eAppendix:

DESCRIPTION OF SAS CODE USED TO FIT JOINT MARGINAL STRUCTURAL COX PROPORATIONAL HAZARDS MODEL USING DATA FROM THE AIDS LINK TO INTRAVENOUS EXPERIENCE COHORT STUDY

Here we provide SAS code to fit the joint marginal structural Cox proportional hazards model described in the main text using data from the AIDS Link to Intravenous Experience (ALIVE) cohort study.[1] The ALIVE data file, ALIVEDATA, contains multiple records per participant where each participant is uniquely identified by the variable, ID. Each record corresponds to a calendar date, VISITY, when a study measurement was taken. The ENTER variable denotes the time since study entry for the VISITY of the current record, while the variable EXIT represents the time since study entry corresponding to the VISITY of the subsequent record. However, for the final record, EXIT represents the time since study entry at exit from follow up due to HIV infection or censoring. A description of the 2-part process used to fit the joint marginal structural Cox model using ALIVEDATA follows. In general, the stabilized weight, $W(t) = W^{X_1}(t)W^{X_2}(t)W^{D_1}(t)W^{D_2}(t)$, described in the main text was estimated in parts 1a through 1e. The stabilized weight estimated in part 1 was used to fit the joint Cox model in part 2. Note in the provided SAS code, $W(t)$ is represented as, w, while $W^{X_1}(t)$, $W^{X_2}(t)$, $W^{D_1}(t)$, and $W^{D_2}(t)$ are shown as wx1, wx2, wd1, and wd2.

In part 1a, two pooled logistic models were fit to ALIVEDATA and used to estimate the conditional probabilities for the numerator and denominator of the weight, wx1. The outcome in both models was PDBIN which was an indicator of whether a participant reported heavy drinking in the prior six months at the current measurement time (PDBIN=1 for yes; PDBIN=0

for no). The *proc logistic* statement here and elsewhere without the *desc* option models the log odds that the outcome is 0, in this case that PDBIN=0.

In the pooled logistic model used to estimate the conditional probabilities for the numerator of wx1, predictors included the time-updated time-specific intercepts (ENTER ENTER1 ENTER2 ENTER3), time-updated indicators of heavy drinking (PDBIN1B) and injection drug use (INJBIN1B) in the preceding six months and collected at the prior measurement time, as well as the time-fixed confounders measured at study entry, Z_1. These time-fixed confounders included age (AGE, AGE1, AGE2, AGE3), an indicator for sex (FEMALE), and years of education (EDY EDY1 EDY2 EDY3). Product terms between the time-fixed confounders (AGEEDY=AGE*EDY, FEMAGE=FEMALE*AGE, and FEMEDY=FEMALE*EDY) were also included as predictors. Note here and elsewhere restricted quadratic splines[2] were used to model the continuous predictors, ENTER, AGE, and EDY where the additional variables beyond the main effect included in the regression (e.g., AGE1, AGE2, and AGE3) correspond to the restricted quadratic spline basis functions. The pooled logistic model outputs the conditional probability that PDBIN=0 (num1) at each record. These outputted probabilities are saved in the dataset, n1. So that the numerator of the weight, wx1, is derived from the conditional probability of reporting the level of alcohol consumption that a participant reported at a given measurement time, for records where PDBIN=1, the data step subsequent to the pooled logistic regression was used to reassign num1 to correspond to the probability that PDBIN=1 in the n1 dataset.

In the pooled logistic model used to estimate the conditional probabilities for the denominator of wx1, predictors included the time-updated time-specific intercepts, the time-updated indicators of heavy drinking and injection drug use collected at the prior measurement

time, the time-fixed confounders included in the pooled logistic model used to estimate the numerator for wx1, as well as the time-updated confounders included in L_1. The time-updated confounders pertained to the preceding six months and were collected at the prior measurement time. These time-updated confounders included indicators for the report of a sexually transmitted infection (STI1B), the number of sexual partners (SEX3CAT1B), male sex with a male (MSM1B), cocaine use (COKE1B), and shooting gallery attendance (GALL1B). The pooled logistic model outputs the conditional probability that PDBIN=0 (den1) at each record. These outputted probabilities are saved in the dataset d1. So that the denominator of the weight, wx1, is derived from the conditional probability of reporting the level of alcohol consumption that a participant reported at a given measurement time, for records where PDBIN=1, the data step subsequent to the pooled logistic regression was used to reassign den1 to correspond to the probability that PDBIN=1 in the d1 dataset.

