

Selection Bias Due to Loss to Follow Up in Cohort Studies

Chanelle J. Howe,^a Stephen R. Cole,^b Bryan Lau,^c Sonia Napravnik,^{b,d} and Joseph J. Eron, Jr.^d

Abstract: Selection bias due to loss to follow up represents a threat to the internal validity of estimates derived from cohort studies. Over the past 15 years, stratification-based techniques as well as methods such as inverse probability-of-censoring weighted estimation have been more prominently discussed and offered as a means to correct for selection bias. However, unlike correcting for confounding bias using inverse weighting, uptake of inverse probability-of-censoring weighted estimation as well as competing methods has been limited in the applied epidemiologic literature. To motivate greater use of inverse probability-of-censoring weighted estimation and competing methods, we use causal diagrams to describe the sources of selection bias in cohort studies employing a time-to-event framework when the quantity of interest is an absolute measure (e.g., absolute risk, survival function) or relative effect measure (e.g., risk difference, risk ratio). We highlight that whether a given estimate obtained from standard methods is potentially subject to selection bias depends on the causal diagram and the measure. We first broadly describe inverse probability-of-censoring weighted estimation and then give a simple example to demonstrate in detail how inverse probability-of-censoring weighted estimation mitigates selection bias and describe challenges to estimation. We then modify complex, real-world data from the University of North Carolina Center for AIDS Research HIV clinical cohort study and estimate the absolute and relative change in the occurrence of death with and without inverse probability-of-censoring weighted correction using the modified University of North Carolina data. We provide SAS code to aid with implementation of inverse probability-of-censoring weighted techniques.

(*Epidemiology* 2016;27: 91–97)

Submitted 13 February 2015; accepted 29 September 2015.

From the ^aDepartment of Epidemiology, Center for Population Health and Clinical Epidemiology, Brown University School of Public Health, Providence, RI; ^bDepartment of Epidemiology, University of North Carolina Gillings School of Global Public Health, Chapel Hill, NC; ^cDepartment of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; and ^dDivision of Infectious Diseases, Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, NC.

Supported by the National Institutes of Health Grant P30 AI50410.

The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Chanelle J. Howe, Department of Epidemiology, Center for Population Health and Clinical Epidemiology, Brown University School of Public Health, 121 South Main Street, Providence, RI 02912. E-mail: chanelle_howe@brown.edu.

Copyright © 2015 Wolters Kluwer Health, Inc. All rights reserved.

ISSN: 1044-3983/16/2701-0091

DOI: 10.1097/EDE.0000000000000409

In cohort studies, a group of individuals are sampled from a source population and followed over time to ascertain the occurrence of an outcome of interest.¹ Such cohort data are often analyzed using a time-to-event framework given the frequent occurrence of loss to follow up. In the analysis of time-to-event data, a common objective is to estimate survival in the source population, as well as how survival differs by levels of exposure. Selection bias due to loss to follow up, also known as informative censoring, represents a threat to the internal validity of estimates derived from cohort studies.² Over the past 15 years, stratification-based techniques such as standard regression adjustment as well as methods such as inverse probability-of-censoring weighted estimation have been more prominently discussed and offered as a means to correct for such selection bias.^{2–9} However, unlike correcting for confounding bias using inverse probability-of-exposure weights,^{7,10} uptake of inverse probability-of-censoring weighted estimation as well as competing methods,^{11–17} including missing data approaches, such a multiple imputation to correct for selection bias has been limited in the applied epidemiologic literature.

This limited uptake may be due to a lack of clarity regarding the sources of selection bias in cohort studies as well as few detailed applications. Lack of clarity regarding the sources of selection bias may also contribute to the limited discussion in the epidemiologic literature concerning the importance of incorporating in the design phase of a cohort study the collection of information necessary to correct analytically for such selection bias.^{9,18} This limited discussion is in stark contrast to the frequently mentioned importance of collecting information on potential confounders as part of the study design.

Therefore, the objectives of this article are, first, to use causal diagrams to describe the sources of selection bias in cohort studies analyzed under a time-to-event framework given the presence of loss to follow up when the quantity of interest is an absolute measure (e.g., absolute risk, survival function) or relative effect measure (e.g., risk difference, risk ratio). The absolute measure describes the occurrence of a certain characteristic or outcome in a single group. By relative effect measure, we mean a measure that compares two or more groups (e.g., exposed vs. unexposed) that is intended to estimate a causal effect or an associational effect when the exposure is not well defined.³ We focus primarily on the risk difference and risk ratio for the relative effect measures of

interest instead of the hazard ratio, which is more commonly estimated in time-to-event analyses, to avoid the selection bias that the hazard ratio is innately subject to.¹⁹ The second objective is to broadly describe inverse probability-of-censoring weighted techniques. Third, we will provide a simple example that demonstrates how inverse probability-of-censoring weighted estimation corrects for selection bias. Fourth, we will discuss related challenges to estimation. Fifth, we will modify more complex, real-world data from the University of North Carolina Center for AIDS Research (UNC CFAR) HIV clinical cohort study and estimate the absolute and relative change in the occurrence of death with and without inverse probability-of-censoring weighted correction for potential selection bias due to loss to follow up using the UNC data. The UNC analyses were performed in SAS, version 9.3, software (SAS Institute, Inc., Cary, NC).

NOTATION

In a cohort of $i = 1$ to n HIV-positive individuals who became infected at least 5 years before study entry, let T_i represent the time in visits from study entry to the occurrence of the event (death), C_i is the time in visits from study entry to censoring due to loss to follow up, M_i is the time in visits from study entry to the administrative end of the study, and Y_i is the observed follow-up time ($Y_i = \min(T_i, C_i, M_i)$) for person i . Defining u to be an index of time in visits since study entry ($u = 1$ to $\max(y_i)$), $A_i(u)$ is a measured indicator of injection drug use in the prior 6 months (1: yes; 0: no), $L_i(u)$ is a measured indicator of heavy alcohol use in the prior 6 months (1: yes, 0: no), $Q_i(u)$ is an unmeasured indicator of CD4 cell count (1: ≥ 200 cells/ μ l, 0: < 200 cells/ μ l), and $Z_i(u)$ is an unmeasured indicator of level of education (1: not college educated, 0: college educated) at time u for person i . Further at time u , $D_i(u)$ is an indicator of loss to follow up (1: lost, 0: otherwise), while $O_i(u)$ is an indicator of developing the event (1: event, 0: otherwise) for person i . Henceforth, i and u will be suppressed when possible.

CAUSAL DIAGRAMS FOR THE SOURCES OF SELECTION BIAS DUE TO LOSS TO FOLLOW UP

Selection bias due to loss to follow up is the absolute or relative bias that arises from how participants are selected out of a given risk set.³ Here and throughout this article, absolute bias refers to bias of an absolute measure, whereas relative bias pertains to the bias of a relative effect measure. We define bias as a difference between the expected value of an estimator (e.g., mean survival, mean log risk ratio) and the true value for the quantity of interest in the study population present at baseline which we henceforth assume represents the source population.²⁰

Hernán et al.^{2,9} outlined a common structure for selection bias based on causal diagrams when the quantity of interest is a relative effect measure and the exposure does not cause the outcome resulting in an equivalence between collider-stratification bias (i.e., bias resulting from conditioning on a collider)

and relative selection bias.^{3,21} Here, we build on this prior work when the exposure causes the outcome and demonstrate that selection bias of a relative effect measure can occur even in the absence of conditioning on a collider. Furthermore, we discuss absolute bias and the fact that whether a given estimate is subject to selection bias depends on the causal diagram and the measure. For some diagrams, both the absolute and relative estimates are unbiased, whereas in others, solely the absolute measure or both the absolute and relative measure may be biased. The diagrams we identify here for when the absolute or relative measure may be biased build on work by Hernán et al.,^{2,9} are informed by theoretical and applied study by Daniel et al.,¹⁸ Greenland and Pearl,²² and Westreich,²³ and have been demonstrated in simulations included in our eAppendix 1 (<http://links.lww.com/EDE/A985>). For those less familiar with relevant definitions as well as the rules of and assumptions encoded in causal diagrams including the definition of a collider, we refer the reader to the Appendix of Hernán et al.²⁴

Figure 1 shows five causal diagrams for the effect of injection drug use, heavy alcohol use, CD4 cell count, and education on loss to follow up and time to death. In each diagram the exposure (if applicable) is injection drug use and a box appears around D given that the analysis is restricted to those participants who remain not lost to follow up at a given time u . Diagram I indicates that losses occur completely at random given that losses are not associated with A , L , or T . Losses that occur completely at random imply that those who are lost represent a simple, uniform random sample of those who were at risk for the event at a given time since study entry. Completely at random losses are considered to be a type of noninformative censoring where losses occur independently of the event of interest. In contrast, diagrams II to V imply that losses do not occur completely at random, meaning that those who are lost to follow up are not a random sample of all participants who are in the risk set at the time a given participant is lost. When who is lost is related to the occurrence of the outcome of interest then losses are considered to be informative.²⁵

In diagram I, given that losses are random with respect to A , L , and T , loss to follow up in the cohort does not induce absolute or relative selection bias when standard survival analysis methods (e.g., discrete-time survival function estimator, discrete-time hazard model) are used for estimation. However, in diagram II, losses are dependent on L ; therefore, loss to follow up is not random. Given that L also predicts T these losses are informative and therefore losses may introduce bias of absolute measures or relative effect measures.

For instance, let us assume that those who engage in heavy alcohol use were more likely to be lost to follow up as well as die than those who do not engage. This prior scenario, which is represented by L being a common cause of D and T in diagram II, would result in nonengagers, who are less likely to die, being more likely to remain in the risk set during follow up. As such, the

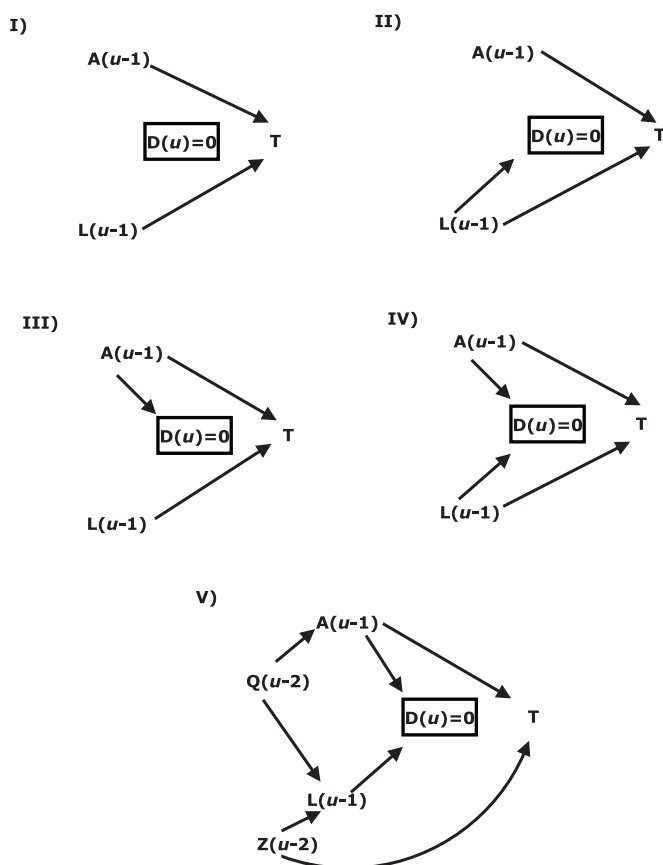


FIGURE 1. Causal diagram depicting five scenarios (I–V) for the effect of injection drug use (A), heavy alcohol use (L), CD4 cell count (Q), and education (Z), on lost to follow up (D) and time to death (T) in a cohort study where u indexes time in visits since study entry and denotes that A , L , Q , Z , and D can vary with time.

survival function in the source population is expected to be overestimated in the analysis sample. The estimated relative effect of injection drug use on death may be biased as well. Such relative bias may occur because of inaccurate estimation of a joint effect. As discussed in our eAppendix 2 (<http://links.lww.com/EDE/A985>) and elsewhere,^{3,6} validly estimating the relative effect of injection drug use on death in the presence of loss to follow up requires accurate estimation of a joint effect. Accurate estimation of a joint effect requires adequately accounting for all common causes of loss and the outcome of interest (e.g., L).^{3,6}

Prior work²³ and simulations (not shown) indicate that when A does not cause T in diagram II, the estimated relative effect (i.e., risk difference, risk ratio, and odds ratio) is not biased. Similar to diagram II, losses should be informative, but in this case, dependent on A in diagram III given that A is a common predictor of D and T . These losses are expected to introduce bias of absolute measures, but will not bias the relative effect of the exposure, injection drug use, given that within strata of injection drug use losses should be random.

In diagrams IV and V, losses are informative related to both A and L . Specifically, in diagram IV, both A and L are common causes of D and T . In diagram V, A is a common cause of D and T , whereas L causes D and shares a common cause with T , the covariate Z . These informative losses are expected to result in selection bias for both the absolute and relative measures. The absolute measure is expected to be biased because D and T are associated via A and L . The relative effect measure is expected to be biased because restricting the analysis sample to those who remain in the risk set opens a noncausal path from A to D to L to T (or A to D to L to Z to T) given that D is a collider. In other words, even within levels of injection drug use losses will be informative. Losses will be informative given that engaging in heavy alcohol use is associated with injection drug use due to restricting the analysis to those who remain under follow up and engaging in heavy alcohol use is associated with time to death.

