WEB MATERIAL

Web Appendix 1. G-formula details

Overview

We estimate the cumulative mortality risk under the observed treatment history (i.e., "no intervention") and each of the dynamic treatment plans using a Monte Carlo simulation of 50,000 patients sampled from the study sample randomly with replacement at baseline. The densities of covariates measured at baseline were estimated using their empirical distributions in the sampled data, and the densities of time-varying covariates were modeled using parametric regression models in the observed data. The values of these covariates under each dynamic treatment plan were imputed using a draw from a density estimated by the regression models after assigning treatment according to the designated treatment plan in the Monte Carlo sample. The probability of cause-specific mortality was estimated for each patient in each person-month using a parametric regression model with treatment assignment and covariate history set to what it would have been under the given dynamic treatment plan.

Details

Let patients be indexed from $i = 1, \ldots 3882$, A represent treatment, $\boldsymbol{Z}$ represent a vector of covariates including CD4 cell count, viral load, and baseline covariates, C represent censoring, and Y represent death.

1. **Write the formula for the cumulative incidence under no intervention on exposure.** Under the assumption that censoring was not informative on later outcome, exposure, or time varying covariates, the cumulative incidence of death due to cause j at time t can be written as Equation 1:

$$F(t,j) = \sum_{\bar{a}_t} \sum_{\bar{z}_t} \sum_{k=0}^{t} P[J = j | Y(k) = 1, \bar{A}(k) = \bar{a}(k), \bar{Z}(k) = \bar{z}(k), Y(k-1) = C(k) = 0]$$

$$\times P[Y(k) = 1 | \bar{A}(k) = \bar{a}(k), \bar{Z}(k) = \bar{z}(k), Y(k-1) = C(k) = 0]$$

$$\times \prod_{s=0}^{k} \begin{array}{c} \times f[a(s) | \bar{z}(s), \bar{a}(s-1), Y(s-1) = C(s) = 0] \\ \times f[z(s) |\, z(s-1), \bar{a}(s-1), Y(s-1) = C(s) = 0] \\ P[Y(s-1) = 0 | \bar{A}(s-1) = \bar{a}(s-1), \bar{Z}(s-1) = \bar{z}(s-1), Y(s-2) = C(s-1) = 0] \end{array}$$

2. **Fit parametric models.** We first fit parametric models (described below) to each component of this joint density of the observed data as specified below. The notation $g(x)$ indicates that x was modeled flexibly using restricted quadratic splines.
    a. Fit a logistic model to estimate whether patient *i* has a detectable viral load at time *s.*

$$logit[P(dvl_{i,s} = 1)] = \\ = \alpha_0 + \boldsymbol{\alpha_1 L}_i^T + \alpha_2 dvl_{i,s-1} + \alpha_3 g(vl_{i,s-1}) + \alpha_4 g(cd4_{i,s-1}) \\ + \alpha_5 A + \alpha_7 g(s)$$

Where ***L*** is a vector of time-fixed covariates, including sex, race, ethnicity, injection drug use, MSM status, age at baseline, year of study entry, CD4 at study entry, viral load at study entry, the product of year and age at study entry.

b. Fit linear regression models to estimate viral load at time s among patients who had a detectable viral load at time s.

$$VL_{i,s} = \alpha_0 + \boldsymbol{\alpha_1 L_i^T} + \alpha_2 dvl_{i,s-1} + \alpha_3 g(vl_{i,s-1}) + \alpha_4 g(cd4_{i,s-1}) + \alpha_5 A + \alpha_7 AIDS_{i,s-1} + \alpha_8 g(s) + \epsilon_i, \quad \epsilon \sim N(0,\sigma)$$

c. Fit linear regression models to estimate CD4 cell count at time s stratified by treatment at time $s-1$:

$$CD4_{i,s} = \alpha_0 + \boldsymbol{\alpha_1} L_i^T + \alpha_2 dvl_{i,s-1} + \alpha_3 dvl_{i,s} + \alpha_4 g(vl_{i,s}) + \alpha_5 g(vl_{i,s-1}) + \alpha_6 g(cd4_{i,s-1}) + \alpha_7 AIDS_{i,s-1} + \alpha_8 g(s) + \epsilon_i, \quad \epsilon \sim N(0,\sigma)$$

d. Fit a logistic regression model to estimate the probability of being treated at time *s.*

$$logit\big[P[A(s)=1|\bar{Z}(s)=\bar{z}(s),\bar{a}(s-1),Y(s-1)=C(s)=0]\big] = \alpha_0 + \boldsymbol{\alpha_1} L_i^T + \alpha_2 dvl_{i,s} + \alpha_3 dvl_{i,s-1} + \alpha_4 g(vl_{i,s}) + \alpha_5 g(cd4_{i,s}) + \alpha_6 AIDS_{i,s} + \alpha_7 g(s)$$

e. Fit a logistic regression model to estimate the probability of death at time *s.*

$$logit\big[P[Y(k)=1|\bar{A}(k)=\bar{a}(k),\bar{Z}(k)=\bar{z}(k),Y(k-1)=C(k)=0]\big] = \alpha_0 + \boldsymbol{\alpha_1} L_i^T + \alpha_2 dvl_{i,k} + \alpha_3 dvl_{i,k-1} + \alpha_4 g(vl_{i,k}) + \alpha_5 g(cd4_{i,s}) + \alpha_6 A + \alpha_7 AIDS_{i,s-1} + \alpha_8 g(k)$$

f. Fit a logistic regression model to estimate the probability of a death at time *k* is due to cause j.

$$logit[P[J=j|Y(k)=1,\bar{A}(k)=\bar{a}(k),\bar{Z}(k)=\bar{z}(k),Y(k-1)=C(k)=0]] = \alpha_0 + \boldsymbol{\alpha_1} L_i^T + \alpha_2 dvl_{i,k} + \alpha_3 dvl_{i,k-1} + \alpha_4 g(vl_{i,k}) + \alpha_5 g(cd4_{i,k}) + \alpha_6 A + \alpha_7 AIDS_{i,s-1} + \alpha_8 g(k)$$

In all models, continuous variables were modeled using restricted quadratic splines (36), and categorical variables were modeled using indicator variables. Addition of interaction terms between treatment status and year did not alter the results.

3. **Draw Monte Carlo Sample:** Draw a large (N = 50,000) Monte Carlo sample from the observed patients at baseline with replacement.
4. **Check form of parametric models:** In the Monte Carlo sample, estimate the cumulative incidence under no intervention on treatment or censoring using the conditional probabilities defined by parameters estimated from the parametric models above and Equation 1. See the figure below, comparing the observed cumulative incidence function to the cumulative incidence function predicted using Equation 1.

5. **Write the formula for the cumulative incidence under each exposure plan.** Write the cumulative incidence of death due to cause j at time t under dynamic treatment regime g as Equation 2:

$$F_g(t,j) = \sum_{\bar{a}_t} \sum_{\bar{z}_t} \sum_{k=0}^{t} P[J = j | Y(k) = 1, \bar{A}(k) = \bar{a}(k), \bar{Z}(k) = \bar{z}(k), Y(k-1) = C(k) = 0]$$

$$\times P[Y(k) = 1 | \bar{A}(k) = \bar{a}(k), \bar{Z}(k) = \bar{z}(k), Y(k-1) = C(k) = 0]$$

$$\times \prod_{s=0}^{k} \begin{matrix} f^g[a(s)|\bar{z}(s), \bar{a}(s-1), Y(s-1) = C(s) = 0] \\ \times f[z(s)|\, z(s-1), \bar{a}(s-1), Y(s-1) = C(s) = 0] \\ P[Y(s-1) = 0 | \bar{A}(s-1) = \bar{a}(s-1), \bar{Z}(s-1) = \bar{z}(s-1), Y(s-2) = C(s-1) = 0] \end{matrix}$$