In part 1b, two pooled logistic models were fit to ALIVEDATA and used to estimate the conditional probabilities for the numerator and denominator of the weight, wx2. The outcome in both models was INJBIN which was an indicator of whether a participant reported any injection drug use in the prior six months at the current measurement time (INJBIN=1 for yes; INJBIN=0 for no). The *proc logistic* statement models the log odds that INJBIN=0.

In the pooled logistic model used to estimate the conditional probability for the numerator of wx2, predictors included the time-updated time-specific intercepts, time-updated indicators of heavy drinking (PDBIN) and injection drug use (INJBIN1B) pertaining to the preceding six months, but collected at the concurrent and prior measurement times, respectively, as well as the time-fixed confounders measured at study entry, Z_2. These time-fixed confounders

were the same as those included in Z_1. The pooled logistic model outputs the conditional probability that INJBIN=0 (num2) at each record. These outputted probabilities are saved in the dataset, n2. So that the numerator of the weight, wx2, is derived from the conditional probability of reporting the level of injection drug use that a participant reported at a given measurement time, for records where INJBIN=1, the data step subsequent to the pooled logistic regression was used to reassign num2 to correspond to the probability that INJBIN=1 in the n2 dataset.

In the pooled logistic model used to estimate the conditional probability for the denominator of wx2, predictors included the time-updated time-specific intercepts, time-updated indicators of heavy drinking and injection drug use pertaining to the preceding six months, but collected at the concurrent and prior measurement times, time-fixed confounders included in the pooled logistic model used to estimate the numerator of wx2, as well as the time-updated confounders included in L_2. These time-updated confounders were the same as those included in L_1. The pooled logistic model outputs the conditional probability that INJBIN=0 (den2) at each record. These outputted probabilities are saved in the dataset d2. So that the denominator of the weight, wx2, is derived from the conditional probability of reporting the level of injection drug use that a participant reported at a given measurement time, for records where INJBIN=1, the data step subsequent to the pooled logistic regression was used to reassign den2 to correspond to the probability that INJBIN=1 in the d2 dataset.

In part 1c, two pooled logistic models were fit to ALIVEDATA and used to estimate the conditional probabilities for the numerator and denominator of the weight, wd1. The outcome in both models was DROP which was an indicator of whether a participant dropped out of the study at a given EXIT time (DROP=1 for yes; DROP=0 for no). The *proc logistic* statement models

the log odds that DROP=0.

In the pooled logistic model used to estimate the conditional probabilities for the numerator in wd1, predictors included the time-updated time-specific intercepts (EXIT EXIT1 EXIT2 EXIT3), time-updated indicators of heavy drinking (PDBIN) and injection drug use (INJBIN) pertaining to the preceding six months and collected at the most recent measurement time, as well as the time-fixed common predictors measured at study entry, Z_3. These time-fixed common predictors were the same as those included in Z_1 and Z_2. However, product terms between the time-fixed common predictors were not included as predictors in the pooled logistic model. The pooled logistic model outputs the conditional probability that DROP=0 (num3) at each record where the numerator of wd1 is the product of conditional probabilities that DROP=0. These outputted probabilities are saved in the dataset, n3.

In the pooled logistic model used to estimate the conditional probabilities for the denominator in wd1, predictors included the time-updated time-specific intercepts, time-updated indicators of heavy drinking and injection drug use pertaining to the preceding six months and collected at the most recent measurement time, time-fixed common predictors included in the pooled logistic model used to estimate the numerator for wd1, as well as the time-updated common predictors included in L_3. These time-updated common predictors were the same covariates as those included L_1 and L_2, but collected at the most recent measurement time. The pooled logistic model outputs the conditional probability that DROP=0 (den3) where the denominator of wd1 is the product of conditional probabilities that DROP=0. These outputted probabilities are saved in the dataset, d3.