USING INVERSE PROBABILITY-OF-CENSORING WEIGHTS TO CORRECT FOR SELECTION BIAS DUE TO LOSS TO FOLLOW UP

Ideally losses to follow up would be minimized during the design and conduct stages of a cohort study by minimizing losses since selection via loss is required to have selection bias and the extent of selection bias is partly dependent on the degree of selection (e.g., percent lost to follow up). However, in most settings, some losses are unavoidable and such losses often do not occur completely at random. Therefore, informed by causal diagrams, nonstandard analytic methods should be considered and perhaps employed to correct for potential bias induced by loss to follow up. Such methods include inverse probability-of-censoring weighted estimation as well as stratification-based techniques including standard regression adjustment that stratify the data to address selection bias.^{2,3}

As noted by Hernán et al.² and described later using the UNC HIV example as well as in our eAppendix 2 (<http://links.lww.com/EDE/A985>) in the case of diagram V, there are situations where stratification-based methods may be insufficient to correct for selection bias, while inverse probability-of-censoring weighted estimation continues to provide unbiased estimates given that necessary assumptions outlined below are met. Furthermore, compared with stratification-based techniques, inverse probability-of-censoring weighted estimation can more readily provide marginal rather than conditional estimates of absolute measures corrected for potential selection bias. Marginal estimates have a preferred interpretation and are easier to display graphically compared with conditional estimates.²⁶ Therefore, the remainder of the article largely focuses on inverse probability-of-censoring weighted estimation rather than stratification-based techniques to address selection bias. Next, we broadly describe the use of inverse probability-of-censoring weights to correct for potential selection bias.

Inverse probability-of-censoring weights can be used to create the pseudo-population that would have been observed had losses to follow up occurred but been random with respect to measured determinants of loss to follow up (depicted in the relevant causal diagram) including the exposure (if applicable). This pseudo-population can be created by re-weighting the contribution of each participant who was not lost to follow up to a given risk set. Specifically, at time u , each participant is typically assigned a stabilized weight $SW(u)$ that is a ratio of the probability that the participant was not lost to follow up through time u conditional on the exposure (if applicable) and the probability that the participant remained not lost to follow up through time u conditional on measured determinants of loss to follow up including the exposure (if applicable). The aforementioned probabilities as well as the weight, $SW(u)$, are often estimated using a pooled logistic regression model for not being lost to follow up.⁹ The $SW(u)$ can then be used to estimate weighted versions of standard survival analysis methods. In our eAppendix 2 (<http://links.lww.com/EDE/A985>), we use a simple example to more thoroughly demonstrate the use of inverse probability-of-censoring weights to reduce selection bias when estimating survival after study entry as well as the change in survival as a function of injection drug use via the risk difference or risk ratio.

For valid estimation of absolute measures and causal relative effect measures using inverse probability-of-censoring weights, the assumptions of exchangeability, positivity, and correct model specification in the outcome and weight model (where appropriate) must hold. Furthermore, the exposure (if applicable) and censoring mechanism must be well defined given that the exposure (if applicable) and censoring mechanism represent points of intervention.^{3,5} When any of the prior assumptions and conditions are not met, the results from using inverse probability-of-censoring weighted estimation may be biased or lack a causal interpretation. Conditional exchangeability assumes that there are no unaccounted for sources of confounding bias (if applicable) and selection bias due to lost to follow up. Positivity requires that there is a nonzero probability of every possible exposure level (if applicable) within every observed combination of the measured confounders. In addition, there must be a nonzero probability of not being lost to follow up at each time that losses occur within every combination of possible exposure levels (if applicable) and observed measured variables that contribute to the selection bias. Lack of positivity can occur for systematic reasons (e.g., a given exposure level is not possible at a specific level of the confounder) or due to random chance (e.g., small sample size).^{5,27,28} Correct model specification means that the model choice, including model form and functional forms between the predictors and the dependent variable (i.e., exposure [if applicable], censoring, or outcome) in all relevant regression models are correct. A well-defined exposure and censoring mechanism does not suffer from interference²⁹ and either corresponds to a single

well-defined intervention or has version irrelevance when more than one well-defined intervention exists.³⁰

To minimize the potential for violations in conditional exchangeability, potential confounders as well as common causes of loss to follow up and the outcome of interest should be considered in the study design phase and included in data collection.^{9,18} Although violations in conditional exchangeability are not testable, sensitivity analyses can be performed to assess the robustness of inference to unmeasured sources of selection bias.^{31,32} In the presence of potential positivity violations, more complex double robust estimators such as targeted minimum loss-based estimation can instead be used for appropriate estimation as long as the outcome distribution is consistently estimated.^{17,28} Correct specification in the weight model can be facilitated using data-adaptive procedures including super and ensemble learning techniques rather than the more commonly used pooled logistic regression model.³³ Even if positivity and correct model specification are not an issue, targeted minimum loss-based estimation with data-adaptive procedures should still be considered given the potential for efficiency gains when measured covariates can predict the outcome well.¹⁷

EXAMPLE: UNIVERSITY OF NORTH CAROLINA CENTER FOR AIDS RESEARCH HIV CLINICAL COHORT STUDY

African Americans have been shown to suffer disproportionately from HIV-related mortality.³⁴ Therefore, here, we use modified data on 2,511 HIV-infected persons in the UNC CFAR HIV clinical cohort to examine the association between African American race and subsequent mortality. We focus on association since African American race is not a well-defined exposure because it does not correspond to any possible well-defined, real-world intervention.^{3,35} The UNC CFAR HIV clinical cohort (henceforth, the cohort) collects relevant information from all HIV-positive patients attending the UNC HIV clinic who provide written informed consent in English or Spanish. All study forms and protocols were approved by the UNC institutional review board. The secondary data analysis below was approved by the institutional review boards at UNC and Brown University. Additional details concerning this clinic cohort are provided elsewhere.³⁶

This analysis uses data on the 2,511 African American and Caucasian patients who attended the UNC HIV clinic during the study period, January 1, 1999, to January 1, 2012, and who had information available on date of birth, gender, insurance status, prior AIDS-defining illness diagnoses, CD4 cell count, and HIV RNA level at least at the first clinic visit during the study period (henceforth, the first clinic visit). The data were modified such that clinic visits as well as assessment and updating of CD4 cell count and HIV RNA level occurred every 6 months subsequent to the first clinic visit. Insurance status and prior AIDS-defining illnesses were assumed to only be known at the first clinic visit. Death dates were coarsened

to only occur at clinic visits. Last observation carried forward methods were used to complete CD4 and HIV RNA measures that were unavailable for a given visit. For the purposes of this simplified example, these completed values were assumed to represent the truth. However, beyond this simplified example, other more sophisticated and potentially less biased techniques for handling missing data should be considered.³⁷ Patients were considered to be lost to follow up 2 years after the last time they were seen at a clinic visit during the study period. Patients who were last seen within 2 years of January 1, 2012, were administratively censored at January 1, 2012.

Diagram I in Figure 2 is a causal diagram for the effect of African American race on time to death among the UNC cohort patients.^{36,38,39} Assuming this causal diagram is correct, then the effect of African American race on time to death is potentially subject to selection bias via the non-causal path from African American race to loss to follow up to covariates that include CD4 cell count, AIDS, HIV RNA level, and insurance to death. Stratification-based methods such as standard regression adjustment for the abovementioned covariates would address this potential selection bias. However, any indirect effect that African American race has on time to death that operates although these covariates may also be removed with standard regression adjustment. Inverse probability-of-censoring weights can account for this selection bias while allowing for estimation

of the effect of African American race on death operating through pathways that include and do not include the mentioned covariates. Simulations that appear in our eAppendix 1 (<http://links.lww.com/EDE/A985>) were performed to confirm the potential selection bias of the effect of African American race on time to death and that inverse probability-of-censoring weights can be used to appropriately reduce such selection bias.

To further demonstrate the impact of informative losses on estimation, the hypothesized causal relationships indicated by diagram I of Figure 2 were created or strengthened by modifying the UNC data. Standard and inverse probability-of-censoring weighted approaches were then used to estimate measures of interest based on the altered UNC data. Table shows observed patient characteristics at the first clinic visit for the modified data. During follow up, 404 patients died, 1,390 patients were lost to follow up, and 717 patients reached the end of study follow up alive.

African American race, insurance status, and ever receiving a diagnosis of an AIDS-defining illness at the first clinic visit, as well as CD4 cell count and HIV RNA level at the prior visit were used to estimate inverse probability-of-censoring weights using pooled logistic regression. Our eAppendix 3 (<http://links.lww.com/EDE/A985>) provides the SAS, version 9.3 code that was used to estimate the aforementioned weights. In the pooled logistic regression model, continuous covariates were fit using linear and quadratic terms while indicator variables were used for noncontinuous covariates. The resultant weights had a mean (standard deviation) of 1.00 (0.37) with a range from 0.33 to 11.30. As shown in Table, the observed distribution of characteristics at the first clinic visit was preserved in the weighted population. However, the sample size at the first clinic visit in Table and the number of deaths in the weighted data compared with the observed data increased by 1 and 66, respectively. The aforementioned increases may indicate model misspecification or nonpositivity⁵ which the alternative stratification-based approaches are subject as well. Assuming all necessary assumptions hold, the diagram that corresponds to this weighted population is shown in diagram II of Figure 2 where censoring due to loss to follow up is random with respect to African American race and all of the other measured covariates.

Risk ratios obtained from the standard and inverse probability-of-censoring weighted survival functions were used to quantify the association between African American race and subsequent death. Figure 3 shows the survival functions and risk ratios comparing African Americans to Caucasians in the observed and weighted populations. Assuming diagram I in Figure 2 is correct, the aforementioned results show that selection bias due to loss to follow up related to the measured exposure and covariates was sizeable. Specifically, informative selection appeared to overestimate survival and alter the association between African American race and subsequent death at later visits.



FIGURE 2. Causal diagram depicting the association between African American race and time to death in the unweighted (I) and weighted (II) data among 2,511 HIV-infected African American and Caucasian men and women with 25,319 total person-visits of follow up where u indexes time in visits since study entry, UNC CFAR HIV clinical cohort, 1999–2012.

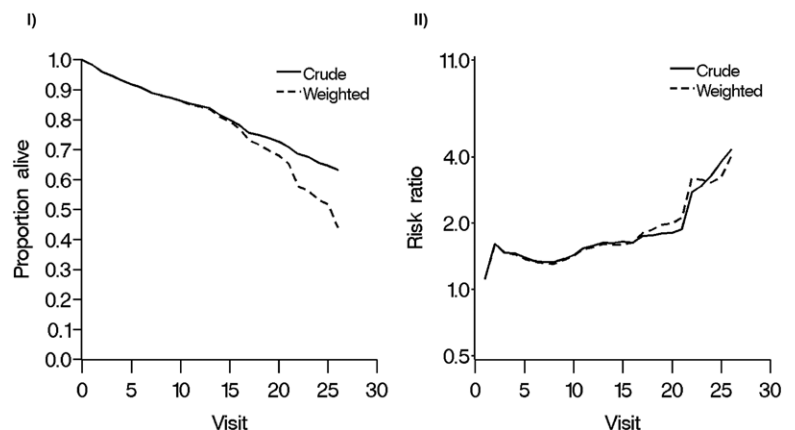
TABLE. Observed and Weighted Characteristics of 2,511 HIV-infected African American and Caucasian Men and Women, UNC CFAR HIV Clinical Cohort, 1999–2012

Characteristic	At First Clinic Visit During Study Period N = 2,511 Patients in Observed Population	At First Clinic Visit During Study Period N = 2,512 Patients in Weighted ^a Population
Age in years, median (quartiles)	39 (32; 46)	39 (32; 46)
Male, % (n)	70 (1,749)	70 (1,749)
African American, % (n)	66 (1,652)	66 (1,659)
Prior AIDS-defining illness diagnosis, % (n)	24 (605)	24 (602)
Prior antiretroviral therapy use, % (n)	79 (1,979)	79 (1,976)
Insurance, % (n)		
Private	25 (639)	25 (634)
Public ^b	38 (947)	38 (944)
Uninsured	37 (925)	37 (934)
CD4 cell count in cells/ μ l, % (n)		
<200	29 (738)	29 (734)
\geq 200	71 (1,773)	71 (1,778)
Detectable HIV-1 RNA level, % (n)		
Yes	62 (1,562)	62 (1,567)
No	38 (949)	38 (945)

^aAccounts for insurance status and receiving a prior AIDS-defining illness diagnosis at the first clinic visit as well as CD4 cell count and HIV RNA level at the prior visit.

^bMedicaid, Medicare, or other US public insurance (e.g., AIDS Drug Assistance Program, Veterans Administration, Department of Defense for prisoners).