6. **Estimate cumulative incidence under each exposure plan:** In the Monte Carlo sample, estimate the cumulative incidence each dynamic treatment plan using the g-formula provided in equation 2.
    a. The distribution of ***L*** in the large Monte Carlo sample approximates the distribution of ***L*** in the observed data.
    b. Estimate CD4 cell count, whether or not viral load is detectable, and viral load for participant i at time s using the coefficients from models a through c above.
    c. Set treatment according to the treatment regime of interest. The value of $A_{i,s}$ drawn from a Bernoulli distribution with probability given below. Specifically, if *x* represents the CD4 cell count threshold for treatment initiation, and *m* is the number of months that CD4 cell count has been below *x*,

For the immediate treatment arm,
$f^g[a(s)|\bar{z}(s), \bar{a}(s-1), Y(s-1) = C(s) = 0] = 1$

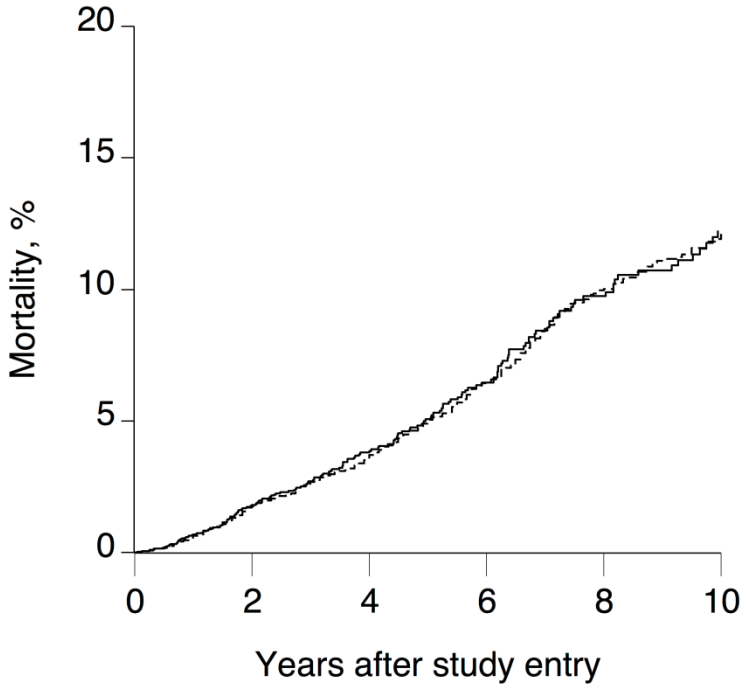
For the delayed treatment arm,

$$f^g[a(s)|\bar{z}(s), \bar{a}(s-1), Y(s-1) = C(s) = 0]$$
$$= \begin{cases} 0 & CD4_{i,s} > x \\ f^{obs}[a(s)|\bar{z}(s), \bar{a}(s-1), Y(s-1) = C(s) = 0] & CD4_{i,s} < x, m < 6 \\ 1 & (CD4_{i,s} < x\,, m \geq 6) \text{ or } A_{i,s-1} = 1 \end{cases}$$

    d. Prevent censoring; set $P(C = 0) = 1\ \forall\ i, s$

7. Perform steps 1 through 4 in 500 bootstrap samples. We use the standard deviation of the 500 estimates as the standard error of the point estimate.