In part 1d, two pooled logistic models were fit to ALIVEDATA and used to estimate the

conditional probabilities for the numerator and denominator of the weight, wd2. The outcome in both models was DIED which was an indicator of whether a participant died at a given EXIT time (DIED=1 for yes; DIED=0 for no). The *proc logistic* statement models the log odds that DIED=0. The specification of the predictors included in these pooled logistic regression models were identical to the specifications used to estimate wd1. The pooled logistic model outputs the conditional probability that DIED=0 (num4 and den4) where the numerator and denominator of wd2 is the product of conditional probabilities that DIED=0. These outputted probabilities are saved in the n4 and d4 datasets.

In part 1e, the data sets with the conditional probabilities are merged with ALIVEDATA to form dataset, e. In the data step that follows the merge, the weights wx1, wx2, wd1, and wd2 are obtained by taking the ratio of the products of the conditional probabilities estimated in parts 1a through 1d. The weight, w is obtained from taking the product of wx1, wx2, wd1, and wd2.

In part 2, the *proc phreg* statement is used to fit a weighted Cox model for the joint effect of heavy drinking and injection drug use on HIV acquisition. The indicators for heavy drinking and any injection drug use in the prior six months and collected at the current measurement time were included as predictors in the weighted Cox model along with the product term between heavy drinking and any injection drug use (PDINJ=PDBIN*INJBIN). The time-fixed covariates included in Z_1, Z_2, and Z_3 were also included as predictors. The outcome was an indicator of HIV infection at a given exit time (HIV=1 if yes, HIV=0 if no). The *covs* option outputs model parameters with robust standard errors. The statement *weight w*, re-weights the data using the weights, w, estimated in part 1. The *test* statement performs a joint Wald test on the product term represented by PDINJ. The output from this Wald test is labeled *pterm*. *Contrast* statements with the *estimate=exp* option are used to estimate hazard ratios for heavy drinking alone, any

injection drug use alone, and both heavy drinking and any injection drug use.

GLOSSARY OF DATA AND VARIABLES USED IN PROVIDED SAS CODE

ALIVEDATA is the ALIVE data file. The variables included in ALIVEDATA appear in uppercase and are defined as:

- ID – Unique participant identifier
- VISITY – Calendar date when study measurement was taken
- ENTER – Time since study entry for the VISITY of the current record
- ENTER1, ENTER2, ENTER3 – Restricted quadratic spline basis functions for ENTER
- EXIT – Time since study entry corresponding to the VISITY of the subsequent record; for the final record, time since study entry at exit from follow up due to HIV infection or censoring
- EXIT1, EXIT2, EXIT3 – Restricted quadratic spline basis functions for EXIT
- PDBIN – Indicator of whether a participant reported heavy drinking in the prior six months at the current measurement time (PDBIN=1 for yes; PDBIN=0 for no)
- PDBIN1B – Indicator of whether a participant reported heavy drinking in the prior six months at the prior measurement time (PDBIN1B=1 for yes; PDBIN1B=0 for no)
- INJBIN – Indicator of whether a participant reported any drug injections in the prior six months at the current measurement time (INJBIN=1 for yes; INJBIN=0 for no)
- INJBIN1B – Indicator of whether a participant reported any drug injections in the prior six months at the prior measurement time (INJBIN1B=1 for yes; INJBIN1B=0 for no)
- PDINJ – PDBIN*INJBIN
- AGE – Age at study entry
- AGE1, AGE2, AGE3 – Restricted quadratic spline basis functions for AGE
- FEMALE – Indicator of female sex (FEMALE=1 for yes; FEMALE=0 for no)

