups (protease inhibitor (PI) vs. no PI) for each level of the covariates among 200 women in AIDS Clinical Trials Group 320 with corresponding 95% confidence interval (CI)

Variable	**AIDS or Death at One Year** Risk Difference (95 % CI)
Race or ethnic group	
White, non-Hispanic	0.01 (-0.25, 0.27)
Black, non-Hispanic	0.04 (-0.21, 0.13)
Hispanic	-[a]
Asian/Other	-[a]
Age group	
18-<30 yr	0.15 (-0.10, 0.39)
30-<40 yr	-0.09 (-0.26, 0.07)
40-<50 yr	0 (-0.20, 0.20)
≥50 yr	-[b]
Injection drug use	
Yes	-[c]
No	-0.04 (-0.17, 0.10)
Baseline CD4 count	
<50 cells/mm^3	0 (-0.26, 0.24)
50-<100 cells/mm^3	-0.05 (-0.26, 0.16)
100-<200 cells/mm^3	-0.02 (-0.16, 0.12)
≥200 cells/mm^3	-[a]

[a]No events in this stratum.
[b]No events in this stratum among those randomized to a PI.
[c]No events in this stratum among those randomized to no PI.

Table 5.5: Characteristics of 6,158 participants in the CFAR Network of Integrated Clinical Systems (CNICS) who were HIV-positive, HAART naive, and had CD4 cell count ≤ 200 cells/mm^3 at the previous visit and 1,156 participants at baseline in AIDS Clinical Trials Group (ACTG) 320 by treatment group (with and without a protease inhibitor (PI))

Variable	**CNICS** (m = 6,158)	**ACTG 320** (n = 1,156)	**ACTG 320 Protease Inhibitor (PI)** ($n_1 = 577$)	**ACTG 320 No Protease Inhibitor (PI)** ($n_0 = 579$)
Male sex - no. (%)	4,909 (80)	956 (83)	471 (82)	485 (84)
Race or ethnic group - no. (%)				
White, non-Hispanic	2,436 (40)	598 (52)	303 (53)	295 (51)
Black, non-Hispanic	2,690 (44)	328 (28)	163 (28)	165 (29)
Hispanic	734 (12)	205 (18)	99 (17)	106 (18)
Asian/Other	298 (5)	25 (2)	12 (2)	13 (2)
Median age - yr (Q1-Q3)	41 (34-47)	38 (33-44)	38 (33-44)	38 (33-44)
Age group - no. (%)				
16-<30 yr	714 (12)	142 (12)	69 (12)	73 (13)
30-<40 yr	2,108 (34)	536 (47)	272 (47)	264 (46)
40-<50 yr	2,315 (38)	350 (30)	169 (29)	181 (31)
≥50 yr	1,021 (17)	128 (11)	67 (12)	61 (11)
Injection drug use - no. (%)	1,241 (20)	184 (16)	91 (16)	93 (16)
Median CD4 count (Q1-Q3)	89 (27-172)	75 (23-137)	80 (24-138)	70 (23-135)
Baseline CD4 count - no. (%)[a]				
<50 cells/mm^3	2,237 (36)	453 (39)	219 (38)	234 (41)
50-<100 cells/mm^3	1,047 (17)	248 (22)	118 (20)	130 (23)
100-<200 cells/mm^3	1,818 (30)	372 (32)	200 (35)	172 (30)
≥200 cells/mm^3	1,056 (17)	82 (7)	40 (7)	42 (7)

[a]One A5202 participant missing baseline CD4 cell count.

Table 5.6: Difference in the estimated risk of AIDS or death at one year between treatment groups (protease inhibitor (PI) vs. no PI) for each level of the covariates among 1,156 participants in AIDS Clinical Trials Group (ACTG) 320 with corresponding 95% confidence interval (CI)

Variable	**AIDS or Death at One Year** Risk Difference (95 % CI)
Sex	
Male	-0.07 (-0.12, -0.01)
Female	-0.03 (-0.15, 0.09)
Race or ethnic group	
White, non-Hispanic	-0.05 (-0.12, 0.02)
Black, non-Hispanic	-0.08 (-0.17, 0.01)
Hispanic	-0.07 (-0.21, 0.07)
Asian/Other	-0.11 (-0.57, 0.35)
Age group	
16-<30 yr	0.04 (-0.09, 0.16)
30-<40 yr	-0.12 (-0.19, -0.05)
40-<50 yr	-0.02 (-0.11, 0.08)
≥50 yr	-0.05 (-0.24, 0.13)
Injection drug use	
Yes	-0.06 (-0.17, 0.04)
No	-0.06 (-0.12, -0.002)
Baseline CD4 count	
<50 cells/mm^3	-0.09 (-0.19, 0.01)
50-<100 cells/mm^3	-0.06 (-0.17, 0.04)
100-<200 cells/mm^3	-0.03 (-0.08, 0.02)
≥200 cells/mm^3	-[a]

[a]No events in this stratum among those randomized to no PI.

Table 5.7: Characteristics of 1,012 women in the Women's Interagency HIV Study (WIHS) who were HIV-positive, ART naive, and had viral load > 1000 copies/ml at the previous visit and 322 women at baseline in AIDS Clinical Trials Group (ACTG) A5202 by treatment group (abacavir-lamivudine (ABC-3TC) vs. tenofovir disoproxil fumarate-emtricitabine (TDF-FTC))