FIGURE 3. Proportion alive (I) and risk ratio for death comparing African Americans to Caucasians (II) by visit among 2,511 HIV-infected men and women with 25,319 total person-visits of follow up, UNC CFAR HIV clinical cohort, 1999–2012. The *solid curve* (crude) does not correct for selection bias, whereas the *dashed curve* (weighted) corrects for selection bias due to loss to follow up dependent on African American race and measured covariates including insurance status and a prior AIDS-defining illness diagnosis at the first clinic visit as well as CD4 cell count and HIV RNA level at the prior visit using inverse probability-of-censoring weights.



DISCUSSION

Here, we used simple notation and causal diagrams to better characterize the sources of selection bias due to attrition in cohort studies when the quantity of interest is an absolute measure or relative effect measure. We discussed that when the exposure causes the outcome, conditioning on a collider is not necessary for selection bias of a relative effect. Instead, selection bias of a relative effect may occur solely due to the existence of a common cause of loss and the outcome. In addition, whether a given estimate obtained from standard methods is subject to selection bias can depend on the measure. For some scenarios, both the absolute and relative estimates obtained from standard methods will be unbiased, whereas in others, solely the absolute measure or both the absolute and relative measures obtained from standard methods may be biased.

Inverse probability-of-censoring weighted estimation was reviewed as a technique to correct for selection bias due to loss to follow up when estimating absolute measures or relative effect measures. Compared with nonstandard techniques, such as stratification-based methods, weighted methods can correct for selection bias in a broader number of scenarios and more readily provide covariate-corrected marginal estimates. However, when necessary assumptions or conditions are potentially violated, alternative techniques such as targeted learning should be considered.^{17,28,33}

ACKNOWLEDGMENTS

The authors thank Dr. Bianca De Stavola for helpful feedback on an earlier draft, Dr. Daniel Westreich for expert advice, Mr. Sam Stinnette for assistance with the UNC CFAR

HIV clinical cohort data, and the rest of the UNC CFAR HIV clinical cohort study staff.

REFERENCES

- Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
- Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
- Hernán MA, Robins J. *Causal Inference Book*. Boca Raton, FL: Chapman & Hall/CRC; Forthcoming 2016.
- Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168:656–664.
- Howe CJ, Cole SR, Chmiel JS, Munoz A. Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *Am J Epidemiol*. 2011;173:569–577.
- Howe CJ, Cole SR, Mehta SH, Kirk GD. Estimating the effects of multiple time-varying exposures using joint marginal structural models: alcohol consumption, injection drug use, and HIV acquisition. *Epidemiology*. 2012;23:574–582.
- Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11:561–570.
- Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*. 2000;56:779–788.
- Hernán MA, McAdams M, McGrath N, Lanoy E, Costagliola D. Observation plans in longitudinal studies with time-varying treatments. *Stat Methods Med Res*. 2009;18:27–52.
- Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
- Barnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology*. 2011;22:27–35.
- Robins J. A new approach to causal inference in mortality studies with sustained exposure periods: application to control of the healthy worker survivor effect. *Math Model*. 1986;7:1393–1512.
- Malani HM. A modification of the redistribution to the right algorithm using disease markers. *Biometrika*. 1995;82:515–526.
- Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics*. 1996;52:137–151.
- Scharfstein DO, Robins JM. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*. 2002;89:617–634.
- Hsu CH, Taylor JM, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Stat Med*. 2006;25:3503–3517.
- Neugebauer R, Schmittdiel JA, van der Laan MJ. Targeted learning in real-world comparative effectiveness research with time-varying interventions. *Stat Med*. 2014;33:2480–2520.
- Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat Methods Med Res*. 2012;21:243–256.
- Hernán MA. The hazards of hazard ratios [commentary]. *Epidemiology*. 2010;21:13–15.
- DeGroot MH, Schervish MJ. Unbiased estimators. In: Lynch D, Guardino K, eds., *Probability and Statistics*. 3rd ed. Boston, MA: Addison-Wesley; 2001:427–433.
- Hernán MA. Caveats and Considerations in Symposium on Selection Bias due to Loss: An Old and Often Ignored Problem Revisited at 2014 Society for Epidemiologic Research Annual Meeting. Available at: <https://epiresearch.org/about-us/archives/video-archives-2/selection-bias-due-to-loss/>. Accessed February 6, 2015.
- Greenland S, Pearl J. Adjustments and their consequences-collapsibility analysis using graphical models. *Int Stat Rev*. 2011;79:401–426.
- Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiology*. 2012;23:159–164.
- Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002;155:176–184.
- Collett D. *Modelling Survival Data in Medical Research*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC; 2003.
- Cole SR, Lau B, Eron JJ, et al. Estimation of the standardized risk difference and ratio in a competing risks framework: application to injection drug use and progression to AIDS after initiation of antiretroviral therapy. *Am J Epidemiol*. 2015;181:238–245.
- Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol*. 2010;171:674–677; discussion 678–681.
- Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*. 2012;21:31–54.
- Hudgens MG, Halloran ME. Toward causal inference with interference. *J Am Stat Assoc*. 2008;103:832–842.
- VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009;20:880–883.
- Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for non-ignorable drop-out using semi-parametric non-response models. *J Am Statist Assoc*. 1999;94:1096–1120.
- Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for non-ignorable drop-out using semi-parametric non-response models [comments and rejoinder]. *J Am Statist Assoc*. 1999;94:1121–1146.
- Gruber S, Logan RW, Jarrin I, Monge S, Hernán MA. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Stat Med*. 2015;34:106–117.
- The Antiretroviral Therapy Cohort Collaboration. Influence of geographical origin and ethnicity on mortality in patients on antiretroviral therapy in Canada, Europe, and the United States. *Clin Infect Dis*. 2013;56:1800–1809.
- VanderWeele TJ, Robinson WR. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*. 2014;25:473–484.
- Howe CJ, Cole SR, Napravnik S, Eron JJ. Enrollment, retention, and visit attendance in the University of North Carolina Center for AIDS Research HIV clinical cohort, 2001–2007. *AIDS Res Hum Retroviruses*. 2010;26:875–881.
- Vourli G, Touloumi G. Performance of the marginal structural models under various scenarios of incomplete marker's values: a simulation study. *Biom J*. 2014;28:201300159.
- Howe CJ, Cole SR, Napravnik S, et al. The role of at-risk alcohol/drug use and treatment in appointment attendance and virologic suppression among HIV(+) African Americans. *AIDS Res Hum Retroviruses*. 2014;30:233–240.
- Howe CJ, Napravnik S, Cole SR, et al. African American Race and HIV virological suppression: beyond disparities in clinic attendance. *Am J Epidemiol*. 2014;179:1484–1492.

eAPPENDIX 1

Simulations were performed in SAS, version 9.4, software (SAS Institute, Inc., Cary, North Carolina) to confirm and demonstrate the expected absence or presence of bias of the absolute measures and relative effect measures for the five causal diagrams shown in Figure 1 where the absolute measure corresponds to survival from death and the relative effect measure corresponds to the risk difference, risk ratio, and odds ratio for the effect of injection drug use on death. The risk difference, risk ratio, and odds ratio were derived from the complement of the survival function¹. For all examined scenarios, 500 simulations of sample size 1,000 were performed. Continuous times from study entry to death and censoring due to loss to follow up were generated from exponential distributions (i.e., $S(t_{continuous}) = \exp\{-(t_{continuous} / \mu_1)\}$ and $S(c_{continuous}) = \exp\{-(c_{continuous} / \mu_2)\}$) where μ_1 and μ_2 (the mean event times) for the baseline survival function for death and censoring due to loss were 20.09 and 6.69, respectively. Continuous times from study entry to death and loss were coarsened, used to generate observed follow up times along with loss to follow up and death indicators, and mapped to monthly study visits where visit 36 was considered to be the administrative end of study follow-up. Discrete-time methods² were used to obtain the true survival functions, risk differences, risk ratios, and odds ratios as well as the corresponding estimates based on standard approaches that do not account for potential selection bias and inverse probability-of-censoring weighted approaches that do account for potential selection bias.

For all simulated scenarios, injection drug use (*A*) and heavy alcohol use (*L*) were assumed to be measured once and correspond to behavior in the 6 months prior to study entry, while, when relevant, CD4 cell count (*Q*) corresponded to one year prior to study entry and education (*Z*) was

the level of education at the time of HIV infection. Further injection drug use was simulated to increase the likelihood of death. For Diagrams I) through IV) times from study entry to death and loss were generated solely as a function of injection drug use, heavy alcohol use, both, or neither. Specifically, times from study entry to death were generated as a function of injection drug use and heavy alcohol use such that injection drug use and heavy alcohol use each independently changed μ_1 by a factor of 0.30 and 0.41, respectively. The prevalence of both injection drug use and heavy alcohol use was set to 50%.

In Diagram I), approximately 29% of the study population was lost where times from study entry to loss were generated independently of injection drug use and heavy alcohol use or any other factor. In Diagram II), approximately 49% of the study population was lost where times from study entry to loss were generated solely as a function of heavy alcohol use such that heavy alcohol use changed μ_2 by a factor of 0.27. In Diagram III), approximately 48% of the study population was lost where times from study entry to loss were generated solely as a function of injection drug use such that injection drug use changed μ_2 by a factor of 0.27. For Diagram IV), approximately 59% of the study population was lost where times from study entry to loss were generated as a function of injection drug use and heavy alcohol use such that injection drug use and heavy alcohol use each independently changed μ_2 by a factor of 0.27 and 0.41, respectively.

For Diagram V) both the prevalence of having a CD4 cell count ≥ 200 cells/microL and not being college educated was set to 50%. Injection drug use was generated as a function of CD4 cell count where individuals with a CD4 cell count ≥ 200 cells/microL were 1.5 times more likely to inject drugs compared to individuals with a CD4 cell count < 200 cells/microL. Heavy alcohol use was generated as a function of CD4 cell count and education. Independent of

education level, persons with a CD4 cell count ≥ 200 cells/microL were three times more likely to engage in heavy alcohol use than persons with a CD4 cell count < 200 cells/microL. Similarly, independent of CD4 cell count, those without a college education were three times more likely to engage in heavy alcohol use than persons with a college education.

Next for Diagram V), time to death was generated as a function of injection drug use and education such that injection drug use and not being college educated each independently changed μ_1 by a factor of 0.27. Time to loss was generated as a function of injection drug use and heavy alcohol use such that injection drug use and heavy alcohol use each independently changed μ_2 by a factor of 0.37. Approximately 53% of the study population was lost.

Figures A1.1 through A1.5 show the bias and mean squared error of the standard estimator that does not account for potential selection bias (i.e., crude) and the inverse probability-of-censoring weighted approach that does for the five causal diagrams shown in Figure 1. For a given estimator the bias was assessed by comparing the mean estimate across the 500 simulations (i.e., mean survival function, mean risk difference, mean log risk ratio, mean log odds ratio) to the true corresponding value. The mean squared error was calculated as the square of the difference between the mean estimate across the 500 simulations and the true value plus the variance of the estimate across the 500 simulations. As expected, in Figure A1.1, there is an absence of bias for the standard estimator for both the absolute measure and relative effect measures. In Figure A1.2, the absolute measure and the risk difference were substantially biased based on the standard estimator. In Figure A1.3, only the absolute measure is biased based on the standard estimator. In Figures A1.4 and A1.5 the absolute measure and all relative effect measures tended to be biased based on the standard estimator. The inverse probability-of-censoring weighted estimator that accounted for the potential selection bias tended to be less

biased and have a smaller mean squared error than the standard estimator. Any bias associated with the inverse probability-of-censoring weighted estimator was likely due to potential violations in the positivity assumption when the percentage of lost to follow up was large. Furthermore, in terms of the survival function, the lower mean squared error for the standard estimator compared to the inverse probability-of-censoring weighted estimator around visit 6 in Figures A1.2 through A1.5 is likely due the crossing of the standard estimator and true survival functions as the standard estimator changes from under to overestimating the true survival function.

An additional 500 simulations of sample size 1,000 were performed in SAS, version 9.4 to confirm and demonstrate the expected presence of bias of the absolute measure and relative effect measures for the causal diagram shown in Graph I) in Figure 2 where the absolute measure corresponds to survival from death and the relative effect measures correspond to the risk difference, risk ratio, and odds ratio for the effect of African American race on death which was simulated such that African American race increased the likelihood of death. Continuous times from study entry to death were generated from a weibull distribution (i.e.,

$S(t_{continuous}) = \exp\{-[t_{continuous} / \theta]^{1/\sigma}\}$) where θ and σ for the baseline survival function for death was 12.18 and 0.8, respectively. The σ was assumed to be the same for the baseline and non-baseline survival functions and corresponded to a monotonically increasing hazard of death.

Continuous times from study entry to censoring due to loss to follow up were generated from an exponential distribution (i.e., $S(c_{continuous}) = \exp\{-[c_{continuous} / \phi]\}$) where ϕ (the mean event time) for the baseline survival function for censoring due to loss was 3.67. Similar to the Figure 1 simulations, data were coarsened, used to generate observed follow up times along with loss to

follow up and death indicators, mapped, and then discrete-time methods were used to obtain the true survival function, risk difference, risk ratio, and odds ratio as well as the corresponding estimates based on standard and inverse probability-of-censoring weighted approaches.