- EDY – Years of education at study entry
- EDY1, EDY2, EDY3 – Restricted quadratic spline basis functions for EDY
- AGEEDY – AGE*EDY
- FEMAGE – FEMALE*AGE
- FEMEDY – FEMALE*EDY
- STI – Indicator for the report of a sexually transmitted infection during the prior six months at the current measurement time
- STI1B – Indicator for the report of a sexually transmitted infection during the prior six months at the prior measurement time
- SEX3CAT – Indicators for the reported number of sexual partners during the prior six months at the current measurement time
- SEX3CAT1B – Indicators for the reported number of sexual partners during the prior six months at the prior measurement time
- MSM – Indicator for the report of male sex with a male during the prior six months at the current measurement time
- MSM1B – Indicator for the report of male sex with a male during the prior six months at the prior measurement time
- COKE – Indicator for the report of cocaine use during the prior six months at the current measurement time
- COKE1B – Indicator for the report of cocaine use during the prior six months at the prior measurement time
- GALL – Indicator for the report of shooting gallery attendance during the prior six months at the current measurement time

- GALL1B – Indicator for the report of shooting gallery attendance during the prior six months at the prior measurement time
- DROP – Indicator of whether a participant dropped out of the study at a given EXIT time (DROP=1 for yes; DROP=0 for no)
- DIED – Indicator of whether a participant died at a given EXIT time (DIED=1 for yes; DIED=0 for no)
- HIV – Indicator of HIV infection at a given EXIT time (HIV=1 for yes; HIV=0 for no)

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

- num1 – Conditional probability of reporting the level of alcohol consumption that a participant reported at a given measurement time in numerator of wx1
- n1 – Dataset that contains num1
- den1 – Conditional probability of reporting the level of alcohol consumption that a participant reported at a given measurement time in denominator of wx1
- d1 – Dataset that contains den1
- num2 – Conditional probability of reporting the level of injection drug use that a participant reported at a given measurement time in numerator of wx2
- n2 – Dataset that contains num2
- den2 – Conditional probability of reporting the level of injection drug use that a participant reported at a given measurement time in denominator of wx2
- d2 – Dataset that contains den2
- num3 – Conditional probability of a participant not dropping out at a given EXIT time in numerator of wd1

- n3 – Dataset that contains num3
- den3– Conditional probability of a participant not dropping out at a given EXIT time in denominator of wd1
- d3 – Dataset that contains den3
- num4 – Conditional probability of a participant not dying at a given EXIT time in numerator of wd2
- n4 – Dataset that contains num4
- den4– Conditional probability of a participant not dying at a given EXIT time in denominator of wd2
- d4 – Dataset that contains den4
- e – Dataset resulting from merging ALIVEDATA, n1, d1, n2, d2, n3, d3, n4, and d4
- wx1 – Exposure weight for alcohol
- wx2 – Exposure weight for injection drug use
- wx3 – Censoring weight for dropout
- wx4 – Censoring weight for death
- w – Inverse probability-of-exposure-and-censoring weight used to fit joint Cox model