Variable	**WIHS** (m = 1,012)	**ACTG A5202** (n = 322)	**ACTG A5202 ABC-3TC** (n_1 = 173)	**ACTG A5202 TDF FTC** (n_0 = 149)
Race or ethnic group - no. (%)				
White, non-Hispanic	171 (17)	57 (18)	30 (17)	27 (18)
Black, non-Hispanic	586 (58)	172 (53)	94 (54)	78 (52)
Hispanic	222 (22)	82 (26)	42 (24)	40 (27)
Asian/Other	33 (3)	11 (3)	7 (4)	4 (3)
Median age - yr (Q1-Q3)	39 (33-44)	39 (31-46)	39 (31-46)	39 (31-46)
Age group - no. (%)				
16-<30 yr	123 (12)	57 (18)	30 (17)	27 (18)
30-<40 yr	435 (43)	110 (34)	62 (36)	48 (32)
40-<50 yr	345 (34)	107 (33)	54 (31)	53 (36)
≥50 yr	109 (11)	48 (15)	27 (16)	21 (14)
Injection drug use - no. (%)	388 (38)	18 (6)	9 (5)	9 (6)
Hepatitis B/C - no. (%)	356 (35)	25 (8)	14 (8)	11 (7)
AIDS diagnosis - no. (%)	373 (37)	62 (19)	39 (23)	23 (15)
CD4 count - no. (%)				
<50 cells/mm^3	102 (10)	61 (19)	38 (22)	23 (15)
50-<100 cells/mm^3	61 (6)	24 (7)	15 (8)	9 (6)
100-<200 cells/mm^3	162 (16)	55 (17)	28 (16)	27 (18)
200-<350 cells/mm^3	295 (29)	130 (40)	65 (38)	65 (44)
≥350 cells/mm^3	392 (39)	52 (16)	4 (6)	25 (17)
Median CD4 count (Q1-Q3)	290 (162-423)	226 (87-313)	209 (60-308)	249 (129-316)
Viral load - no. (%)				
<50,000 cp/ml	552 (55)	187 (58)	93 (54)	94 (63)
50,000-<100,000 cp/ml	144 (14)	62 (19)	33 (19)	29 (20)
100,000-<300,000 cp/ml	193 (19)	38 (12)	24 (14)	14 (9)
300,000-<500,000 cp/ml	55 (5)	9 (3)	6 (3)	3 (2)
≥500,000 cp/ml	68 (7)	26 (8)	17 (10)	9 (6)
Median log10 viral load (Q1-Q3)	4.61 (4.04-5.11)	4.58 (4.07-4.93)	4.62 (4.10-5.09)	4.55 (4.04-4.86)

Table 5.8: Difference in the estimated risk of virolgic failure at week 48 between treatment groups (abacavir-lamivudine (ABC-3TC) vs. tenofovir disoproxil fumarate-emtricitabine (TDF-FTC)) for each level of the covariates among 322 women in AIDS Clinical Trials Group A5202 with corresponding 95% confidence interval (CI)

Variable	**Virologic Failure at Week 48** Risk Difference (95 % CI)
Race or ethnic group	
White, non-Hispanic	-[a]
Black, non-Hispanic	0.03 (-0.13, 0.19)
Hispanic	-0.06 (-0.19, 0.08)
Asian/Other	-[b]
Age group	
18-<30 yr	0.07 (-0.13, 0.26)
30-<40 yr	0.03 (-0.16, 0.22)
40-<50 yr	0.10 (-0.10, 0.30)
≥50 yr	-[c]
Injection drug use	
Yes	-[c]
No	0.04 (-0.07, 0.15)
Hepatitis B/C	
Yes	-0.02 (-0.45, 0.41)
No	0.04 (-0.06, 0.15)
AIDS diagnosis	
Yes	0.10 (-0.16, 0.37)
No	0.02 (-0.09, 0.13)
CD4 count	
<50 cells/mm^3	0.11 (-0.16, 0.38)
50-<100 cells/mm^3	-0.14 (-0.66, 0.39)
100-<200 cells/mm^3	0.00 (-0.15, 0.16)
200-<350 cells/mm^3	0.01 (-0.15, 0.16)
≥350 cells/mm^3	0.11 (-0.22, 0.43)
Viral load	
<50,000 cp/ml	0.06 (-0.06, 0.18)
50,000-<100,000 cp/ml	-[d]
100,000-<300,000 cp/ml	0.16 (-0.27, 0.60)
300,000-<500,000 cp/ml	0 (-1.07, 1.07)
≥500,000 cp/ml	-[b]

[a]No events before week 48 in the TDF-FTC arm.
[b]No events in the TDF-FTC arm.
[c]No events in the ABC-3TC arm.
[d]No events before week 48 in the ABC-3TC arm.

Table 5.9: Characteristics of 12,302 participants in the CFAR Network of Integrated Clinical Systems (CNICS) who were HIV-positive, ART naive, and had viral load > 1000 copies/ml at the previous visit and 1,857 participants at baseline in AIDS Clinical Trials Group (ACTG) A5202 by treatment group (abacavir-lamivudine (ABC-3TC) vs. tenofovir disoproxil fumarate-emtricitabine (TDF-FTC))

Variable	**CNICS** (m = 12,302)	**ACTG A5202** (n =1,857)	**ACTG A5202 ABC-3TC** (n_1 = 928)	**ACTG A5202 TDF FTC** (n_0 = 929)
Male sex - no. (%)	10,063 (82)	1,535 (83)	755 (81)	780 (84)
Race or ethnic group[a] - no. (%)				
White, non-Hispanic	5,567 (45)	746 (46)	363 (39)	383 (41)
Black, non-Hispanic	4,682 (38)	615 (33)	317 (34)	298 (32)
Hispanic	1,420 (12)	429 (23)	214 (23)	215 (23)
Asian/Other	633 (5)	62 (3)	31 (3)	31 (3)
Median age - yr (Q1-Q3)	39 (31-46)	38 (31-45)	38 (30-45)	39 (31-45)
Age group - no. (%)				
16-<30 yr	2,454 (20)	404 (22)	201 (22)	203 (22)
30-<40 yr	4,225 (34)	625 (34)	335 (36)	290 (31)
40-<50 yr	3,896 (32)	573 (31)	273 (29)	300 (32)
≥50 yr	1,727 (14)	255 (14)	119 (13)	136 (15)
Injection drug use - no. (%)	2,042 (17)	162 (9)	77 (8)	85 (9)
Hepatitis B/C - no. (%)	2,245 (18)	165 (9)	75 (8)	90 (10)
AIDS diagnosis - no. (%)	2,834 (23)	312 (17)	172 (19)	140 (15)
CD4 count- no. (%)[b] - no. (%)				
<50 cells/mm^3	2,000 (16)	339 (18)	176 (19)	163 (18)
50-<100 cells/mm^3	920 (7)	150 (8)	74 (8)	76 (8)
100-<200 cells/mm^3	1,692 (14)	311 (17)	159 (17)	152 (16)
200-<350 cells/mm^3	3,262 (27)	656 (35)	312 (34)	344 (37)
≥350 cells/mm^3	4,428 (36)	400 (22)	207 (22)	193 (21)
Median CD4 count (Q1-Q3)	271 (109-427)	230 (90-334)	229 (84-338)	230 (96-330)
Viral load - no. (%)				
<50,000 cp/ml	6,450 (52)	1,000 (54)	492 (53)	508 (55)
50,000-<100,000 cp/ml	1,861 (15)	391 (21)	196 (21)	195 (21)
100,000-<300,000 cp/ml	2,232 (18)	203 (11)	106 (11)	97 (10)
300,000-<500,000 cp/ml	744 (6)	72 (4)	38 (4)	34 (4)
≥500,000 cp/ml	1,015 (8)	191 (10)	96 (10)	95 (10)
Median log10 viral load (Q1-Q3)	4.64 (3.95-5.18)	4.66 (4.33-5.01)	4.66 (4.31-5.06)	4.65 (4.34-4.96)