Approximately 68% of the study population was lost. Times from study entry to death were generated as a function of African American race and a single binary composite variable (1=yes, 0=no) representing the net effect of CD4 cell count, a prior AIDS-defining illness diagnosis, HIV RNA level, and insurance status at study entry on time to death. African American race and the composite variable each independently changed θ by a factor of 0.49 and 0.57, respectively. Similarly, times from study entry to loss were generated as a function of African American race and the single binary composite variable. African American race and the composite variable each independently changed ϕ by a factor of 0.58 and 0.50, respectively. African Americans were assumed to be twice as likely to have a value of 1 for the composite variable.

Figure A1.6 shows the bias and mean squared error of the standard estimator that does not account for potential selection bias (i.e., crude) as well as the inverse probability-of-censoring weighted approach that does attempt to account for the potential selection bias in the causal diagram shown in Graph I) of Figure 2 based on African American race and the composite variable. The bias and mean squared error were assessed and calculated as done in Figures A1.1 through A1.5. In Figure A1.6, both the absolute measure and relative effect measures tended to be biased based on the standard estimator that ignores the potential selection bias. The inverse probability-of-censoring weighted estimator that accounted for the potential selection bias tended to be less biased and have a smaller mean squared error than the standard estimator that ignored the potential selection bias. In terms of the survival function, the lower mean squared error for the standard estimator compared to the inverse probability-of-censoring weighted estimator

around visit 6 in Figure A1.6 is likely due the crossing of the standard estimator and true survival functions as the standard estimator changes from under to overestimating the true survival function.

Figure A1.1 Bias and mean squared error for absolute measure (i.e., survival) and relative effect measure (i.e., risk difference, log risk ratio, and log odds ratio) for Diagram I) in Figure 1 based on 500 simulations each with a sample size of 1,000.

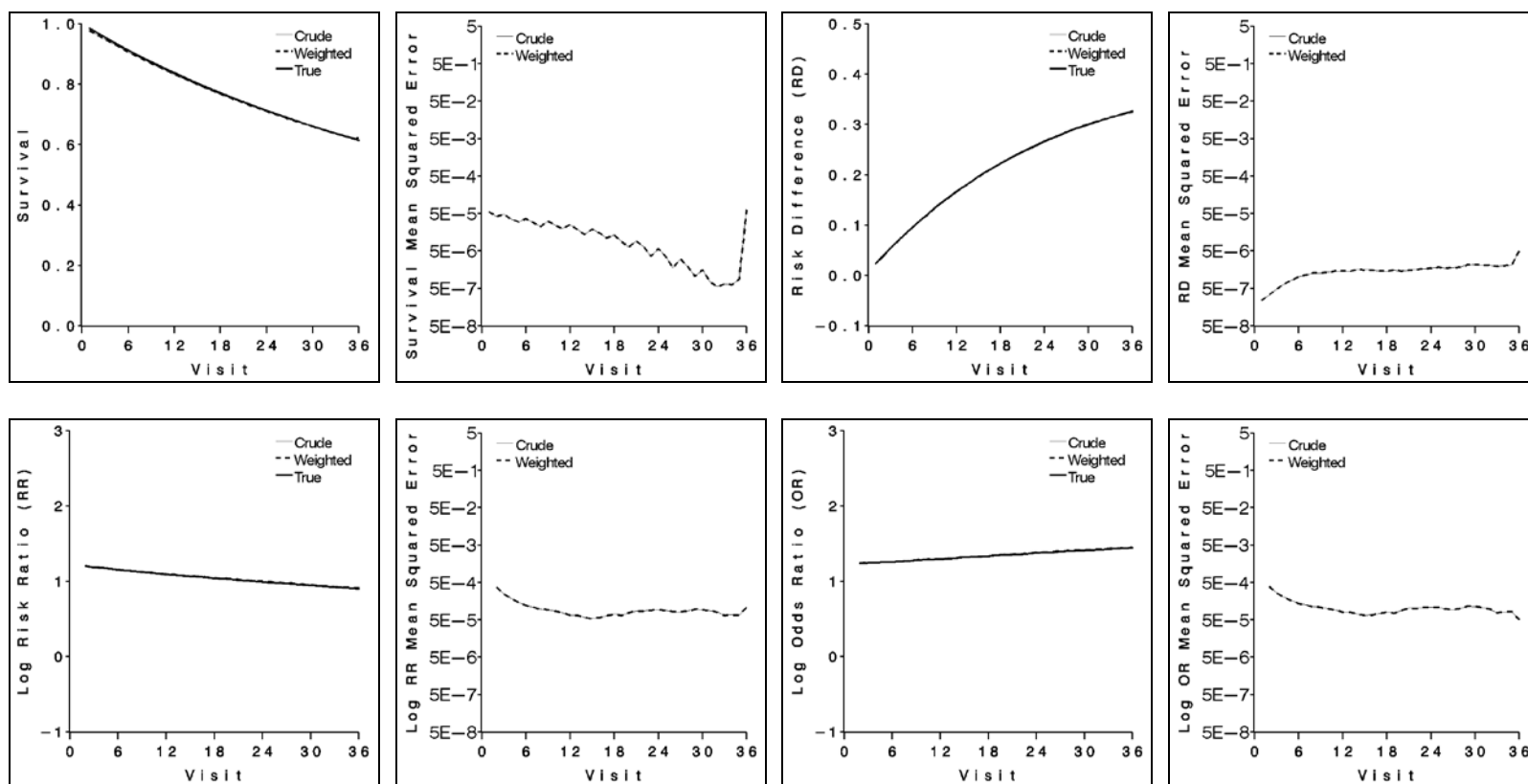
Figure A1.2 Bias and mean squared error for absolute measure (i.e., survival) and relative effect measure (i.e., risk difference, log risk ratio, and log odds ratio) for Diagram II) in Figure 1 based on 500 simulations each with a sample size of 1,000.

Figure A1.3 Bias and mean squared error for absolute measure (i.e., survival) and relative effect measure (i.e., risk difference, log risk ratio, and log odds ratio) for Diagram III) in Figure 1 based on 500 simulations each with a sample size of 1,000.

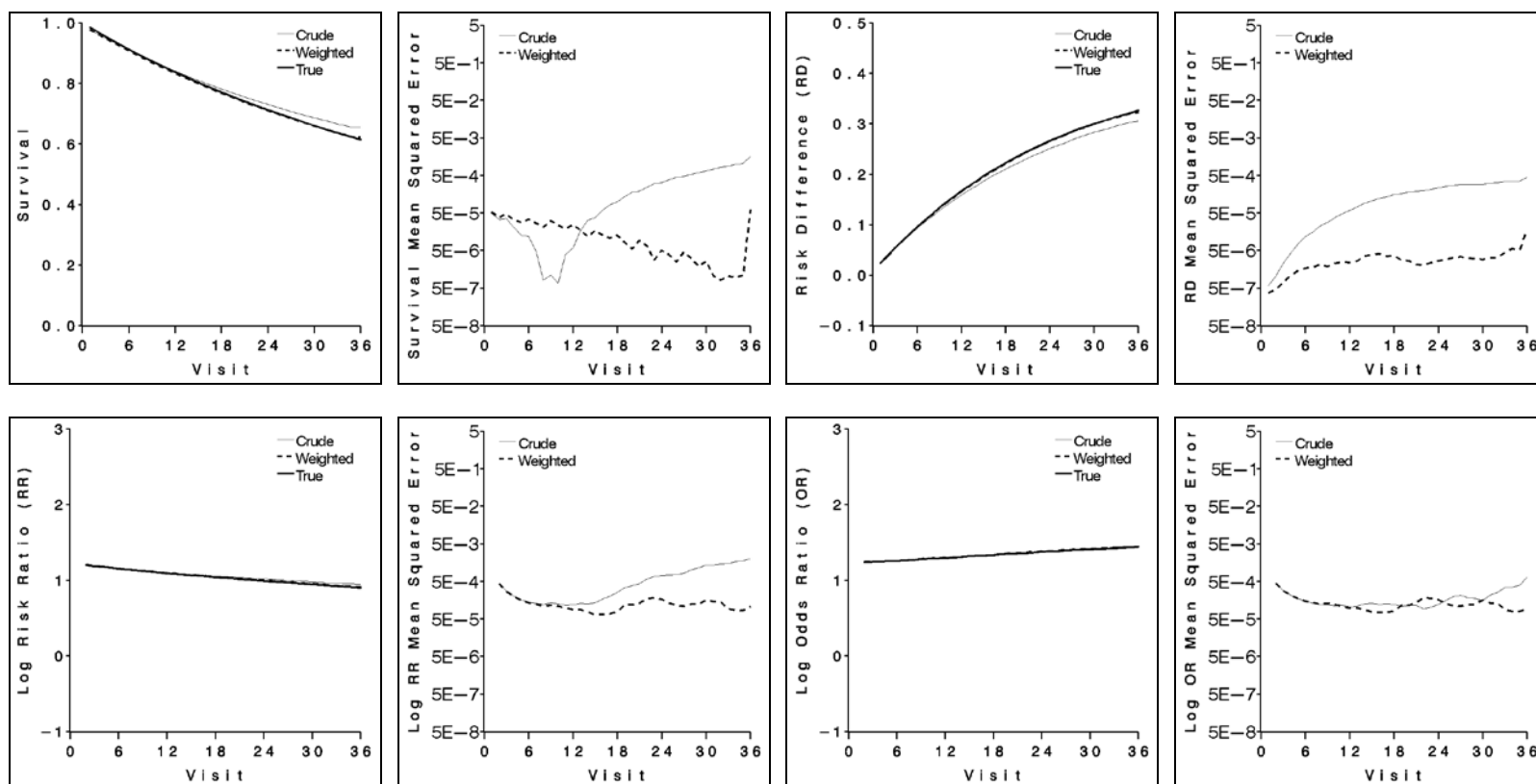
Figure A1.4 Bias and mean squared error for absolute measure (i.e., survival) and relative effect measure (i.e., risk difference, log risk ratio, and log odds ratio) for Diagram IV) in Figure 1 based on 500 simulations each with a sample size of 1,000.

Figure A1.5 Bias and mean squared error for absolute measure (i.e., survival) and relative effect measure (i.e., risk difference, log risk ratio, and log odds ratio) for Diagram V) in Figure 1 based on 500 simulations each with a sample size of 1,000.

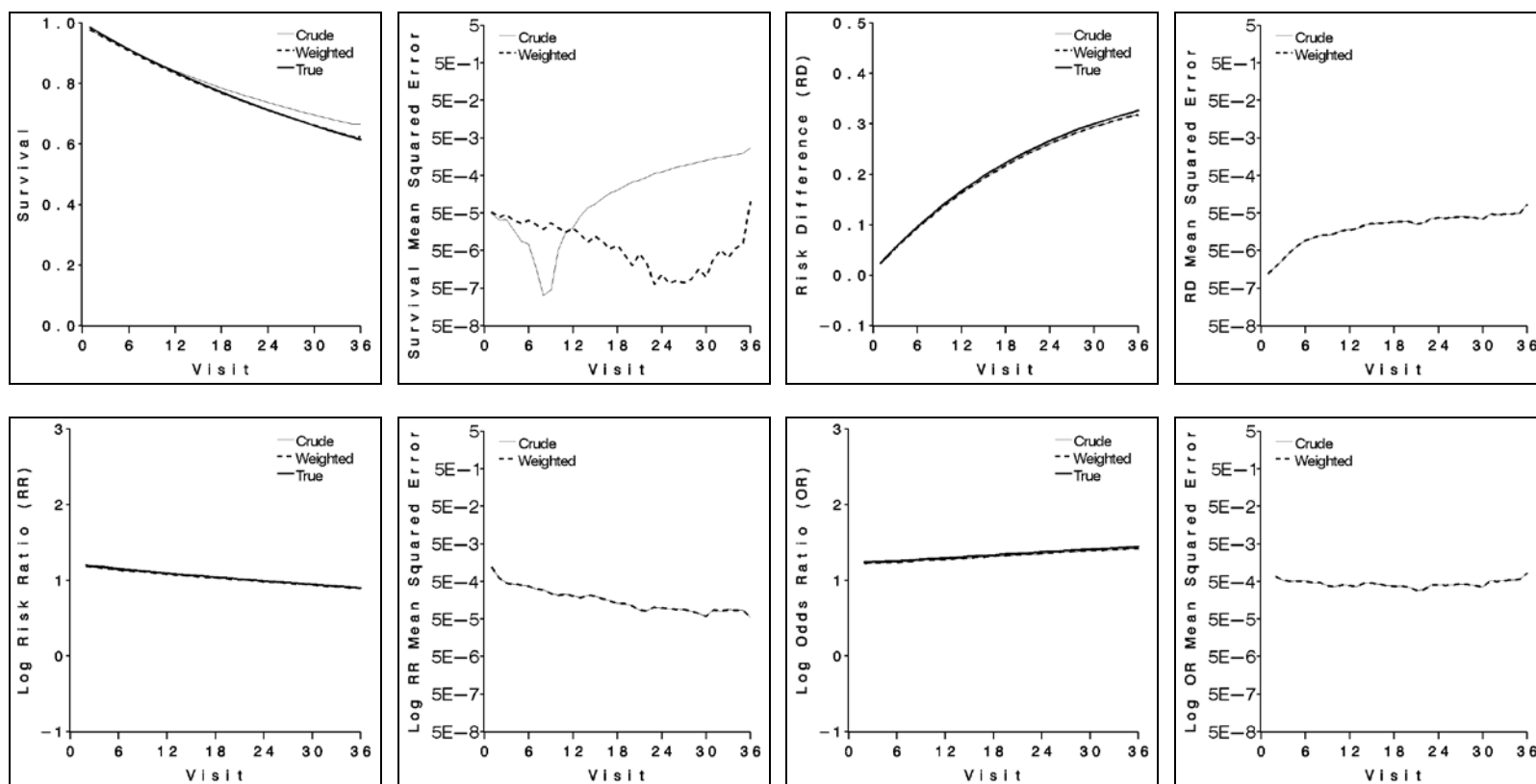
Figure A1.6 Bias and mean squared error for absolute measure (i.e., survival) and relative effect measure (i.e., risk difference, log risk ratio, and log odds ratio) for Diagram I) in Figure 2 based on 500 simulations each with a sample size of 1,000.