SAS (VERSION 9.2) CODE

```
/***Part 1: Estimating Stabilized Weight, w;***/

/**Part 1a: estimating conditional probabilities for stabilized weight, wx1 using pooled logistic regression;**/
/*Numerator for wx1;*/
*Sorting records in ALIVEDATA by ID and VISITY;
proc sort data=ALIVEDATA; by ID VISITY; run;
*Modeling the log odds that PDBIN=0 and outputting corresponding probabilities as num1 into n1 dataset;
proc logistic data=ALIVEDATA;
        model PDBIN=ENTER ENTER1 ENTER2 ENTER3 PDBIN1B INJBIN1B AGE AGE1 AGE2 AGE3 FEMALE EDY
                        EDY1 EDY2 EDY3 AGEEDY FEMAGE FEMEDY;
        output out=n1(keep=ID VISITY PDBIN num1) p=num1;
run;
*For records where PDBIN=1, reassigning num1 to correspond to the probability that PDBIN=1 in n1 dataset;
data n1;
        set n1;
        if PDBIN=1 then num1=1-num1;
        keep ID VISITY PDBIN num1;
run;

/*Denominator for wx1;*/
*Modeling the log odds that PDBIN=0 and outputting corresponding probabilities as den1 into d1 dataset;
proc logistic data=ALIVEDATA;
        class SEX3CAT1B/param=ref desc;
        model PDBIN=ENTER ENTER1 ENTER2 ENTER3 PDBIN1B INJBIN1B AGE AGE1 AGE2 AGE3 FEMALE EDY
                        EDY1 EDY2 EDY3 AGEEDY FEMAGE FEMEDY STI1B SEX3CAT1B MSM1B COKE1B GALL1B;
        output out=d1(keep=ID VISITY PDBIN den1) p=den1;
run;
*For records where PDBIN=1, reassigning den1 to correspond to the probability that PDBIN=1 in d1 dataset;
data d1;
        set d1;
```

```
        if PDBIN=1 then den1=1-den1;
        keep ID VISITY PDBIN den1;
run;

/**Part 1b: estimating conditional probabilities for stabilized weight, wx2 using pooled logistic regression;** /
/*Numerator for wx2;*/
*Modeling the log odds that INJBIN=0 and outputting corresponding probabilities as num2 into n2 dataset;
proc logistic data=ALIVEDATA;
        model INJBIN=ENTER ENTER1 ENTER2 ENTER3 PDBIN INJBIN1B AGE AGE1 AGE2 AGE3 FEMALE EDY
                        EDY1 EDY2 EDY3 AGEEDY FEMAGE FEMEDY;
        output out=n2(keep=ID VISITY INJBIN num2) p=num2;
run;
*For records where INJBIN=1, reassigning num2 to correspond to the probability that INJBIN=1 in n2 dataset;
data n2;
        set n2;
        if INJBIN=1 then num2=1-num2;
        keep ID VISITY INJBIN num2;
run;

/*Denominator for wx2 ;*/
*Modeling the log odds that INJBIN=0 and outputting corresponding probabilities as den2 into d2 dataset;
proc logistic data=ALIVEDATA;
        class SEX3CAT1B/param=ref desc;
        model INJBIN=ENTER ENTER1 ENTER2 ENTER3 PDBIN INJBIN1B AGE AGE1 AGE2 AGE3 FEMALE EDY
                        EDY1 EDY2 EDY3 AGEEDY FEMAGE FEMEDY STI1B SEX3CAT1B MSM1B COKE1B GALL1B;
        output out=d2(keep=ID VISITY INJBIN den2) p=den2;
run;
*For records where INJBIN=1, reassigning den2 to correspond to the probability that INJBIN=1 in d2 dataset;
data d2;
        set d2;
        if INJBIN=1 then den2=1-den2;
        keep ID VISITY INJBIN den2;
```

```
run;

/**Part 1c: Estimating conditional probabilities for stabilized weight, wd1  using pooled logistic regression;**/
/*Numerator for wd1;*/
*Modeling the log odds that DROP=0 and outputting corresponding probabilities as num3 into n3 dataset;
proc logistic data=ALIVEDATA;
        model DROP=EXIT EXIT1 EXIT2 EXIT3 PDBIN INJBIN AGE AGE1 AGE2 AGE3 FEMALE EDY
                        EDY1 EDY2 EDY3;
        output out=n3(keep=ID VISITY num3) p=num3;
run;

/*Denominator for wd1;*/
*Modeling the log odds that DROP=0 and outputting corresponding probabilities as den3 into d3 dataset;