[a]Five A5202 participants were missing race.
[b]One A5202 participant was missing CD4.

Table 5.10: Difference in the estimated risk of virolgic failure at week 48 between treatment groups (abacavir-lamivudine (ABC-3TC) vs. tenofovir disoproxil fumarate-emtricitabine (TDF-FTC)) for each level of the covariates among 1,857 participants in ACTG A5202 with corresponding 95% confidence interval (CI)

Variable	**Virologic Failure at Week 48** Risk Difference (95 % CI)
Sex	
Male	0.06 (0.01, 0.10)
Female	0.04 (-0.06, 0.14)
Race or ethnic group	
White, non-Hispanic	0.07 (0.02, 0.13)
Black, non-Hispanic	0.04 (-0.04, 0.12)
Hispanic	0.01 (-0.06, 0.08)
Asian/Other	-[a]
Age group	
18-<30 yr	0.09 (0.004, 0.17)
30-<40 yr	0.08 (0.004, 0.15)
40-<50 yr	0.03 (0.04, 0.10)
≥50 yr	-0.02 (-0.10, 0.07)
Injection drug use	
Yes	0.11 (-0.03, 0.24)
No	0.05 (0.01, 0.09)
Hepatitis B/C	
Yes	-0.003 (-0.16, 0.16)
No	0.06 (0.02, 0.10)
AIDS diagnosis	
Yes	0.12 (0.002, 0.23)
No	0.04 (-0.002, 0.08)
CD4 count	
<50 cells/mm^3	0.15 (0.03, 0.27)
50-<100 cells/mm^3	0.11 (-0.03, 0.26)
100-<200 cells/mm^3	0.01 (-0.08, 0.09)
200-<350 cells/mm^3	0.03 (-0.02, 0.08)
≥350 cells/mm^3	0.02 (-0.06, 0.11)
Viral load	
<50,000 cp/ml	0.02 (-0.03, 0.06)
50,000-<100,000 cp/ml	0.02 (-0.06, 0.10)
100,000-<300,000 cp/ml	0.16 (0.02, 0.30)
300,000-<500,000 cp/ml	0.14 (-0.11, 0.39)
≥500,000 cp/ml	0.18 (0.03, 0.33)

[a]No events in the TDF-FTC arm.

Table 5.11: Results for the risk difference of the time-to-event outcomes with corresponding 95% confidence intervals in two AIDS Clinical Trials Group studies, where the sampling score model included variables associated with trial participation, the outcome, or effect modifiers (with a linear term for continuous variables) and all pairwise interactions

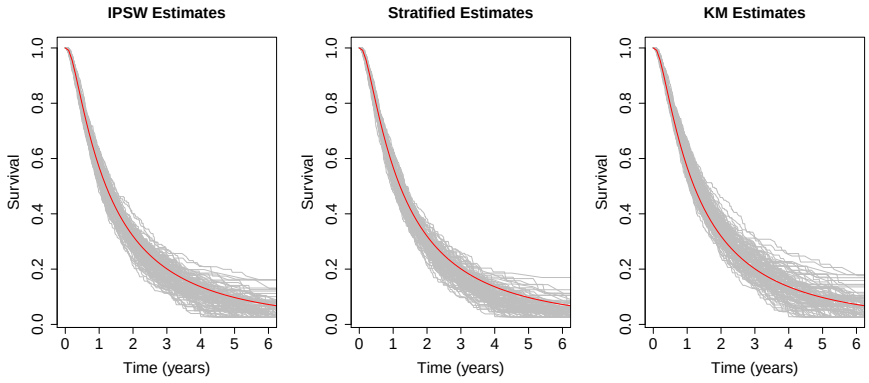
Cohort	**Trial**	**ITT**	**IPSW**	**Stratified**
WIHS	320[a]	-0.03 (-0.11, 0.05)	-0.02 (-0.08, 0.04)	-0.01 (-0.12, 0.09)
WIHS	A5202[b]	0.04 (-0.03, 0.11)	0.08 (-0.12, 0.28)	0.12 (-0.10, 0.26)
CNICS	320	-0.06 (-0.10, -0.02)	-0.05 (-0.09, -0.01)	-0.06 (-0.09, -0.02)
CNICS	A5202	0.05 (0.03, 0.08)	0.06 (0.02, 0.10)	0.07 (0.02, 0.10)

[a]For 320, the treatment contrast was PI vs. no PI.
[b]For A5202, the treatment contrast was ABC-3TC vs. TDF-FTC.