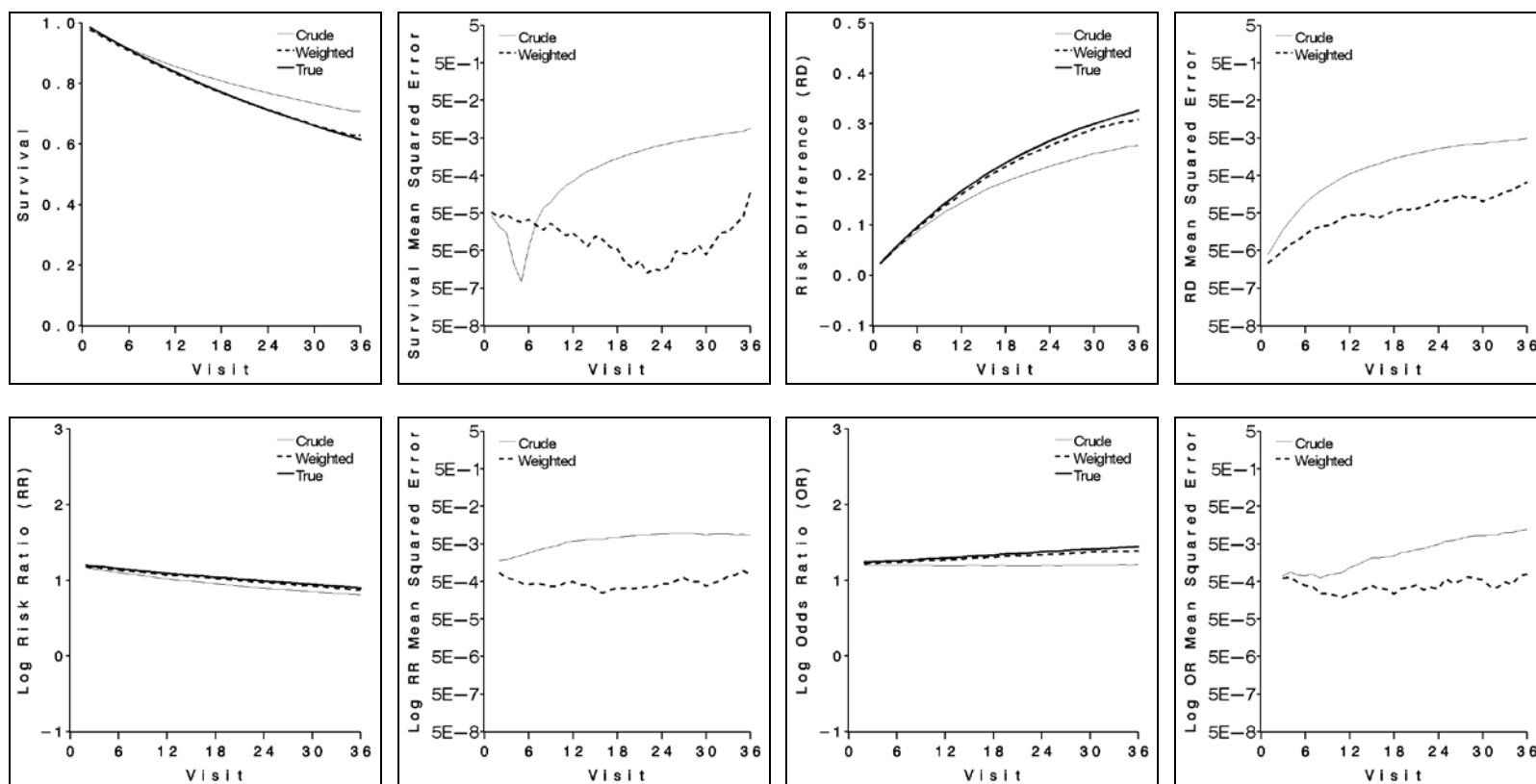
RD, risk difference; RR, risk ratio; OR, odds ratio



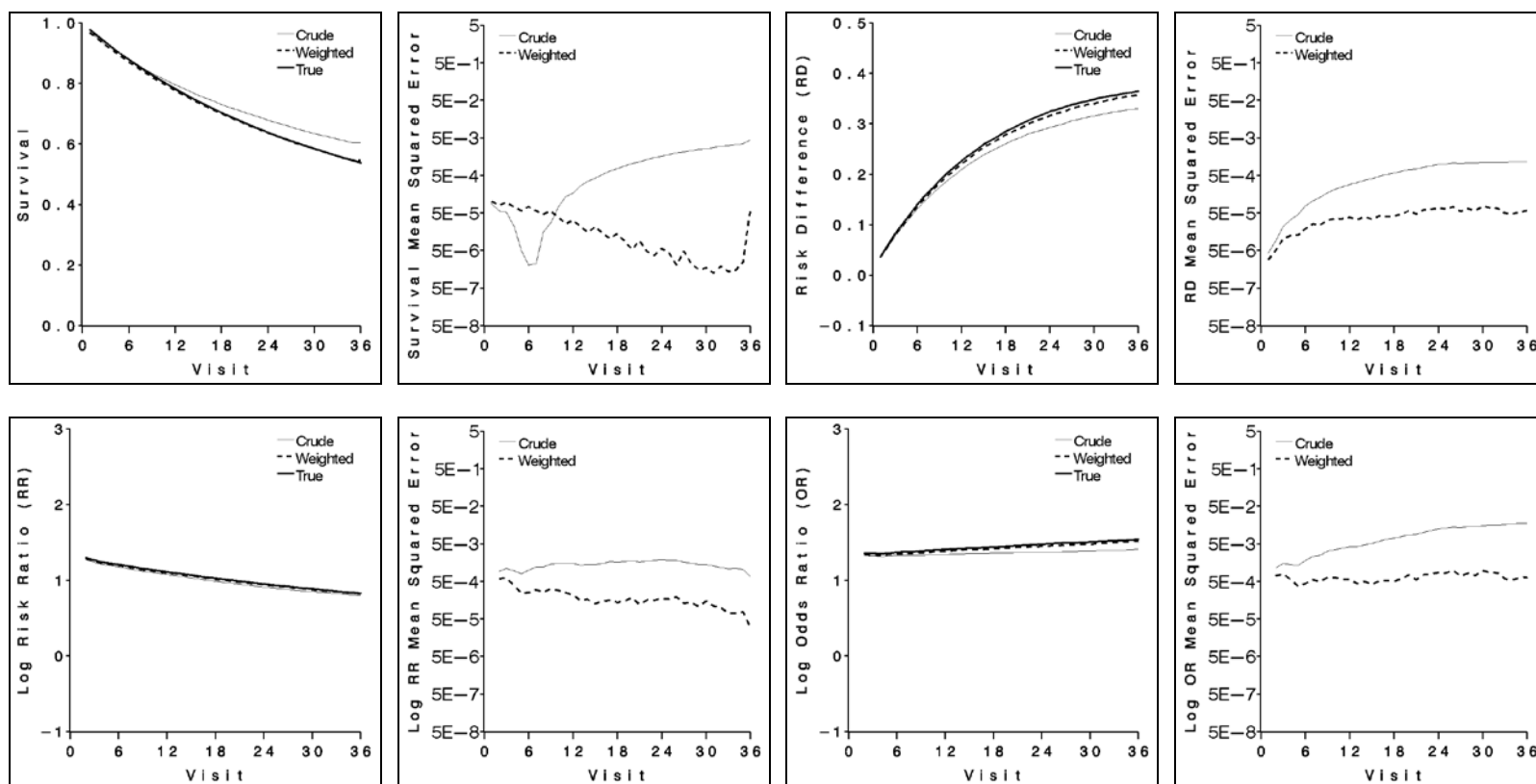
RD, risk difference; RR, risk ratio; OR, odds ratio



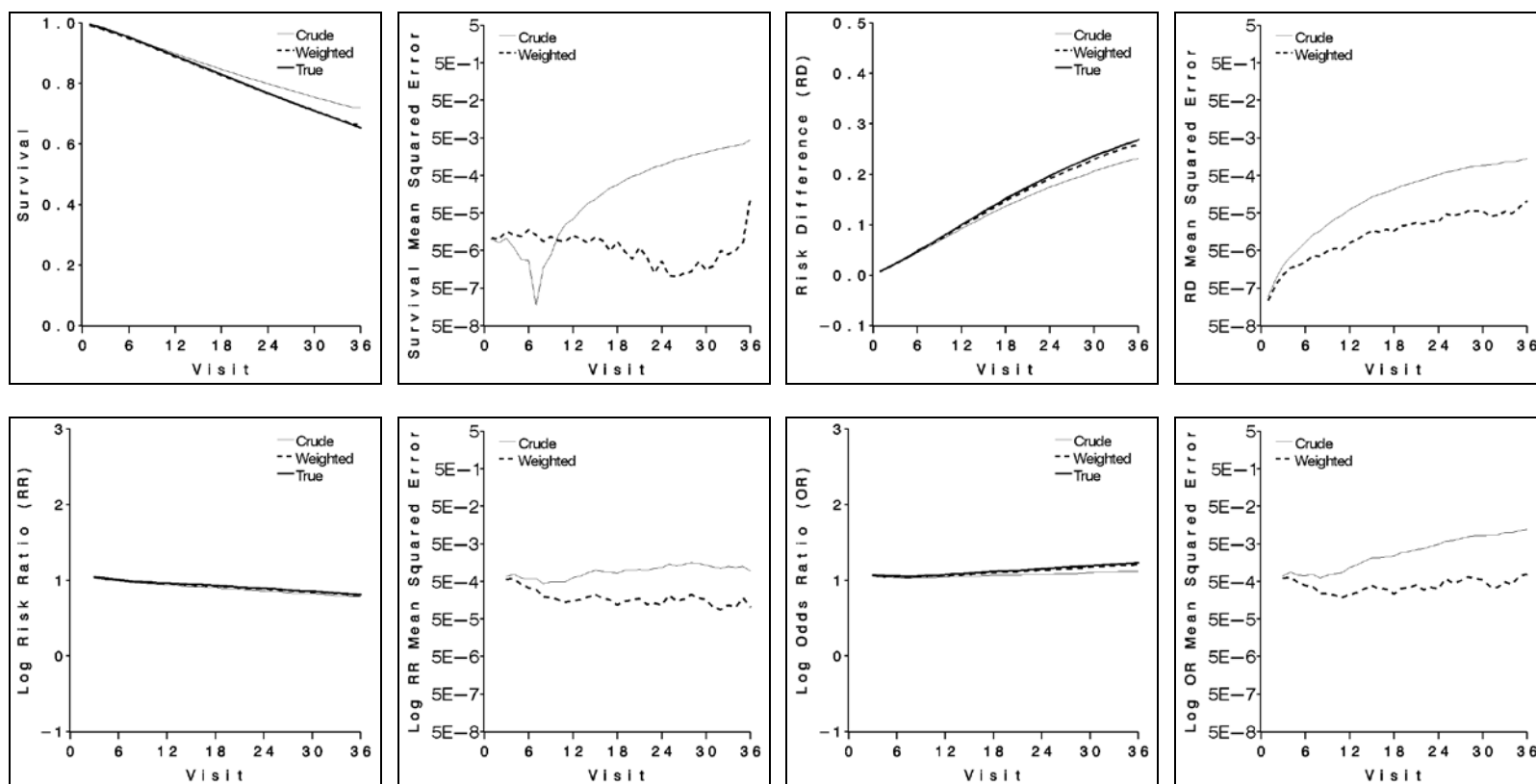
RD, risk difference; RR, risk ratio; OR, odds ratio



RD, risk difference; RR, risk ratio; OR, odds ratio



RD, risk difference; RR, risk ratio; OR, odds ratio



RD, risk difference; RR, risk ratio; OR, odds ratio

eAPPENDIX 2

Figure A2.1 shows the follow up data of 8 participants in a cohort study with loss to follow up. The objective of the study was to estimate survival after study entry as well as the difference in survival as a function of injection drug use via the risk difference or risk ratio. The 8 participants were randomly sampled from a larger population. If Diagram IV) in Figure 1 is assumed to represent the causal relationships between injection drug use at study entry, heavy alcohol use at study entry, loss to follow up after study entry, and time to death in the analysis sample then methods such as inverse probability-of-censoring weighted estimation should be used to correct for potential bias of absolute and relative effect measures.

