Estimating survival in left filtered data: simulation and application

Kevin Chen

Stat 256: Causal Inference (Fall 2021)

Introduction

The United Auto Workers-General Motors (UAW-GM) Cohort Study is a longitudinal occupational cohort study established in the early 1980s to study the health effects of metalworking fluids (Eisen et al. 1992, 2001). Metalworking fluids (MWF) are complex mixtures of fluids used in industrial metalworking operations to lubricate and cool machinery and parts. The three major classes of MWF are straight, soluble, and synthetic metalworking fluids (Byers 2006). Possible routes of human exposure include absorption through skin, inhalation or aerosols, and ingestion of droplets.

A central concern in the analysis of occupational cohorts is the potential for the healthy worker survivor effect (HWSE), the phenomenon by which healthy individuals remain at work, while less healthy individuals leave work – possibly in response to exposure-related health decline. In the presence of the HWSE, those with the highest cumulative occupational exposures are also those who are less at risk of disease. Thus, standard measures of association would show an inverse relationship between occupational exposure and poor health outcomes (Arrighi and Hertz-Picciotto 1994). The HWSE is an example of time-varying confounding affected by past exposure. Previous studies have attempted to assess the presence of the HWSE in observed data by assessing so-called path-specific associations using Cox proportional hazards modeling (Naimi et al. 2013; Garcia et al. 2017). However, these measures of associations are themselves subject to the confounding they seek to quantify.

If sequential ignorability of exposure status at each point in follow-up and positivity can be attained conditional on covariates, then causal methods can be applied to account for the HWSE. Past studies have applied causal methods capable of accounting for time-varying confounding affected by past exposure to the study of MWF exposures and cancer mortality outcomes in the UAW-GM Cohort Study (Garcia et al. 2018; Izano et al. 2019), but the study of cancer incidence outcomes is further problematized by incomplete observation of cancer incidence outcomes over the study period. We wish to make inferences about the carcinogenicity of MWF exposure over an individual's lifetime starting three years after they enter the workforce. The UAW-GM cohort includes those hired roughly between 1938 and 1982. However, cancer incidence reporting at the Michigan Cancer Registry did not begin until 1985. In particular, we have the presence of *left filtering* in the UAW-GM

Cohort Study when cancer incidence is the outcome of interest: before 1985, both cancer incidence status and time of cancer incidence are unknown. Observation of the complete cancer incidence outcome vector over the study period is conditional on an individual surviving to 1985 cancer-free.

In the presence of the HWSE, left filtering implies outcome misclassification that is informative of true cancer status. As part of her dissertation research, Izano (2017) conducted a quantitative bias analysis for the estimation of survival curves in the presence of left filtering and the HWSE. She simulated data compatible with the HWSE and estimated cancer-free survival curves using an inverse probability of treatment and censoring weighted Kaplan-Meier (WKM) estimator and the WKM with an Aalen filter for left-filtering (AWKM) (Andersen et al. 1993; Xie and Liu 2005). Data were simulated under five different scenarios. The causal estimands of interest were the survival curves under interventions: (1) always exposed at work with no censoring due to death and (2) never exposed at work with no censoring due to death. For each intervention, three survival curves were produced: (1) the true survival curve, (2) the WKM survival curve, and (3) AWKM survival curve. The bias in mean survival difference calculated by AWKM was negligible when the true 20-year risk of cancer was under 0.25.

Here, we replicate the simulation and bias analysis presented in Chapter 3 of Izano (2017), embed the problem in the non-parametric structural causal approach of Pearl (1995), comment on the results, and apply the WKM and AWKM estimators to the UAW-GM Cohort.

Methods

Causal model

The UAW-GM Cohort data includes person-year level exposure, outcome, and covariate data starting three years after hire. To emulate the shape of the data for this longitudinal cohort, we consider 20 years of data over time indexed by years since hire. Notation representing the variables of interest are presented in Table 1.

Assume we have $n = 50\,000$ iid units in X with

$$X_i(t) = \left(R_i = 0, W_i, S_i, \bar{H}_i(t), \bar{N}_i(t), \bar{A}_i(t), \bar{a}_i(t), \bar{Y}_i^*(t) = \bar{Y}_i(t)\right).$$

The bar notation to indicates variable history where $\bar{X}_i(t) = \left(X_i(k)\right)_{k=1}^{t}$. In the case of exposure, $\bar{A}_i(t)$ is the indicator of whether unit i was ever exposed through time t. Note that true cancer status $Y^*(t)$ not observed until $t \geq R$, after the start of the registry. Call X the full data, where we have $R = 0$ for all. In the observed data X^{obs}, we cannot assume $R = 0$ for all, and susceptibility S is not known:

$$X_i^{\text{obs}}(t) = \left(R_i, W_i, \bar{N}_i(t), \bar{A}_i(t), \bar{a}_i(t), \bar{Y}_i(t)\right).$$

Table 1: Descriptions of variables.