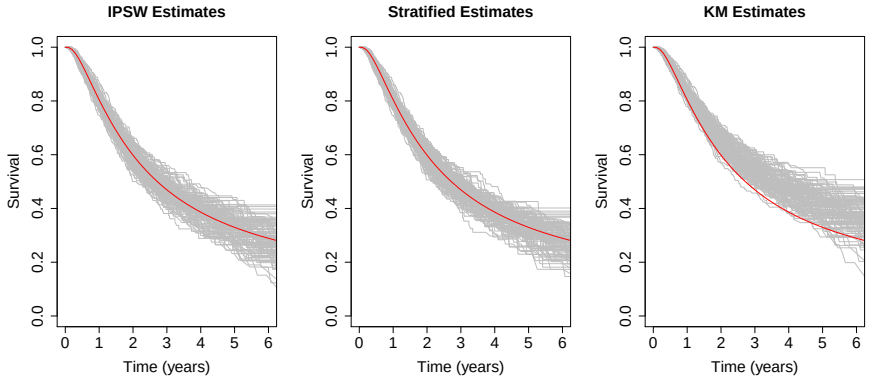
Table 5.12: Results for the risk ratio of the time-to-event outcomes with corresponding 95% confidence intervals in two AIDS Clinical Trials Group studies, where the sampling score model included variables associated with trial participation, the outcome, or effect modifiers (with a linear term for continuous variables) and all pairwise interactions

Cohort	**Trial**	**ITT**	**IPSW**	**Stratified**
WIHS	320[a]	0.70 (0.23, 2.09)	0.58 (0.09, 3.84)	0.85 (0.12, 2.84)
WIHS	A5202[b]	1.43 (0.73, 2.78)	1.80 (0.54, 6.05)	2.32 (0.54, 6.00)
CNICS	320	0.51 (0.34, 0.77)	0.52 (0.31, 0.86)	0.48 (0.32, 0.85)
CNICS	A5202	1.83 (1.33, 2.52)	1.83 (1.23, 2.72)	2.02 (1.16, 2.88)

[a]For 320, the treatment contrast was PI vs. no PI.
[b]For A5202, the treatment contrast was ABC-3TC vs. TDF-FTC.

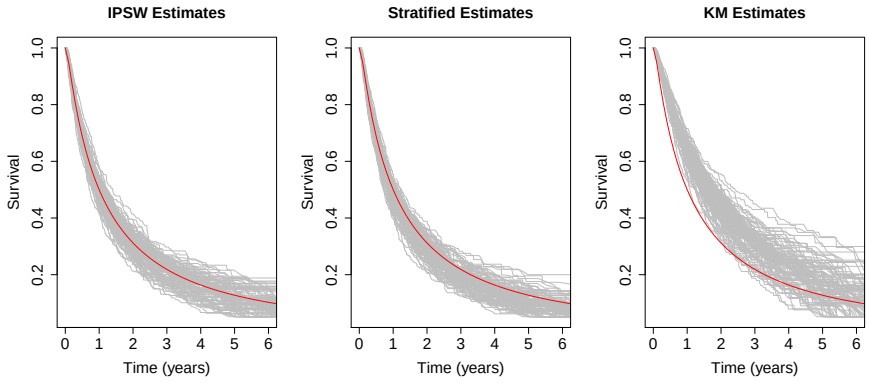


(a) Simulation results with a binary covariate for $X = 0$

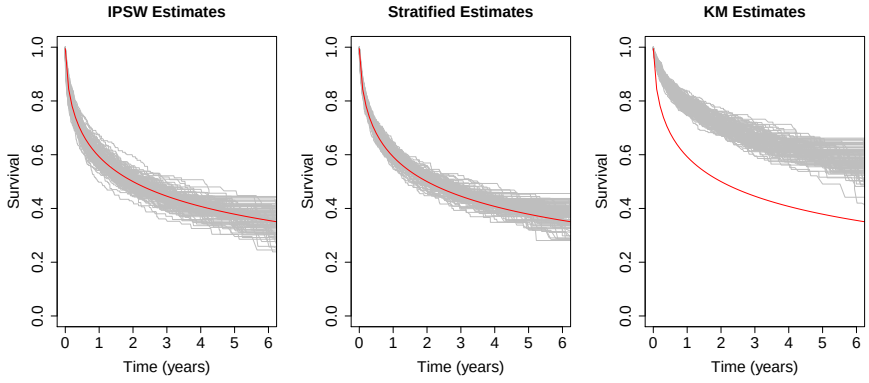


(b) Simulation results with a binary covariate for $X = 1$

Figure 5.1: Comparison of the distributions of the inverse probability of sampling weighted estimator and the Kaplan-Meier estimator based on 100 simulated datasets with a right-censored outcome, independent censoring and one binary covariate for $\boldsymbol{\beta} = (-7, 0.6)$ and $e^{\alpha} = 4$ (red line is the true survival curve in the target population)



(a) Simulation results with a continuous covariate for $X = 0$



(b) Simulation results with a continuous covariate for $X = 1$

Figure 5.2: Comparison of the distributions of the inverse probability of sampling weighted estimator and the Kaplan-Meier estimator based on 100 simulated datasets with a right-censored outcome, independent censoring and one continuous covariate for $\boldsymbol{\beta} = (-7, 0.6)$ and and $e^{\alpha} = 6$ (red line is the true survival curve in the target population)

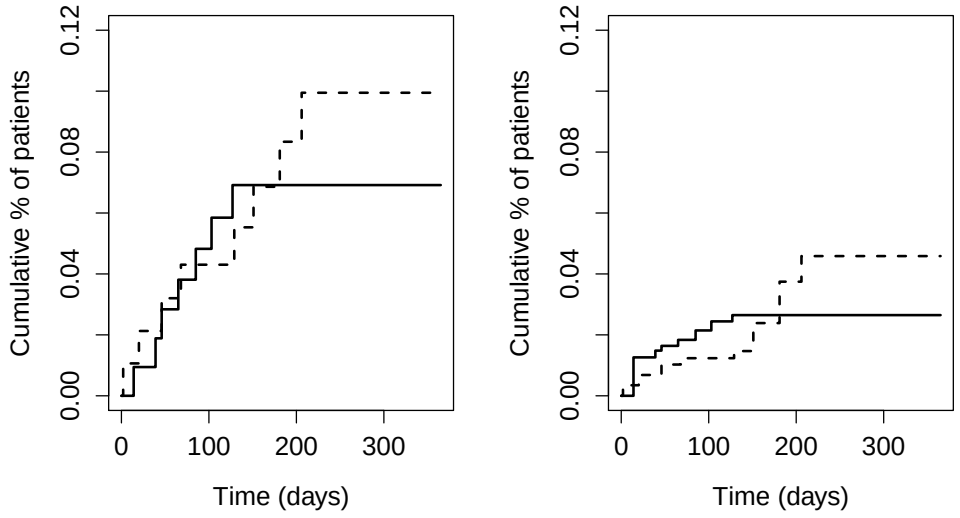


Figure 5.3: Complement of the Kaplan-Meier survival curves among women randomized to a regimen with a protease inhibitor (PI) (solid curves) and without a PI (dashed curves) in AIDS Clinical Trial Group 320 Study using intent-to-treat (left panel) and inverse probability of sampling weighted estimators (right panel). Representative cohort based on data from the Women's Interagency HIV Study.