The inverse probability-of-censoring weighted method can be used to appropriately estimate absolute and relative effect measures by creating the pseudo-population that would have been observed had losses to follow up not occurred among the 8 participants. In the absence of losses in the pseudo-population the direct arrows from A and L to D would be removed from Diagram IV) in Figure 1 yielding Diagram I). This pseudo-population is created by re-weighting the contribution of each participant who was not lost to follow up to a given risk set. Specifically, at time u each participant is assigned a weight $W(u)$ that is inversely proportional to the estimated probability that the participant remained not lost to follow up through time u conditional on measured determinants of loss to follow up including the exposure (if applicable). This conditional probability and weight, $W(u)$ can be estimated non-parametrically in the context of low-dimensional follow up data. When data are high-dimensional due to a large amount of losses at different follow up times or losses are informative based on many (or continuous) covariates, a pooled logistic regression model for not being lost to follow up can be fit and used to estimate $W(u)$ ³.

Non-parametric methods were used to estimate $W(u)$ for the Figure A2.1 example where $W(u)$ is defined as

$$W(u) = \begin{cases} \prod_{k=1}^u \frac{1}{P[D(k)=0 | \bar{D}(k-1)=0, \bar{O}(k-1)=0, A, L]}, & \text{if } D(u) = 0 \\ 0, & \text{if } D(u) = 1 \end{cases} \quad (1).$$

Note to estimate $W(u)$ based on equation (1), the data in Figure A2.1 were coarsened into 1-unit time intervals to correspond to study visits where $u = 1, 2, \dots, 15$. Use of pooled logistic regression models in the context of high dimensional data may also require a similar coarsening of the follow up data. For observed times prior to lost to follow up, the denominator of $W(u)$ is a participant's probability of remaining not lost to follow up through time u given A and L where $\bar{D}(k-1) = 0$ and $\bar{O}(k-1) = 0$ in equation (1) respectively specify that the participant was not lost to follow up and did not develop the event prior to time k . To preclude loss to follow up in the pseudo-population weighted by $W(u)$, $W(u) = 0$ for times at or after lost to follow up.

Both A and L were used to solely estimate the denominator of $W(u)$ so that neither A nor L determined D in the pseudo-population where loss to follow up did not occur. $W_i(u)$ is the number of participants who are like individual i in terms of their values of A and L that would have been in the risk set at time u in the absence of losses. Individuals who are not lost but have the highest probability of being lost are more greatly up-weighted in the pseudo-population to represent their peers with the same values for A and L who were lost.

Figure A2.2 shows stratification-based non-parametric methods to estimate $W(u)$. The data were first stratified by every observed combination of the levels of A and L and $W(u)$ was estimated for each participant at all relevant u 's using equation (1). Based on $W(u)$, participant

5 in the $A = 0$ and $L = 0$ stratum represents 1 individual for $1 \leq u \leq 3$, then after participant 2 is lost to follow up at $u = 4$, for $4 \leq u \leq 11$ participant 5 represents 1.5 individuals, their self plus half of participant 2. The other half of participant 2 is represented by participant 6. Similar to participant 5, prior to $u = 4$ participant 6 represented 1 person.

Figure A2.3 shows the re-weighted follow up data based on $W(u)$ after study entry among the 8 participants in the previously described cohort study. In the observed data in Figure A2.1 there were 3 deaths, 2 losses to follow up, and 3 persons who reached the administrative end of the study alive. However, in the pseudo-population there are 4.5 deaths, 0 losses to follow up, and 3.5 persons who reached the administrative end of the study alive. Therefore, the pseudo-population based on $W(u)$ is the follow up data that would have been observed in the absence of losses.

To help preserve the amount of information in the observed data (e.g., the number of events) and minimize the variability of the weights due to non-positivity, weights are typically stabilized by replacing the numerator of 1 in equation (1) with the conditional probability of not being lost to follow up given the exposure (if applicable), in this case A ^{3,4}. Here we define the stabilized weight as

$$SW(u) = \prod_{k=1}^u \frac{P[D(k) = 0 \mid \bar{D}(k-1) = 0, \bar{O}(k-1) = 0, A]}{P[D(k) = 0 \mid \bar{D}(k-1) = 0, \bar{O}(k-1) = 0, A, L]} \quad (2).$$

For observed times prior to or at lost to follow up, the denominator of $SW(u)$ is a participant's probability of remaining not lost to follow up through time u given A and L . Similarly, the numerator is a participant's probability of remaining not lost to follow up through time u given A . To allow individuals who are lost to receive a non-zero weight when they exit from the risk set and in turn be able to calculate the number of losses in the pseudo-population weighted by

$SW(u)$, $SW(u) = 0$ only for times after loss to follow up. Assigning $SW(u) = 0$ only for times after loss to follow up is also consistent with the commonly made assumption in discrete-time survival analysis that losses occur after an event occurs when there are tied event and censoring times⁵.

This stabilization creates the pseudo-population that would have been observed had losses been random with respect to L or any other variable solely used in the denominator of equation (2). Individuals who are not lost and have a higher probability of being lost than other cohort members with the same level of A but a different level of L receive a higher weight in terms of their contribution to the risk set pseudo-population. Conversely, individuals who are not lost and have a lower probability of being lost than other cohort members with the same level of A but a different level of L receive a lower weight in terms of their contribution to the risk set pseudo-population. In the case of the Figure A2.1 data, the stabilized pseudo-population in Figure A2.3 based on equation (2) should now correspond to Diagram III) rather than Diagram IV) in Figure 1.

Similar to the unstabilized weight, $W(u)$, both non-parametric and parametric methods can be used to estimate the stabilized weight, $SW(u)$. Non-parametric stratification was used to estimate $SW(u)$ for the Figure A2.1 example based on equation (2). Similar to non-parametrically estimating $W(u)$, the data were first stratified by every observed combination of the levels of A and L as well as solely by the observed level of A . Next, $SW(u)$ was estimated for each participant at all relevant u 's using equation (2) based on the stratified data.

Figure A2.3 shows the re-weighted follow up data based on $SW(u)$ derived from equation (2) after study entry among the 8 previously described cohort participants. In the pseudo-population

based on $SW(u)$ there are 3.5 deaths, 2.5 losses to follow up, and 2.5 persons who reached the administrative end of the study alive. As the sample size increases and assuming all necessary assumptions are met, the number of events, losses, and persons who reached the administrative end of the study event-free in the pseudo-population should approach the number of events, losses, and persons who reached the administrative end of the study event-free in the observed data.

Equation (4) shows the pooled logistic regression model that can be used to parametrically estimate the denominator of the unstabilized weights for a given participant at time u . Similarly, the pooled logistic regression models in equations (3) and (4) can be used to parametrically estimate the numerator and denominator of the stabilized weights shown in equation (2).

$$\text{logit } P[D(k) = 0 | \bar{D}(k-1) = 0, \bar{O}(k-1) = 0, A] = \alpha_{0k} + \alpha_1 A \quad (3)$$

$$\text{logit } P[D(k) = 0 | \bar{D}(k-1) = 0, \bar{O}(k-1) = 0, A, L] = \beta_{0k} + \beta_1 A + \beta_2 L \quad (4)$$

In equations (3) and (4) $\text{logit } p = \ln[p / (1 - p)]$. The parameters α_{0k} and β_{0k} are the time specific intercepts without and with inclusion of L in the pooled model, respectively, where k is time. The parameter α_1 is the log hazard odds ratio for the effect of A on D , β_1 is the log hazard odds ratio for the effect of A on D adjusting for L , and β_2 is the log hazard odds ratio for the effect of L on D adjusting for A .

Assuming necessary assumptions are met, equations (5) and (6) which have been adapted from Robins and Finkelstein⁶ may be used to estimate the inverse probability-of-censoring

weighted survival function, $\hat{S}(u)$, corrected for potential selection bias due to losses to follow up.

$$\hat{\rho}(k) = \frac{\sum_{i \in V_k} \hat{W}_i(k)}{\sum_{i \in \tau_k} \hat{W}_i(k)} \quad (5)$$

$$\hat{S}(u) = \prod_{k=1}^u [1 - \hat{\rho}(k)] \quad (6)$$

In equations (5) and (6) τ_k is the subset of the cohort at entry that is in the risk set at time k , while V_k is the subset of τ_k that develops the event at time k . Note $S\hat{W}_i(k)$ can replace $\hat{W}_i(k)$ in equations (5) and (6) as long as A is not used when estimating the numerator of the weights in equation (2). Standard errors for $\hat{S}(u)$ can be obtained via bootstrapping, recalculating the weights on each resample⁷. When a relative effect estimate such as the risk difference or risk ratio is the quantity of interest, the inverse probability-of-censoring weighted risk difference and risk ratio can be obtained via $[1 - \hat{S}_{A=1}(u)] - [1 - \hat{S}_{A=0}(u)]$ and $[1 - \hat{S}_{A=1}(u)] / [1 - \hat{S}_{A=0}(u)]$, respectively, where $\hat{S}_{A=1}(u)$ is the inverse probability-of-censoring weighted survival function solely estimated among those who engaged in injection drug use in the 6 months prior to study entry, while $\hat{S}_{A=0}(u)$ is the inverse probability-of-censoring weighted survival function solely estimated among those who did not engage in injection drug use in the 6 months prior to study entry^{1,8}.

Note that although unbiased relative effect estimates can be obtained via inverse probability-of-censoring weighted methods without removing the arrow from A to D in Diagram IV) as done with the stabilized weights estimated for Figure A2.3 based on Equation (2), the exposure A must

at least be used to estimate the denominator of the weights given that the quantity of interest corresponds to a joint effect. Specifically, the effect of intervening on both the exposure A and on censoring due to loss to follow up^{4,9}. However, if A is a common cause of the event time and losses to follow up like in Diagram IV), then the direct arrow from A to D must be removed through the weighting to obtain an unbiased absolute measure even though the direct arrow from A to D does not need to be removed to obtain an unbiased relative effect measure since within strata of A losses should be random. The arrow from A to D could be removed by excluding A from the numerator in Equation (2) yielding Equation (7)

$$SW(u) = \prod_{k=1}^u \frac{P[D(k) = 0 \mid \bar{D}(k-1) = 0, \bar{O}(k-1) = 0]}{P[D(k) = 0 \mid \bar{D}(k-1) = 0, \bar{O}(k-1) = 0, A, L]} \quad (7).$$

The weights in Equation (7) create the pseudo-population that would have been observed had losses been random with respect to both A and L . Individuals who are not lost but have a higher probability of being lost, given their observed levels of A and L , compared to the entire cohort receive a higher weight in terms of their contribution to the risk set pseudo-population. Conversely, individuals who are not lost but have a lower probability of being lost, given their observed levels of A and L , compared to the entire cohort receive a lower weight in terms of their contribution to the risk set pseudo-population. In the case of the Figure A2.1 data the pseudo-population in Figure A2.3 based on equation (7) would now correspond to Diagram I) rather than Diagram IV) in Figure 1. Similar to the equation (2) stabilized weights, as the sample size increases and assuming all necessary assumptions are met, the number of events, losses, and persons who reached the administrative end of the study event-free in the equation (7) pseudo-population should approach the number of events, losses, and persons who reached the administrative end of the study event-free in the observed data.

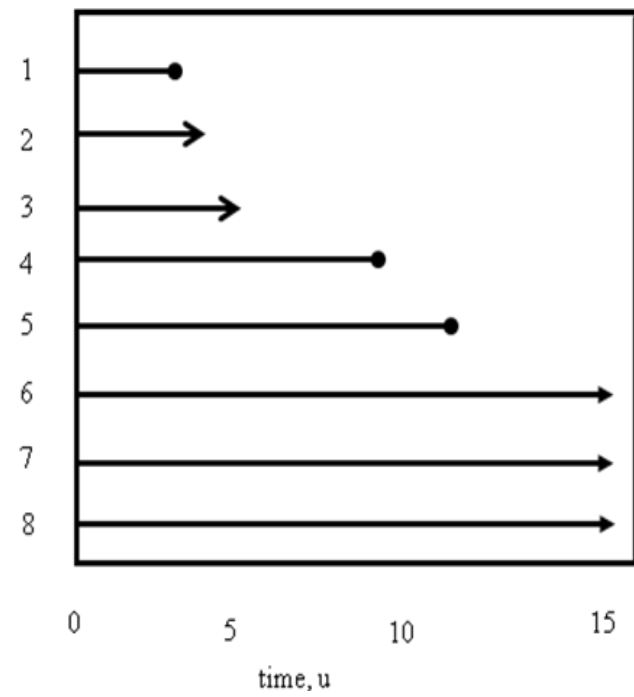
Also important to note is that stratification-based methods such as stratification or standard regression adjustment could also be used to address potential absolute or relative selection bias in Diagrams II) to IV) by conditioning the analysis on either A or L or both. However, in the case of Diagram V) stratification-based methods cannot be used to address potential relative selection bias since although conditioning on L blocks the non-causal path between A and T via D , the non-causal path between A and T via Q , L , and Z is now open and cannot be addressed given that Q and Z are not measured. The weights in Equation (7) could be used to address potential absolute and relative selection bias in Diagram V), however.