Appendix 1 Figure: Observed cumulative incidence of mortality (solid line) and cumulative incidence of mortality estimated under no intervention (dashed line) among 3882 patients who entered care with a CD4 cell count over 500 cells/mm^3 between January 1, 1998 and December 31, 2014 at 8 US clinical sites, followed for death up to 10 years



Web Appendix 2. Accounting for missing cause of death data

Under the assumption that cause of death data is missing at random, conditional on the covariates included in the cause of death model described above, the parametric g-formula will provide consistent estimates of the cumulative incidence of death due to cause j even in the presence of missing cause of death data. To account for the missing data, one simply fits the cause of death model among those who have cause of death information in Step 1. This regression model should contain the union of the sets of covariates needed for exchangeability between participants following each exposure plan at time t and for exchangeability between participants with and without cause of death information. Then, in Step 4, the regression coefficients estimated using the procedure outlined above are used to estimate each individual's probability of dying due to cause j by time t, even among individuals missing cause of death in the study. This approach provides a consistent estimator because we appropriately assume that the conditional probabilities of dying due to cause j (conditional on all of the covariates in the model) are the same for individuals with and without cause of death information. This procedure can be used to reconstruct the cumulative incidence of death due to cause j under no intervention on exposure plan or under the exposure plans of interest.

Simulations indicated that this approach to account for missing data yielded very little bias in estimates of the cumulative incidence of death due to cause j under the natural course or under the exposure plans of interest when the model for cause of death fit in Step 1 included all joint predictors of cause of death and missingness. Estimates became more imprecise as the amount of missing data increased. A second set of simulations shows that bias will occur when missingness of cause of death information is affected by a predictor of cause of death that is not included in the cause of death model in Step 1.

To understand why this procedure provides consistent estimates of the risk function for cause of death j, it is useful to view the g-formula as an algorithm that imputes the potential outcomes. The

models fit in Step 1 act as the imputation models that are used to predict the missing potential outcomes for the same group of subjects under alternative exposure plans. For more details on the parallels between multiple imputation and the g-formula for causal inference, see (13).

Simulations for missing data

Simulations to examine the performance of our method for handling missing data were built off the same data generating mechanism as simulations to assess the performance of the proposed approach for handling outcome misclassification. Additionally, an indicator of observing the cause of death, R_i, was generated based on covariates. In the first set of simulations (Web Table 1), R_i was determined by X_i and Z_i. In the second set of simulations (Web Table 2), R_i was determined by X_i, Z_i, and hypothetically unobserved variable U_i. Specifically, individuals with $U = 1$ were 3 times as likely to have their cause of death observed than participants with $U = 0$. In the second set of simulations, U_i was also a predictor of dying due to cause j (individuals with $U_i = 1$ had 20% higher probability of dying due to cause j as individuals with $U_i = 0$). If $R_i = 1$ then the observed cause of death $Jo_i = J_i$. If $R_i = 0$ then Jo_i was missing.

Web Table 1. Simulation results illustrating the performance of the parametric g-formula to estimate the effect of intervention X in a cohort of 3000 patients with varying amounts of missing cause of death data, where missingness depends on measured covariates included in the cause of death model in 1000 simulation experiments

	Proportion of deaths missing cause				
	0	5	25	50	75
Cumulative incidence of death due to cause j (natural course)	26.2	26.2	26.2	26.2	26.2
Cumulative incidence (set $X = 1$)	32.6	32.6	32.6	32.6	32.6
Cumulative incidence (set $X = 0$)	23.8	23.8	23.8	23.8	23.7
Difference	8.8	8.8	8.8	8.8	8.9
Bias	0.00	0.00	0.00	0.00	0.02
Standard deviation of bias	0	0.16	0.45	0.63	0.98
Mean Squared error	0	0.03	0.20	0.39	0.95

Web Table 2. Simulation results illustrating the performance of the parametric g-formula to estimate the effect of intervention X in a cohort of 3000 patients with varying amounts of missing cause of death data, where missingness depends on an unmeasured predictor of the outcome in addition to measured covariates included in the cause of death model in 1000 simulation experiments

	Proportion of deaths missing cause				
	0	5	25	50	75
Cumulative incidence of death due to cause j (natural course)	17.5	17.5	18.0	18.5	19.3
Cumulative incidence (set $X = 1$)	21.5	21.6	21.5	21.8	22.4
Cumulative incidence (set $X = 0$)	14.5	14.8	15.4	15.9	16.8
Difference	6.9	6.8	6.1	5.8	5.6
Bias	0	-0.10	-0.83	-1.04	-1.28
Standard deviation of bias	0	0.48	0.73	0.82	0.97
Mean Squared error	0	0.28	1.23	1.77	2.58