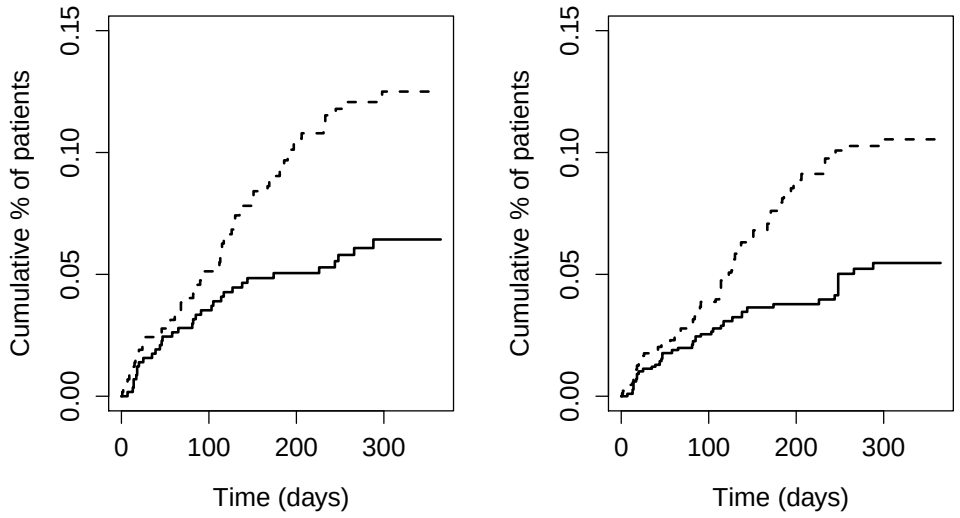


Figure 5.4: Complement of the Kaplan-Meier survival curves among participants randomized to a regimen with a protease inhibitor (PI) (solid curves) and without a PI (dashed curves) in AIDS Clinical Trial Group 320 Study using intent-to-treat (left panel) and inverse probability of sampling weighted estimators (right panel). Representative cohort based on data from the Center for AIDS Research Network of Integrated Clinical Systems.

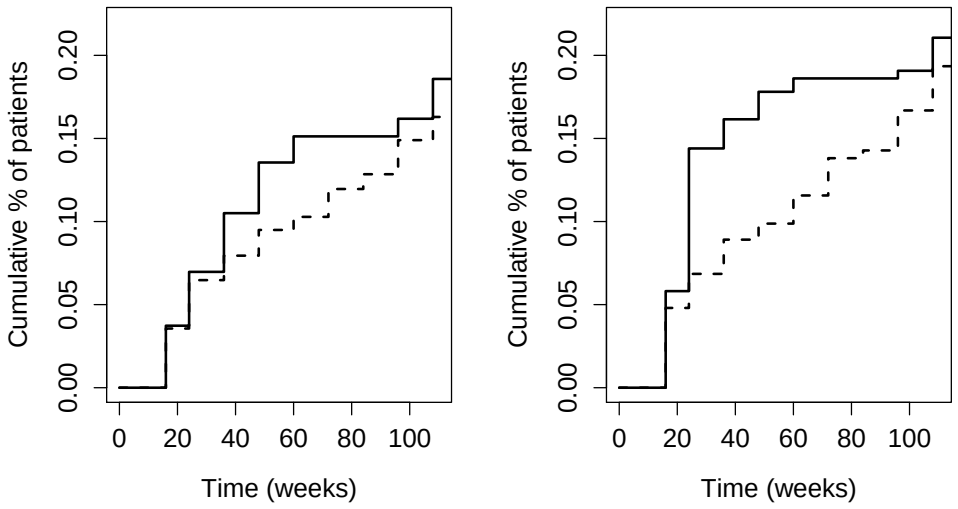


Figure 5.5: Complement of the Kaplan-Meier survival curves among those randomized to abacavir-lamivudine (ABC-3TC) (solid curves) and tenofovir disoproxil fumarate-emtricitabine (TDF-FTC) (dashed curves) in AIDS Clinical Trial Group A5202 Study using intent-to-treat (left panel) and inverse probability of sampling weighted estimators (right panel). Representative cohort based on data from the Women's Interagency HIV Study.

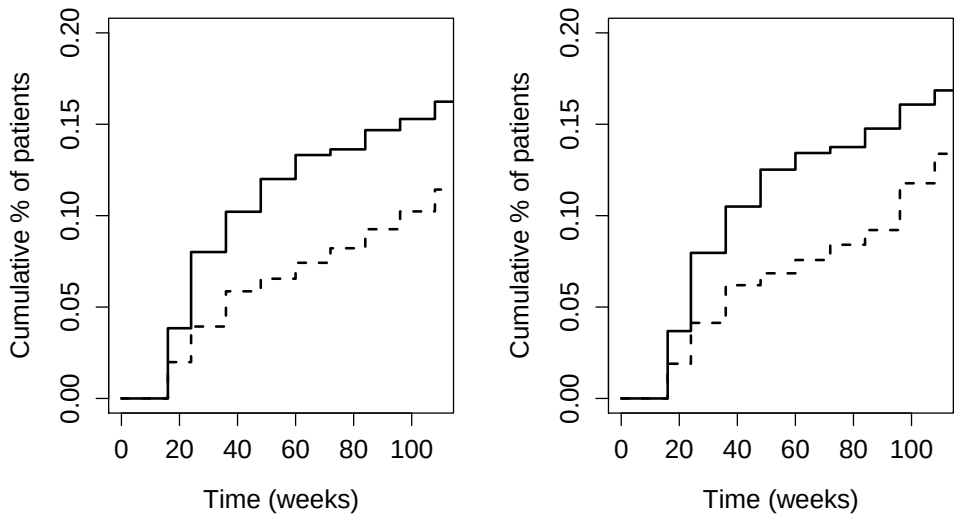


Figure 5.6: Complement of the Kaplan-Meier survival curves among participants randomized to abacavir-lamivudine (ABC-3TC) (solid curves) and tenofovir disoproxil fumarate-emtricitabine (TDF-FTC) (dashed curves) in AIDS Clinical Trial Group A5202 Study using intent-to-treat (left panel) and inverse probability of sampling weighted estimators (right panel). Representative cohort based on data from the Center for AIDS Research Network of Integrated Clinical Systems.