Figure A2.1. Follow up data after study entry among 8 participants in a cohort study. The time scale is visits since study entry where time is indexed by u . In the left table, i is the subject identifier, Y is the observed follow up time, D is an indicator of loss to follow up when $u = y$, O is an indicator of the occurrence of the event when $u = y$, A is an indicator of injection drug use in the prior 6 months at study entry, and L is an indicator of heavy alcohol use in the prior 6 months at study entry. In the right diagram, open right arrows represent censoring due to loss to follow up, closed right arrows represent censoring due to reaching the administrative end of the study, and closed dots represent the occurrence of the event.

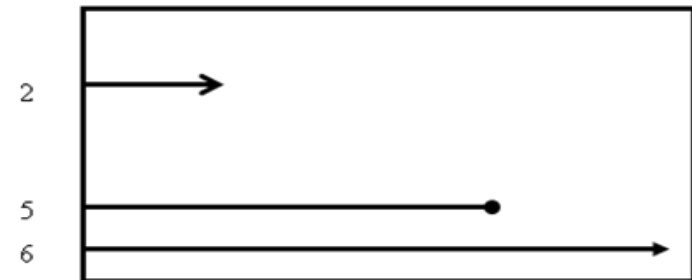
Figure A2.2. Follow up data after study entry among 8 participants in a cohort study by level of injection drug use (A) and heavy alcohol use (L). The time scale is visits since study entry where time is indexed by u . In the left table, i is the subject identifier, u_{enter} is the time from study entry at the start of a defined time interval, u_{exit} is the time from study entry at the end of a defined time interval, D is an indicator of loss to follow up at u_{exit} , O is an indicator of the occurrence of the event at u_{exit} , A is an indicator of injection drug use in the prior 6 months at study entry, L is an indicator of heavy alcohol use in the prior 6 months at study entry, and $W(u_{enter} \leq u \leq u_{exit})$ is the weight between u_{enter} and u_{exit} . In the right diagram, open right arrows represent censoring due to loss to follow up, closed right arrows represent censoring due to reaching the administrative end of the study, and closed dots represent the occurrence of the event.

Figure A2.3. Re-weighted follow up data after study entry among 8 participants in a cohort study. The time scale is visits since study entry where time is indexed by u . In the top diagram, open right arrows represent censoring due to loss to follow up, closed right arrows represent censoring due to reaching the administrative end of the study, and closed dots represent the occurrence of the event. In the bottom table, i is the subject identifier, u_{enter} is the time from study entry at the start of a defined time interval, u_{exit} is the time from study entry at the end of a defined time interval, D is an indicator of loss to follow up at u_{exit} , O is an indicator of the occurrence of the event at u_{exit} , A is an indicator of injection drug use in the prior 6 months at study entry, L is an indicator of heavy alcohol use in the prior 6 months at study entry, $W(u_{enter} \leq u \leq u_{exit})$ is the unstabilized weight between u_{enter} and u_{exit} , and $SW(u_{enter} \leq u \leq u_{exit})$ is the stabilized weight between u_{enter} and u_{exit} based on equation (2).

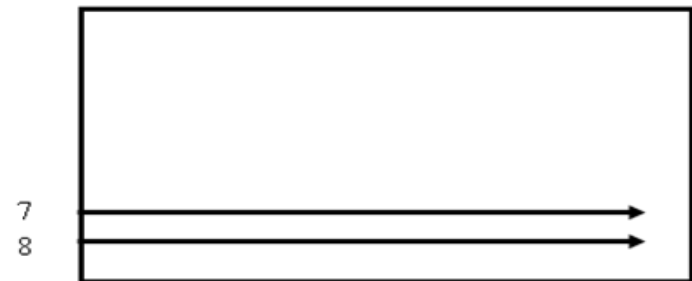
i	Y	D	O	A	L
1	3	0	1	1	1
2	4	1	0	0	0
3	5	1	0	1	1
4	9	0	1	1	1
5	11	0	1	0	0
6	15	0	0	0	0
7	15	0	0	1	0
8	15	0	0	1	0



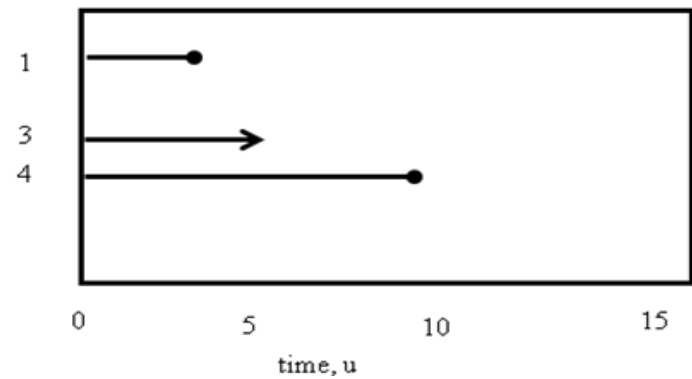
i	$u_{\text{enter}}, u_{\text{exit}}$	D	O	A	L	$W(u_{\text{enter}} \leq u \leq u_{\text{exit}})$
2	1,3	0	0	0	0	$1/(3/3)=1$
2	4	1	0	0	0	0
5	1,3	0	0	0	0	$1/(3/3)=1$
5	4,11	0	1	0	0	$1*1/(2/3)=1.5$
6	1,3	0	0	0	0	$1/(3/3)=1$
6	4,15	0	0	0	0	$1*1/(2/3)=1.5$

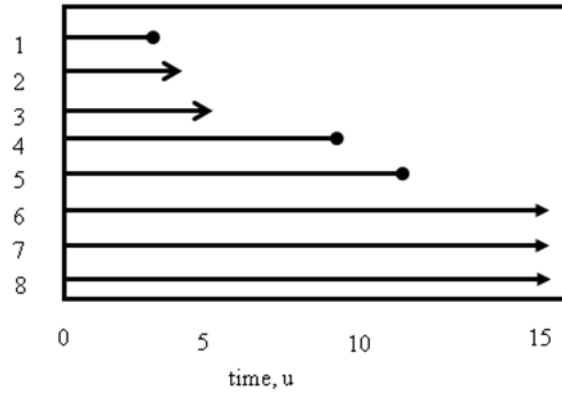


i	$u_{\text{enter}}, u_{\text{exit}}$	D	O	A	L	$W(u_{\text{enter}} \leq u \leq u_{\text{exit}})$
7	1,15	0	0	1	0	$1/(2/2)=1$
8	1,15	0	0	1	0	$1/(2/2)=1$



i	$u_{\text{enter}}, u_{\text{exit}}$	D	O	A	L	$W(u_{\text{enter}} \leq u \leq u_{\text{exit}})$
1	1,3	0	1	1	1	$1/(3/3)=1$
3	1,4	0	0	1	1	$1/(3/3)=1$
3	5	1	0	1	1	0
4	1,4	0	0	1	1	$1/(3/3)=1$
4	5,9	0	1	1	1	$1*1/(1/2)=2$





i	$u_{\text{enter}}, u_{\text{exit}}$	D	O	A	L	$W(u_{\text{enter}} \leq u \leq u_{\text{exit}})$	$SW(u_{\text{enter}} \leq u \leq u_{\text{exit}})$
1	1,3	0	1	1	1	$1/(3/3)=1$	$(5/5)/(3/3)=1$
2	1,3	0	0	0	0	$1/(3/3)=1$	$(3/3)/(3/3)=1$
2	4	1	0	0	0	0	$1*(2/3)/(2/3)=1$
3	1,4	0	0	1	1	$1/(3/3)=1$	$(5/5)/(3/3)=1$
3	5	1	0	1	1	0	$1*(3/4)/(1/2)=1.5$
4	1,4	0	0	1	1	$1/(3/3)=1$	$(5/5)/(3/3)=1$
4	5,9	0	1	1	1	$1*1/(1/2)=2$	$1*(3/4)/(1/2)=1.5$
5	1,3	0	0	0	0	$1/(3/3)=1$	$(3/3)/(3/3)=1$
5	4,11	0	1	0	0	$1*1/(2/3)=1.5$	$1*(2/3)/(2/3)=1$
6	1,3	0	0	0	0	$1/(3/3)=1$	$(3/3)/(3/3)=1$
6	4,15	0	0	0	0	$1*1/(2/3)=1.5$	$1*(2/3)/(2/3)=1$
7	1,4	0	0	1	0	$1/(2/2)=1$	$(5/5)/(2/2)=1$
7	5,15	0	0	1	0	$1*1/(2/2)=1$	$1*(3/4)/(2/2)=0.75$
8	1,4	0	0	1	0	$1/(2/2)=1$	$(5/5)/(2/2)=1$
8	5,15	0	0	1	0	$1*1/(2/2)=1$	$1*(3/4)/(2/2)=0.75$

eAPPENDIX 3

DESCRIPTION OF SAS CODE USED TO EXAMINE THE ASSOCIATION BETWEEN AFRICAN AMERICAN RACE AND SUBSEQUENT DEATH USING MODIFIED DATA FROM THE UNIVERSITY OF NORTH CAROLINA CENTER FOR AIDS RESEARCH HIV CLINICAL COHORT STUDY

Here we provide the SAS code that was used to examine the association between African American race and subsequent death as described in the main text using modified data from the University of North Carolina (UNC) Center for AIDS Research (CFAR) HIV Clinical Cohort Study¹⁰. The UNC data file, UNCDATA, contains multiple records per participant where each participant is uniquely identified by the variable, ID. Each record corresponds to a clinic visit denoted by the variable, VISIT. A description of the 2-part process that was used to estimate the stabilized weighted survival function and risk ratios for the entire study population in Figure 3 of the main text follows. In general, the stabilized weights, $SW(u)$, for the entire study population described in the main text were estimated in parts 1a through 1c. The stabilized weights estimated in part 1 were used to estimate the weighted survival function and risk ratios for the entire study population in part 2. Note in the provided SAS code, $SW(u)$ is represented as, sw.

In parts 1a and 1b, two pooled logistic models were fit to UNCDATA and used to estimate the conditional probabilities for the numerator and denominator of the weight, sw. The outcome in both models was DROPOUT which was an indicator of whether a participant was lost to follow up at a given VISIT (DROPOUT=1 for yes; DROPOUT=0 for no). The *proc logistic* statement models the log odds that DROPOUT=0. In the pooled logistic model used to estimate the conditional probabilities for the numerator in sw in part 1a, predictors solely included the time-updated time-specific intercepts (VISIT VISIT_SQ). The pooled logistic model outputs the

conditional probability that DROPOUT=0 (n_{drop}) at each record where the numerator of sw is the product of conditional probabilities that DROPOUT=0. These outputted probabilities are saved in the dataset, n_data .

In the pooled logistic model used to estimate the conditional probabilities for the denominator in sw in part 1b, predictors included the time-updated time-specific intercepts, time-fixed variables for African American race (AA), health insurance (INS1STVISIT), prior AIDS-defining illnesses (AIDS1STVISIT) as well as time-updated variables for CD4 cell count (CD4BELOW200) and HIV-1 RNA level (DETECTABLERN). The pooled logistic model outputs the conditional probability that DROPOUT=0 (d_{drop}) where the denominator of sw is the product of conditional probabilities that DROPOUT=0. These outputted probabilities are saved in the dataset, d_data . In part 1c, the data sets with the conditional probabilities are merged with UNCDATA to form the dataset, $uncdata_merged$. Next, the weight sw is obtained by taking the ratio of the products of the conditional probabilities estimated in parts 1a and 1b.

In part 2a, the *proc freq* statement is used to calculate the weighted risk set size, r_{sw} , and weighted number of deaths that occur among participants in the weighted risk set, d_{sw} , at each VISIT to be included in the denominator and numerator of equation (5) in our eAppendix 2, respectively, among African American and Caucasian participants. As shown in several *data steps* in the included part 2a SAS code, the output from equation (5), ρ_{sw} , can be used to obtain the weighted survival function for the entire study population, $surv_{sw}$, via equation (6) in our eAppendix 2. In part 2b, code similar to the code used in part 2a, was used to estimate the weighted survival functions for African Americans (aa_surv_{sw}) and Caucasians (c_surv_{sw}) and in turn the weighted risk ratios (rr_{sw}) for the entire study population.