proc logistic data=ALIVEDATA;
        class SEX3CAT/param=ref desc;
        model DROP=EXIT EXIT1 EXIT2 EXIT3 PDBIN INJBIN AGE AGE1 AGE2 AGE3 FEMALE EDY
                        EDY1 EDY2 EDY3 STI SEX3CAT MSM COKE GALL;
        output out=d3(keep= ID VISITY den3) p=den3;
run;

/**Part 1d: estimating conditional probabilities for stabilized weight, wd2 using pooled logistic regression;**/
/*Numerator for wd2;*/
*Modeling the log odds that DIED=0 and outputting corresponding probabilities as num4 into n4 dataset;
proc logistic data=ALIVEDATA;
        model DIED=EXIT EXIT1 EXIT2 EXIT3 PDBIN INJBIN AGE AGE1 AGE2 AGE3 FEMALE EDY
                        EDY1 EDY2 EDY3;
        output out=n4(keep= ID VISITY num4) p=num4;
run;

/*Denominator for wd2;*/
*Modeling the log odds that DIED=0 and outputting corresponding probabilities as den4 into d4 dataset;
proc logistic data=ALIVEDATA;
```

```
	class SEX3CAT/param=ref desc;
	model DIED=EXIT EXIT1 EXIT2 EXIT3 PDBIN INJBIN AGE AGE1 AGE2 AGE3 FEMALE EDY
		EDY1 EDY2 EDY3 STI SEX3CAT MSM COKE GALL;
	output out=d4(keep= ID VISITY den4) p=den4;
run;

/**Part 1e: calculating cumulative probabilities for stabilized weights and in turn, w;**/
*Sorting records in ALIVEDATA and all generated datasets by ID and VISITY;
proc sort data=ALIVEDATA; by ID VISITY; run;
proc sort data=n1; by ID VISITY; run;
proc sort data=d1; by ID VISITY; run;
proc sort data=n2; by ID VISITY; run;
proc sort data=d2; by ID VISITY; run;
proc sort data=n3; by ID VISITY; run;
proc sort data=d3; by ID VISITY; run;
proc sort data=n4; by ID VISITY; run;
proc sort data=d4; by ID VISITY; run;

*Merging ALIVEDATA with generated datasets by ID and VISITY;
data e; merge ALIVEDATA n1 d1 n2 d2 n3 d3 n4 d4; by ID VISITY; run;
*Sorting merged dataset, e;
proc sort data=e; by ID VISITY; run;
*Calculating wx1, wx2, wd1, and wd2;
data e;
	set e;
	by ID VISITY;
	retain x1num x1den x2num x2den d1num d1den d2num d2den;
	if first.id then do; x1num=1; x1den=1; x2num=1; x2den=1; d1num=1; d1den=1; d2num=1; d2den=1; end;
	x1num=x1num*num1;
	x1den=x1den*den1;
	wx1=x1num/x1den;
	x2num=x2num*num2;
```

```
        x2den=x2den*den2;
        wx2=x2num/x2den;
        d1num=d1num*num3;
        d1den=d1den*den3;
        wd1=d1num/d1den;
        d2num=d2num*num4;
        d2den=d2den*den4;
        wd2=d2num/d2den;
run;
*Calculating w;
data e;
        set e;
        w=wx1*wx2*wd1*wd2;
        label num1= den1= num2= den2= num3= den3= num4= den4=;
run;

/***Part 2: Fitting Joint Marginal Structural Cox Proportional Hazards Model;***/

proc phreg data=e covs;
        model (ENTER, EXIT)*HIV(0)=PDBIN INJBIN PDINJ AGE AGE1 AGE2 AGE3 FEMALE EDY
                                        EDY1 EDY2 EDY3/ties=efron rl;
        weight w;
        pterm: test  PDINJ;
        contrast ' PDBIN=1 and INJBIN=0 vs PDBIN =0 and INJBIN=0' PDBIN 1 / estimate=exp;
        contrast ' PDBIN=0 and INJBIN=1 vs PDBIN =0 and INJBIN=0' INJBIN 1  / estimate=exp;
        contrast ' PDBIN=1 and INJBIN=1 vs PDBIN=0 and INJBIN=0' PDBIN 1 INJBIN 1 PDINJ 1 / estimate=exp;
run;
```

REFERENCES

1. Vlahov D, Anthony J, Muñoz A, et al. The ALIVE Study, a longitudinal study of HIV-1 infection in intravenous drug users: description of methods. *NIDA Res Monogr* 1991;**109**:75-100.

2. Howe CJ, Cole SR, Westreich DJ, Greenland S, Napravnik S, Eron JJ, Jr. Splines for Trend Analysis and Continuous Confounder Control [Research Letter]. *Epidemiology* 2011;**22**(6):874-875.