CHAPTER 6: CONCLUSION

To summarize, estimating causal effects may require consideration of both internal and external validity. For estimating effects in a study sample, addressing both confounding and selection bias are necessary. In Chapter 3, we continued the effort of improving understanding and utilization of causal inference methods to address confounding and selection bias in observational studies. We summarized the literature for IP-weighted Cox models through a comparison to the traditional Cox model and provided an illustrative example in HIV/AIDS research.

Estimation of causal effects in the target population requires both internal and external validity. We developed and applied methods for assessing generalizability of internally valid results. Following Cole and Stuart (2010) and Stuart et al. (2011), we considered an inverse probability of sampling weighted estimator in Chapter 4 for generalizing trial results to a target population, where the parameter of interest is a difference in average potential outcomes in a target population. In Chapter 5, we considered this estimator for right-censored data defined as an inverse weighted Kaplan-Meier estimator and empirically evaluated the performance of the nonparametric bootstrap standard error.

There are several future directions for this research. The appropriate method for choosing the covariates for the sampling score model remains an open question; however, methodology developed for treatment propensity scores may be extend for sampling score models (Brookhart et al., 2006; VanderWeele and Shpitser, 2011). In the case that there is residual treatment confounding in the trial, additional methodology will be needed to estimate effects. We suggest using an inverse probability of treatment weight; however, the statistical properties of this method need to be formally shown. The sampling score model was assumed to be correct; however, this may not always happen in practice. This method could be extended using a doubly-robust approach to address this concern (Bang and Robins, 2005). Quantitative methods for causal inference and generalizability is a growing field of statistical research. We hope that this body of work will be useful in strengthening the statistical rigor of these methods and increasing interest in quantitative methods for causal inference and generalizability.

APPENDIX A: REVIEW OF THE STANDARD (UNWEIGHTED) COX PROPORTIONAL HAZARDS MODEL

Let uppercase letters denote random variables and lowercase letters possible realizations of random variables or constants. Let $i = 1, \ldots, n$ index the study participants. Let T_i be the time from baseline to AIDS diagnosis or death, D_i be the time from baseline to study drop out, and C_i be the time from baseline to administrative censoring. In practice, only the minimum of T_i, D_i, and C_i is observed, denoted by $T_i^* = \min(T_i, D_i, C_i)$. See (Cole and Hudgens, 2010) for a review of univariate survival analysis methods.

The Cox proportional hazards regression model (Cox, 1972) is one of the most widely used statistical methods in biomedical research. The univariate Cox model is defined as $h_i(t) = h_0(t)\exp(\beta X_i)$, where $h_i(t)$ is the hazard function for individuals with covariate X_i, $h_0(t)$ is the reference hazard at time t for those with $X_i = 0$, and β is the log hazard ratio for a one unit change in X_i.

Heuristically, Cox regression may be understood as a series of logistic regression models, where at each ordered survival time, the log odds of the event are regressed on the exposure groups and any covariates (Efron, 1977). The Cox model is a semiparametric model because no assumption is placed on the probability distribution for the reference survival time distribution. Equivalently, the function $h_0(t)$ is left arbitrary. The parameters of a Cox model are estimated using maximum partial likelihood (Cox, 1975). Assuming no tied survival times, participant i who had the event at time t contributes the term $\exp(\beta X_i)/\sum_{j\in R(t)}\exp(\beta X_j)$ to the partial likelihood function, where $R(t)$ is the set of participants at risk at time t. For the case of a single covariate X_i, the partial likelihood is defined as simply a product of these individual contributions for events, or $L(\beta) = \prod_{i=1}^{n}\left[\frac{\exp(\beta X_i)}{\sum_{j\in R(T_i)}\exp(\beta X_j)}\right]^{Y_i}$, where Y_i is an event indicator (i.e., $T_i^* = T_i$). Only events contribute to the numerator of the likelihood due to the exponent Y_i. There are several ways to handle tied survival times, including methods ascribed to Peto and Breslow (Peto and Peto, 1972; Breslow, 1974), Efron (Efron, 1977) and an exact approach (Kalbfleisch and Prentice, 2002), which all return the same results if there are no ties. In the presence of moderate ties and if time is truly continuous, Efron's approximation performs well compared to the other approaches (Hertz-Picciotto and Rockhill, 1997).

One of the central assumptions of the Cox model is that the ratios of the hazards defined by levels of the covariates are constant over time. This is the proportional hazards assumption. The proportional hazards assumption can be assessed by fitting the model $h(t) = h_0(t)\exp(\beta_1 X_i + \beta_2 X_i t)$ and testing the null hypothesis that $\beta_2 = 0$, where $X_i \times t$ is a product of the covariate and time t.

In general, a $1-\alpha$ Wald confidence interval (CI) for the hazard ratio is defined as $\exp\left(\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\beta})}\right)$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of a standard normal distribution and $\hat{V}(\hat{\beta})$ is the estimated variance of $\hat{\beta}$. A Wald test statistic is defined as $\left(\hat{\beta}/\sqrt{\hat{V}(\hat{\beta})}\right)^2$ and is chi-squared distributed with 1 degree of freedom under the null hypothesis $\beta = 0$.