GLOSSARY OF DATA AND VARIABLES USED IN PROVIDED SAS CODE

UNCDATA is the UNC CFAR HIV clinical cohort data file. The variables included in UNCDATA appear in uppercase and are defined as:

- ID – Unique participant identifier
- VISIT – Clinic visit
- VISIT_SQ – The square of VISIT (i.e., VISIT*VISIT)
- MAXVISIT – Maximum number of clinic visits for a given participant
- GENDER – Categorical variable for participant's gender (i.e., male or female)
- AA – Indicator of whether a participant is African American (AA=1 for yes; AA=0 for no)
- CD4BELOW200 – Indicator of whether a participant had a CD4 cell count below 200 cells/microL at prior VISIT (CD4BELOW200=1 for yes; CD4BELOW200=0 for no)
- DETECTABLERNAL – Indicator of whether a participant had a detectable HIV-1 RNA level at prior VISIT (DETECTABLERNAL =1 for yes; DETECTABLERNAL=0 for no)
- INS1STVISIT – Categorical variable for participant's health insurance type at the first clinic visit (i.e., private, public, or uninsured)
- AIDS1STVISIT – Indicator of whether a participant had a prior diagnosis of an AIDS-defining illness at the first clinic visit (AIDS1STVISIT =1 for yes; AIDS1STVISIT =0 for no)
- ART – Indicator of whether a participant had previously used antiretroviral therapy at prior VISIT (ART=1 for yes; ART=0 for no)
- AGE – Participant's age at prior VISIT in years
- DROPOUT – Indicator of whether a participant was lost to follow up at a given VISIT (DROPOUT=1 for yes; DROPOUT=0 for no)

- DIED – Indicator of whether a participant died at a given VISIT (DIED=1 for yes; DIED=0 for no)

The data and variables generated from UNCDATA appear in lowercase and are defined as:

- sw – Censoring weight for loss to follow up for entire study population
- n_drop – Conditional probability of a participant not dropping out at a given VISIT in numerator of sw
- n_data – Dataset that contains n_drop
- d_drop – Conditional probability of a participant not dropping out at a given VISIT in denominator of sw
- d_data – Dataset that contains d_drop
- uncd_data_merged – Dataset resulting from merging UNCDATA, n_data, and d_data
- r_sw – Weighted risk set size at a given VISIT for entire study population
- d_sw – Weighted number of deaths at a given VISIT for entire study population
- rho_sw – Weighted proportion at a given VISIT for entire study population
- surv_sw – Weighted survival estimate at a given VISIT for entire study population
- aa_r_sw – Weighted risk set size at a given VISIT for African Americans
- aa_d_sw – Weighted number of deaths at a given VISIT for African Americans
- aa_rho_sw – Weighted proportion at a given VISIT for African Americans
- aa_surv_sw – Weighted survival estimate at a given VISIT for African Americans
- c_r_sw – Weighted risk set size at a given VISIT for Caucasians
- c_d_sw – Weighted number of deaths at a given VISIT for Caucasians

- c_rho_sw – Weighted proportion at a given VISIT for Caucasians
- c_surv_sw – Weighted survival estimate at a given VISIT for Caucasians
- rr_sw – Weighted risk ratio estimate at a given VISIT for entire study population

SAS (VERSION 9.3) CODE

/**Using Equation (7) in our eAppendix 2 to estimate stabilized weights for loss to follow up, sw, and in turn the stabilized weighted survival function and risk ratios for the entire study population in Figure 3 of the main text;*/

/**Part 1a: Estimating conditional probabilities for numerator using pooled logistic regression;*/

*Modeling the log odds that DROPOUT=0 and outputting corresponding probabilities as n_drop into n_data dataset;

```
proc logistic data=UNCDATA;  
    model DROPOUT=VISIT VISIT_SQ;  
    output out=n_data (keep=ID VISIT n_drop) p=n_drop;  
run;
```

/**Part 1b: Estimating conditional probabilities for denominator using pooled logistic regression;*/

*Modeling the log odds that DROPOUT=0 and outputting corresponding probabilities as d_drop into d_data dataset;

```
proc logistic data=UNCDATA;  
    class INS1STVISIT;  
    model DROPOUT=VISIT VISIT_SQ AA CD4BELOW200 INS1STVISIT AIDS1STVISIT DETECTABLERN;  
    output out=d_data (keep=ID VISIT d_drop) p=d_drop;  
run;
```

/**Part 1c: Calculating cumulative probabilities for stabilized weights (sw);*/

*Sorting records in UNCDATA and all generated datasets (n_data and d_data) by ID and VISIT;

```
proc sort data=UNCDATA; by ID VISIT; run;  
proc sort data=n_data; by ID VISIT; run;  
proc sort data=d_data; by ID VISIT; run;
```

*Merging UNCDATA with generated datasets (n_data and d_data) by ID and VISIT and calculating sw;

```
data uncd_data_merged;  
    merge UNCDATA n_data d_data;  
    by ID VISIT;  
    retain num_drop den_drop;  
    if first.ID then do; num_drop=1; den_drop=1; end;
```

```

    num_drop=num_drop*n_drop;
    den_drop=den_drop*d_drop;
    sw=num_drop/den_drop;
run;

*Assessing distribution of sw;
proc means data= uncddata_merged n min mean max std p1 p25 p50 p75 p99;
    var sw;
run;

*Calculating sample size, total person-visits, and number of deaths in observed population;
proc means data= uncddata_merged n nmiss sum;
    where VISIT=MAXVISIT;
    var VISIT;
run;
proc freq data=uncdata_merged;
    tables DIED;
run;

*Examining distribution of characteristics at the first clinic visit in observed population;
proc freq data= uncddata_merged;
    where VISIT=1;
    tables  AA GENDER CD4BELOW200 AIDS1STVISIT ART INS1STVISIT DETECTABLERNA;
run;
proc means data= uncddata_merged n nmiss p25 p50 p75;
    where VISIT=1;
    var AGE;
run;

*Calculating total person-visits and number of deaths in weighted population;
proc means data= uncddata_merged n nmiss sum;
    weight sw;

```

```

        where VISIT=MAXVISIT;
        var VISIT;
run;
proc freq data=uncdata_merged;
    weight sw;
    tables DIED;
run;

*Calculating sample size and examining the distribution of characteristics at the first clinic visit in weighted population;
proc freq data=uncdata_merged;
    weight sw;
    where VISIT=1;
    tables  AA GENDER CD4BELOW200 AIDS1STVISIT ART INS1STVISIT DETECTABLERNA;
run;
proc means data=uncdata_merged n nmiss p25 p50 p75;
    weight sw;
    where VISIT=1;
    var AGE;
run;

/****Part 2a: Estimating weighted survival function for entire study population using equations (5) and (6) in our eAppendix 2;****/
*Calculating weighted risk set size, r_sw, and weighted number of deaths, d_sw, at each VISIT;
proc freq data=uncdata_merged noprint; weight sw; tables VISIT*DIED / out=surv_sw; run;
proc sort data=surv_sw; by VISIT DIED; run;
data surv_sw;
    set surv_sw;
    by VISIT DIED;
    retain r_sw;
    if first.VISIT then do; d_sw=0; r_sw=0; end;
    if DIED=0 then r_sw=count; else if DIED=1 then do; r_sw=r_sw+count; d_sw=count; end;
    if last.VISIT then output;
    keep VISIT r_sw d_sw;

```

```
run;
data surv_sw;
    set surv_sw;
    z=1;
run;
```

*Calculating weighted proportion, rho_sw, and survival function, surv_sw, at each VISIT;

```
proc sort data=surv_sw; by z VISIT; run;
data surv_sw;
    set surv_sw;
    by z;
    retain surv_sw;
    if first.z then do surv_sw=1; end;
    rho_sw=d_sw/r_sw;
    rhoc_sw=1-rho_sw;
    surv_sw=surv_sw*rhoc_sw;
    drop rhoc_sw;
run;
```

*Creating dataset with extra record for VISIT=0 so that survival function starts at 1;

```
data extra;
    input VISIT;
    datalines;
    0
run;
```

*Merging in extra record into weighted survival function dataset;

```
proc sort data=surv_sw; by VISIT; run;
proc sort data=extra; by VISIT; run;
data surv_sw_merged;
    merge surv_sw extra;
    by VISIT;
```

```

        if VISIT=0 then surv_sw=1;
        drop z;
run;

/**Part 2b: Estimating weighted risk ratios for entire study population based on weighted survival functions estimated using
equations (5) and (6) in our eAppendix 2 among African Americans (aa_surv_sw) as well as among Caucasians (c_surv_sw);***/
*Calculating weighted risk set size, aa_r_sw, and weighted number of deaths, aa_d_sw, among African Americans at each VISIT;
proc freq data= unccdata_merged noprint; where AA=1; weight sw; tables VISIT*DIED / out=aa_surv_sw; run;
proc sort data=aa_surv_sw; by VISIT DIED; run;
data aa_surv_sw;
    set aa_surv_sw;
    by VISIT DIED;
    retain aa_r_sw;
    if first.VISIT then do; aa_d_sw=0; aa_r_sw=0; end;
    if DIED=0 then aa_r_sw=count; else if DIED=1 then do; aa_r_sw=aa_r_sw+count; aa_d_sw=count; end;
    if last.VISIT then output;
    keep VISIT aa_r_sw aa_d_sw;
run;
data aa_surv_sw;
    set aa_surv_sw;
    z=1;
run;

*Calculating weighted proportion, aa_rho_sw, and survival function, aa_surv_sw, among African Americans at each VISIT;
proc sort data= aa_surv_sw; by z VISIT; run;
data aa_surv_sw;
    set aa_surv_sw;
    by z;
    retain aa_surv_sw;
    if first.z then do aa_surv_sw=1; end;
    aa_rho_sw=aa_d_sw/aa_r_sw;
    aa_rhoc_sw=1-aa_rho_sw;

```



```

aa_surv_sw=aa_surv_sw*aa_rhoc_sw;
drop aa_rhoc_sw;
run;

*Calculating weighted risk set size, c_r_sw, and weighted number of deaths, c_d_sw, among Caucasians at each VISIT;
proc freq data= unccdata_merged noprint; where AA=0; weight sw; tables VISIT*DIED / out=c_surv_sw; run;
proc sort data=c_surv_sw; by VISIT DIED; run;
data c_surv_sw;
    set c_surv_sw;
    by VISIT DIED;
    retain c_r_sw;
    if first.VISIT then do; c_d_sw=0; c_r_sw=0; end;
    if DIED=0 then c_r_sw=count; else if DIED=1 then do; c_r_sw=c_r_sw+count; c_d_sw=count; end;
    if last.VISIT then output;
    keep VISIT c_r_sw c_d_sw;
run;
data c_surv_sw;
    set c_surv_sw;
    z=1;
run;

*Calculating weighted proportion, c_rho_sw, and survival function, c_surv_sw, among Caucasians at each VISIT;
proc sort data= c_surv_sw; by z VISIT; run;
data c_surv_sw;
    set c_surv_sw;
    by z;
    retain c_surv_sw;
    if first.z then do c_surv_sw=1; end;
    c_rho_sw=c_d_sw/c_r_sw;
    c_rhoc_sw=1-c_rho_sw;
    c_surv_sw=c_surv_sw*c_rhoc_sw;
    drop c_rhoc_sw;

```

```
run;
```

```
*Merging datasets with weighted survival functions among African Americans and Caucasians;
```

```
proc sort data= aa_surv_sw; by z VISIT; run;
```

```
proc sort data= c_surv_sw; by z VISIT; run;
```

```
data aa_c_surv_sw;
```

```
    merge aa_surv_sw c_surv_sw;
```

```
    by z VISIT;
```

```
run;
```

```
*Calculating weighted risk ratios (rr_sw) comparing risk among African Americans to risk among Caucasians;
```

```
data aa_c_surv_sw;
```

```
    set aa_c_surv_sw;
```

```
    if c_surv_sw <1 then rr_sw=(1-aa_surv_sw)/(1- c_surv_sw); else rr_sw=.;
```

```
run;
```

REFERENCES

1. Cole SR, Hudgens MG, Brookhart MA, Westreich D. Risk. *Am J Epidemiol*. 2015;**181**(4):246-50. doi: 10.1093/aje/kwv001. Epub 2015 Feb 5.
2. Singer JD, Willett JB. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, New York: Oxford University Press, Inc, 2003.
3. Hernán MA, McAdams M, McGrath N, Lanoy E, Costagliola D. Observation plans in longitudinal studies with time-varying treatments. *Stat Methods Med Res* 2009;**18**(1):27-52.
4. Hernán MA, Robins J. *Causal Inference Book*. Boca Raton: Chapman & Hall/CRC, Forthcoming 2016.
5. Allison PD. *Survival Analysis Using SAS: A Practical Guide, Second Edition*. Second ed SAS Institute 2010.
6. Robins JM, Finkelstein DM. Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests. *Biometrics* 2000;**56**(3):779-788.
7. Wasserman L. The Bootstrap. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. New York, New York: Springer Science+Business Media, Inc, 2003.
8. Cole SR, Lau B, Eron JJ, et al. Estimation of the standardized risk difference and ratio in a competing risks framework: application to injection drug use and progression to AIDS after initiation of antiretroviral therapy. *Am J Epidemiol*. 2015;**181**(4):238-45. doi: 10.1093/aje/kwu122. Epub 2014 Jun 24.

9. Howe CJ, Cole SR, Mehta SH, Kirk GD. Estimating the effects of multiple time-varying exposures using joint marginal structural models: alcohol consumption, injection drug use, and HIV acquisition. *Epidemiology*. 2012;**23**(4):574-82. doi: 10.1097/EDE.0b013e31824d1ccb.
10. Howe CJ, Cole SR, Napravnik S, Eron JJ. Enrollment, retention, and visit attendance in the University of North Carolina Center for AIDS Research HIV clinical cohort, 2001-2007. *AIDS Res Hum Retroviruses*. 2010;**26**(8):875-81. doi: 10.1089/aid.2009.0282.