Web Appendix 3. Penalized maximum likelihood methods to handle sparse data in cause of death model

Parametric models for each component of the joint likelihood are often high dimensional. For the parametric g-formula to provide consistent estimates of the counterfactual risk functions, we must assume exchangeability between participants receiving each treatment plan at time t conditional on a set of measured variables. Exposure, covariates, and outcomes must be predicted under each treatment plan conditional on the variables required for conditional exchangeability, meaning that parametric models for each component of the joint likelihood must include these variables as covariates. In addition, all parametric models used to model exposure, covariates, and outcomes must be correctly specified. To reduce the probability of misspecifying one of these models, we let these models be as flexible as possible by, for example, modeling continuous variables using restricted quadratic splines. In some cases, the number of covariates may be large relative to the number of end points in a particular model.

In the model for cause of death, the number of parameters necessary to flexibly model the required covariates was large $(p = 17)$ relative to the number of AIDS-related deaths $(n = 36)$. As a result, estimates of parameters from this model were highly variable and unstable. To reduce the variability of these estimates, we applied a penalty to each of the parameters estimated using this model. This penalty pulled each of the p parameter estimates $\hat{\beta}$ towards values $\boldsymbol{m} = (m_1, \ldots m_p)$, effectively introducing a small about of bias in exchange for a reduction in the variance of the parameter estimates (37,38). We used a quadratic penalty such that the form of the penalized log likelihood was $\ln\{L(\boldsymbol{\beta}; \boldsymbol{j})\} - r/2 \times \sum_{k=1}^{p}(\beta_k - m_k)^2$, where r is a tuning parameter governing the amount of shrinkage towards $\boldsymbol{m}$. In the example, we set $\boldsymbol{m} = 0$ and $r = 1/10$. SAS code to apply such a penalty to a logistic regression model is provided by Cole et al (38).

Web Appendix 4. Simulations to evaluate the performance of the modified g-formula to account for outcome misclassification

We illustrate the proposed approaches to handle misclassified cause of death data in the time-fixed setting. In each scenario, let i index hypothetical patients from 1 to 3,000 in each of 2,000 simulated cohorts. Measured confounder $Z = 1$ for 30% of patients, and patients with $Z = 1$ had 7 times the odds of being exposed to the time-fixed exposure $(X = 1)$ as patients with $Z = 0$. An indicator of death due to any cause (Y_i) was generated for each patient based on X_i and Z_i, and the true cause of death J_i for simulated subjects with $Y_i = 1$ was generated based on $X_i, \text{and } Z_i$.

A misclassified version of the cause of death, J_i', was generated for each patient based on J_i and misclassification probabilities reflecting the misclassification probabilities for cause of death in the application described in the main paper. When $J_i = 1, J_i'$ was drawn from a Bernoulli distribution with probability equal to the sensitivity. When $J_i = 0, J_i'$ was drawn from a Bernoulli distribution with probability $1 - \text{specificity}$. The parametric g-formula was used to estimate the difference in the risk of death due to $J = 1$ under exposure plans $x = 1$ and $x = 0$. We assessed the performance of the proposed approach under a range of values for sensitivity and specificity, assuming that sensitivity and specificity were known. Specifically, we compared bias in the risk difference, standard deviation of the bias, and mean squared error between the standard g-formula approach and the modified g-formula approach under a range of scenarios. Bias was defined as 100 times the average of the estimated risk difference from simulation k minus the true risk difference, or $100 \times E[\widehat{RD}_k - RD]$. Mean squared error was defined as the squared bias plus the squared standard deviation of the bias.

Web Figure 1. Full sensitivity analysis graphical results