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 function. 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., $S(t_{\text{continuous}}) = \exp \left\{ -\left( t_{\text{continuous}} / \mu_1 \right) \right\}$ and $S(c_{\text{continuous}}) = \exp \left\{ -\left( c_{\text{continuous}} / \mu_2 \right) \right\}$ where $\mu_1$ and $\mu_2$ (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 methods 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 $\mu_1$ 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 $\mu_2$ 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 $\mu_1$ 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 $\mu_2$ by a factor of 0.27 and 0.41, respectively.

For Diagram V) both the prevalence of having a CD4 cell count $\geq 200$ 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 $\geq 200$ cells/microL where 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 \( \geq 200 \) 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 \( \mu_t \) 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 \( \mu_2 \) 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.,

$$S(t_{\text{continuous}}) = \exp\{-[t_{\text{continuous}} / \theta]^{\sigma}\}$$

where \( \theta \) and \( \sigma \) for the baseline survival function for death was 12.18 and 0.8, respectively. The \( \sigma \) 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.,

$$S(c_{\text{continuous}}) = \exp\{-[c_{\text{continuous}} / \phi]\}$$

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

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

(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 $u = 1, 2, ..., 15$. 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

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

(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 times \(^5\).

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).

$\logit P[D(k) = 0 \mid \overline{D}(k-1) = 0, \overline{O}(k-1) = 0, A] = \alpha_{0k} + \alpha_i A$  \hspace{1cm} (3)$

$\logit P[D(k) = 0 \mid \overline{D}(k-1) = 0, \overline{O}(k-1) = 0, A, L] = \beta_{0k} + \beta_i A + \beta_2 L$  \hspace{1cm} (4)$

In equations (3) and (4) $\logit p = \ln[p / (1 - p)]$. The parameters $\alpha_{0k}$ and $\beta_{0k}$ are the time specific intercepts without and with inclusion of $L$ in the pooled model, respectively, where $k$ is time. The parameter $\alpha_i$ is the log hazard odds ratio for the effect of $A$ on $D$, $\beta_i$ is the log hazard odds ratio for the effect of $A$ on $D$ adjusting for $L$, and $\beta_2$ 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 Finkelstein $^6$ 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.

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

(5)

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

(6)

In equations (5) and (6), $\tau_k$ is the subset of the cohort at entry that is in the risk set at time $k$, while $V_k$ is the subset of $\tau_k$ that develops the event at time $k$. Note $\hat{S}(u)$ 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 resample. 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 $\left[1 - \hat{S}_{A=1}(u)\right] - \left[1 - \hat{S}_{A=0}(u)\right]$ and $\left[1 - \hat{S}_{A=1}(u)\right] / \left[1 - \hat{S}_{A=0}(u)\right]$, 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.

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)

\[
SW(u) = \prod_{k=1}^{n} \frac{P[D(k) = 0 | D(k-1) = 0, O(k-1) = 0]}{P[D(k) = 0 | D(k-1) = 0, O(k-1) = 0, A, L]}
\]

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, $u_{\text{enter}}$ is the time from study entry at the start of a defined time interval, $u_{\text{exit}}$ is the time from study entry at the end of a defined time interval, $D$ is an indicator of loss to follow up at $u_{\text{exit}}$, $O$ is an indicator of the occurrence of the event at $u_{\text{exit}}$, $A$ 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 $W(u_{\text{enter}} \leq u \leq u_{\text{exit}})$ is the weight between $u_{\text{enter}}$ and $u_{\text{exit}}$. 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, \( u_{\text{enter}} \) is the time from study entry at the start of a defined time interval, \( u_{\text{exit}} \) is the time from study entry at the end of a defined time interval, \( D \) is an indicator of loss to follow up at \( u_{\text{exit}} \), \( O \) is an indicator of the occurrence of the event at \( u_{\text{exit}} \), \( A \) 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, 
\[ W(u_{\text{enter}} \leq u \leq u_{\text{exit}}) \] is the unstabilized weight between \( u_{\text{enter}} \) and \( u_{\text{exit}} \), and 
\[ SW(u_{\text{enter}} \leq u \leq u_{\text{exit}}) \] is the stabilized weight between \( u_{\text{enter}} \) and \( u_{\text{exit}} \) based on equation (2).
<table>
<thead>
<tr>
<th>i</th>
<th>Y</th>
<th>D</th>
<th>O</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( i )</td>
<td>( u_{\text{enter}} )</td>
<td>( u_{\text{exit}} )</td>
<td>( D )</td>
<td>( O )</td>
<td>( A )</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2</td>
<td>1,3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1,3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>4,11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1,3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>4,15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1,15</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1,15</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1,3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1,4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1,4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5,9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$i$</td>
<td>$u_{enter} - u_{exit}$</td>
<td>$D$</td>
<td>$O$</td>
<td>$A$</td>
<td>$L$</td>
</tr>
<tr>
<td>-----</td>
<td>----------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>1</td>
<td>1,3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1,3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1,4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1,4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5,9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1,3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>4,11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1,3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>4,15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1,4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>5,15</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1,4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>5,15</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
APPENDIX 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 Study. 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 \texttt{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 (DETECTABLERNA). 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, unndata_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), $rho\_sw$, 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)
- **DETECTABLERNA** – Indicator of whether a participant had a detectable HIV-1 RNA level at prior VISIT (DETECTABLERNA=1 for yes; DETECTABLERNA=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
• unndata_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 DETECTABLERNA;
    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 unndata_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=uncdata_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=uncdata_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=uncdata_merged;
    where VISIT=1;
    tables AA GENDER CD4BELOW200 AIDS1STVISIT ART INS1STVISIT DETECTABLERNA;
run;
proc means data=uncdata_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=uncdata_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= uncdatano drop=VARIABLE; 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;
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= uncedata_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= uncdata_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