APPENDIX B: SANDWICH ESTIMATOR OF THE VARIANCE OF THE IPSW ESTIMATOR

The empirical sandwich-type estimator is used to approximate the asymptotic variance of the IPSW estimator. Substituting the following empirical estimates for their corresponding quantities in equation (4.2) produces a consistent sandwich estimator of the variance when $\boldsymbol{\beta}$ is known. Let $\hat{\boldsymbol{\theta}}^* = (\hat{\mu}_1, \hat{\mu}_0)$ and $\boldsymbol{\theta}_0^* = (\mu_1, \mu_0)$. Define the following matrices: $\hat{\mathbf{A}}^* = (n+m)^{-1} \sum_i \frac{\partial}{\partial \boldsymbol{\theta}_0^*} \Psi_\Delta^*(Y_i, \mathbf{Z}_i, X_i, S_i, \hat{\boldsymbol{\theta}}^*)$ and $\hat{\mathbf{B}}^* = (n+m)^{-1} \sum_i \Psi_\Delta^*(Y_i, \mathbf{Z}_i, X_i, S_i, \hat{\boldsymbol{\theta}}^*) \Psi_\Delta^{*T}(Y_i, \mathbf{Z}_i, X_i, S_i, \hat{\boldsymbol{\theta}}^*)$. $\hat{\boldsymbol{\theta}}^*$ is asymptotically normally distributed with mean $\boldsymbol{\theta}_0^*$ and covariance matrix $\hat{\Sigma}_\theta^* = \hat{\mathbf{A}}^{*^{-1}} \hat{\mathbf{B}}^* \hat{\mathbf{A}}^{*-T}$. When $\boldsymbol{\beta}$ is known, the estimator of the large sample variance of $\hat{\Delta}_{IPW}$ is

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

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

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

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

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

BIBLIOGRAPHY

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

Bang, H. and Robins, J. M. (2005), "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–973.

Bareinboim, E. and Pearl, J. (2013), "A general algorithm for deciding transportability of experimental results," *Journal of Causal Inference*, 1, 107–134.

Breslow, N. (1974), "Covariance analysis of censored survival data," *Biometrics*, 30, 89–99.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Starmer, T. (2006), "Variable selection for propensity score models," *American Journal of Epidemiology*, 163, 1149–1156.

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

CDC (2012), "HIV diagnosis data are estimates from all 50 states, the District of Columbia, and 6 U.S. dependent areas," *HIV Surveillance Supplemental Report*, 17.

Cole, S. R. and Frangakis, C. E. (2009), "The consistency statement in causal inference: a definition or an assumption?" *Epidemiology*, 20, 3–5.

Cole, S. R. and Hernan, M. A. (2002), "Fallibility in estimating direct effects," *International Journal of Epidemiology*, 31, 163–165.

— (2004), "Adjusted survival curves with inverse probability weights," *Computer Methods and Programs in Biomedicine*, 75, 45–49.

— (2008), "Constructing inverse probability weights for marginal structural models," *American Journal of Epidemiology*, 168, 656–664.

Cole, S. R., Hernan, M. A., Robins, J. M., Anastos, K., Chmiel, J., and Detels, R. (2003), "Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models," *American Journal of Epidemiology*, 158, 687–694.

Cole, S. R. and Hudgens, M. G. (2010), "Survival analysis in infectious disease research: describing events in time," *AIDS*, 24, 2423.

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

Collett, D. (2003), *Modelling Survival Data in Medical Research*, Boca Raton: CRC Press.

Cox, D. R. (1972), "Regression models and life-tables," *Journal of the Royal Statistical Society.Series B (Methodological)*, 34, 187–220.

— (1975), "Partial likelihood," *Biometrika*, 62, 269–276.

D'Agostino, R. B., Lee, M., Belanger, A. J., Cupples, L. A., Anderson, K., and Kannel, W. B. (1990), "Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart Study," *Statistics in Medicine*, 9, 1501–1515.

Efron, B. (1977), "The efficiency of Cox's likelihood function for censored data," *Journal of the American Statistical Association*, 72, 557–565.

Efron, B. and Tibshriani, R. (1994), *An Introduction to the Bootstrap*, London: Chapman Hall.

Fisher, R. A. (1973), *Statistical Methods for Research Workers*, New York: Hafner, 14th ed.

Frangakis, C. (2009), "The calibration of treatment effects from clinical trials to target populations," *Clinical Trials*, 6, 136–140.

Frangakis, C. E. and Rubin, D. B. (1999), "Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes," *Biometrika*, 86, 365–379.

— (2002), "Principal stratification in causal inference," *Biometrics*, 58, 21–29.

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

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

Greenhouse, J. B., Kaizar, E. E., Kelleher, K., Seltman, H., and Gardner, W. (2008), "Generalizing from clinical trial data: A case study. The risk of suicidality among pediatric antidepressant users," *Statistics in Medicine*, 27, 1801–1813.

Greenland, S. (1996), "Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference," *Epidemiology*, 7, 498–501.

Greenland, S. and Morgenstern, H. (2001), "Confounding in health research," *Annual Review of Public Health*, 22, 189–212.

Greenland, S., Pearl, J., and Robins, J. M. (1999a), "Causal diagrams for epidemiologic research," *Epidemiology*, 10, 37–48.

Greenland, S., Robins, J. M., and Pearl, J. (1999b), "Confounding and collapsibility in causal inference," *Statistical Science*, 14, 29–46.

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

Hernan, M. A. (2010), "The hazards of hazard ratios," *Epidemiology*, 21, 13–15.

Hernan, M. A., Brumback, B., and Robins, J. M. (2000), "Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men," *Epidemiology*, 11, 561–570.

— (2001), "Marginal structural models to estimate the joint causal effect of nonrandomized treatments," *Journal of the American Statistical Association*, 96, 440–448.

Hernan, M. A. and Cole, S. R. (2009), "Invited commentary: Causal diagrams and measurement bias," *American Journal of Epidemiology*, 170, 959–962.

Hernan, M. A., Hernadez-Diaz, S., and Robins, J. M. (2013), "Randomized trials analyzed as observational studies," *Annals of Internal Medicine*, 159, 560–562.

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

Hertz-Picciotto, I. and Rockhill, B. (1997), "Validity and efficiency of approximation methods for tied survival times in Cox regression," *Biometrics*, 1151–1156.

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

Holland, P. W. (1986), "Statistics and causal inference," *Journal of the American Statistical Association*, 81, 945–960.

Horvitz, D. G. and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685.

Howe, C. J., Cole, S. R., Chmiel, J. S., and Munoz, A. (2011a), "Limitation of inverse probability of censoring weights in estimating survival in the presence of strong selection bias," *American Journal of Epidemiology*, 173, 569–577.

Howe, C. J., Cole, S. R., Westreich, D. J., Greenland, S., Napravnik, S., and Jr, J. J. E. (2011b), "Splines for trend analysis and continuous confounder control," *Epidemiology*, 22, 874.

Hudgens, M. G. and Halloran, E. M. (2006), "Causal vaccine effects on binary postinfection outcomes," *Journal of the American Statistical Association*, 101, 2281–2298.

Hudgens, M. G., Hoering, A., and Self, S. G. (2003), "On the analysis of viral load endpoints in HIV vaccine trials," *Statistics in Medicine*, 22, 21–29.

Kalbfleisch, J. D. and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, Hoboken: Wiley-Interscience.

Kang, J. D. and Schafer, J. L. (2007), "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical Science*, 22, 523–539.

Kaplan, E. L. and Meier, P. (1958), "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, 53, 457–481.

Kaufman, J. S. (2010), "Marginalia: comparing adjusted effect measures," *Epidemiology*, 21, 490–493.

Kish, L. (1992), "Weighting for unequal P," *Journal of Official Statistics*, 8, 183–200.

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

Lau, B., Cole, S. R., and Gange, S. J. (2009), "Competing risk regression models for epidemiologic data," *American Journal of Epidemiology*, 170, 244–256.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010), "Improving propensity score weighting using machine learning," *Statistics in Medicine*, 29, 337–346.

— (2011), "Weight trimming and propensity score weighting," *PloS One*, 6, e18174.

Lin, D. and Wei, L.-J. (1989), "The robust inference for the Cox proportional hazards model," *Journal of the American Statistical Association*, 84, 1074–1078.

Little, R. J. and Rubin, D. B. (2000), "Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches," *Annual Review of Public Health*, 21, 121–145.

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

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

Mickey, R. M. and Greenland, S. (1989), "The impact of confounder selection criteria on effect estimation," *American Journal of Epidemiology*, 129, 125–137.

Morgan, S. L. and Winship, C. (2007), *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, New York: Cambridge University Press.

Neyman, J., Dabrowska, D., and Speed, T. (1992), "On the application of probability theory to agricultural experiments," *Statistical Science*, 5, 465–472.

Nieto, F. J. and Coresh, J. (1996), "Adjusting survival curves for confounders: a review and a new method," *American Journal of Epidemiology*, 143, 1059–1068.

O'Muircheartaigh, C. and Hedges, L. V. (2013), "Generalizing from unrepresentative experiments: a stratified propensity score approach," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63, 195210.

Pearl, J. (2001), "Direct and indirect effects," in *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., pp. 411–420.

— (2010), "On the consistency rule in causal inference: axiom, definition, assumption, or theorem?" *Epidemiology*, 21, 872–875.

Petersen, M. L., Wang, Y., van der Laan, M. J., and Bangsberg, D. R. (2006), "Assessing the effectiveness of antiretroviral adherence interventions: Using marginal structural models to replicate the findings of randomized controlled trials," *JAIDS*, 43, S96–S103.

Peto, R. and Peto, J. (1972), "Asymptotically efficient rank invariant test procedures," *Journal of the Royal Statistical Society.Series A (General)*, 185–207.

Robins, J. (1989), "The control of confounding by intermediate variables," *Statistics in Medicine*, 8, 679–701.

Robins, J. M. (1998), "Marginal structural models," in *Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA, pp. 1–10.

— (2000), *Marginal structural models versus structural nested models as tools for causal inference*, Springer, Statistical Models in Epidemiology, the Environment, and Clinical Trials, pp. 95–133.

Robins, J. M. and Finkelstein, D. M. (2000), "Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted log rank tests," *Biometrics*, 56, 779–788.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000), "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 11, 550–560.

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

Rothman, K. J. (1986), *Modern Epidemiology*, Philadelphia: MA Little, Brown and Company.

Rothman, K. J., Gallacher, J. E., and Hatch, E. E. (2013), "Why representativeness should be avoided," *International Journal of Epidemiology*, 42, 1012–1014.

Rubin, D. B. (1980), "Randomization analysis of experimental data: The Fisher randomization test comment," *Journal of the American Statistical Association*, 75, 591–593.

— (1990), "Comment: Neyman (1923) and causal inference in experiments and observational studies," *Statistical Science*, 5, 472–480.

Sato, T. and Matsuyama, Y. (2003), "Marginal structural models as a tool for standardization," *Epidemiology*, 14, 680–686.

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

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

Scott, A. and Wild, C. (2002), "On the robustness of weighted methods for fitting models to case–control data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 207–219.

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

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

Stuart, E. A., Bradshaw, C. P., and Leaf, P. J. (2014), "Assessing the generalizability of randomized trial results to target populations," *Prevention Science*, 1–11.

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

Thompson, S. (2012), *Sampling*, Hoboken, NJ: John Wiley and Sons.

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

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

VanderWeele, T. J. (2009a), "Concerning the consistency assumption in causal inference," *Epidemiology*, 20, 880–883.

— (2009b), "Marginal structural models for the estimation of direct and indirect effects," *Epidemiology*, 20, 18–26.

VanderWeele, T. J. and Shpitser, I. (2011), "A new criterion for confounder selection," *Biometrics*, 67, 1406–1413.

Weisberg, H. I., Hayden, V. C., and Pontes, V. P. (2009), "Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality?" *Clinical Trials*, 6, 109–118.

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

Xie, J. and Liu, C. (2005), "Adjusted Kaplan Meier estimator and log rank test with inverse probability of treatment weighting for survival data," *Statistics in Medicine*, 24, 3089–3110.

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