Variable	Description
R	Time until start of registry
W	Baseline covariates
S	Susceptibility to effects of metalworking fluid exposure
$H(t)$	Adverse health status at time t
$N(t)$	Employment status at time t
$A(t)$	Metalworking fluid exposure at time t
$D(t)$	Mortality status at time t
$Y^*(t)$	Cancer status at time t
$Y(t)$	Observed Cancer status at time t
$t = \{1, 2, \dots, 20\}$	Time, indexed in years after hire

Under the causal model, we assume the following non-parametric structural equations:

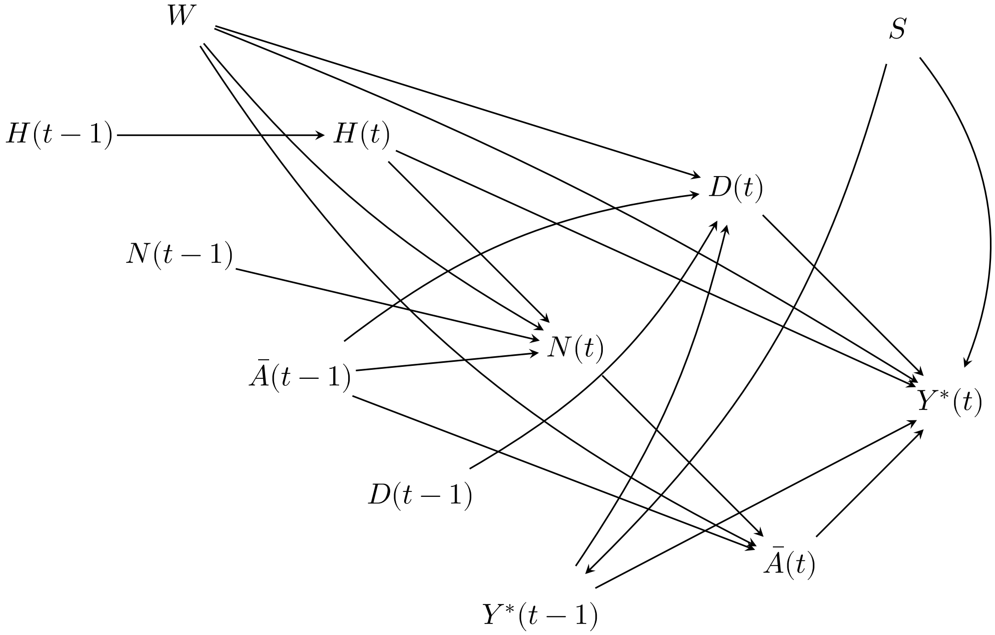
$$\begin{aligned}
R &= f_R\left(U_R\right) \\
W &= f_W\left(U_W\right) \\
S &= f_S\left(U_S\right) \\
H(t) &= f_{H(t)}\left(H(t-1), U_{H(t)}\right) \\
N(t) &= f_{N(t)}\left(W, N(t-1), H(t), A(t-1), U_{N(t)}\right) \\
A(t) &= f_{A(t)}\left(W, \bar{A}(t-1), N(t), U_{A(t)}\right) \\
D(t) &= f_{D(t)}\left(W, \bar{A}(t-1), D(t-1), Y^*(t-1), U_{D(t)}\right) \\
Y^*(t) &= f_{Y^*(t)}\left(W, S, H(t), \bar{A}(t), D(t), Y^*(t-1), U_{Y^*(t)}\right) \\
Y(t) &= Y^*(t) \times \mathbb{1}\left[Y^*(\lfloor R \rfloor) = 0\right] \times \mathbb{1}\left[D(t) = 0\right] \ .
\end{aligned}$$

The exogenous variables (errors) $U = \left(U_R, U_W, U_S, U_{H(t)}, U_{N(t)}, U_{A(t)}, U_{D(t)}, U_{Y^*(t)}\right)_{t=1}^{T}$ are mutually independent. Exposure status is a time-varying indicator; let exposure history $\bar{A}(t)$ be summarized as being ever-exposed or not with $\bar{A}(t) = \mathbb{1}\left[\sum_{k=1}^{t} \mathbb{1}\left[A(k) = 1\right] > 0\right]$. The outcome of interest is a survival outcome, so $Y^*(t-1) = 1 \Rightarrow Y^*(t) = 1$. The observed outcome $Y(t)$ at time t is a function of true outcome status and the time points delimiting left and right censoring. An abbreviated directed acyclic graph (DAG) representing the causal relationships encoded in the equations above is presented in Figure 1.

Simulation

Parametric relationships between the variables were imposed in simulations. For the $n = 50\,000$ units over $T = 20$ years, we have:

Figure 1: Directed acyclic graph representing the causal relationships encoded in the non-parametric structural equation model at time t.



- $U_j \overset{\text{iid}}{\sim} \text{uniform}\,[0,1]$ for all j
- In full data $R = 0$ otherwise $R \sim \text{uniform}\,[0, 30]$
- $W = \mathbb{1}\,[U_W \leq p_W] \sim \text{Bernoulli}\,(p_W)$
- $S = \mathbb{1}\,[U_S \leq p_S] \sim \text{Bernoulli}\,(p_S)$
- If $H(t-1) = 1$, then $H(t) = 1$ otherwise $H(t) = \mathbb{1}\,\left[U_{H(t)} \leq p_H\right] \sim \text{Bernoulli}\,(p_H)$
- if $N(t-1) = 0$ then $N(t) = 0$ otherwise

$$N(t) \sim \text{Bernoulli}\left\{\text{logit}\,\left(\beta_0^N + \beta_W^N W + \beta_H^N H(t) + \beta_A^N A(t-1) \times \mathbb{1}\,[t > 1] + U_{N(t)}\right)\right\}$$

- If $N(t) = 0$ then $A(t) = 0$ otherwise

$$A(t) \sim \text{Bernoulli}\left\{\text{logit}\,\left(\left(\beta_0^A + \beta_W^A W\right) \times \mathbb{1}\,[t = 1] + \beta_A^A A(t-1) \times \mathbb{1}\,[t > 1] + U_{A(t)}\right)\right\}$$

- If $D(t-1) = 1$ then $D(t) = 1$ otherwise

$$D(t) \sim \text{Bernoulli}\left\{\text{logit}\,\begin{pmatrix} \beta_0^D + \beta_W^D W + \beta_{\bar{A}}^D \bar{A}(t-1) \times \mathbb{1}\,[t > 1] \\ +\beta_{\bar{Y}}^D \displaystyle\sum_{k=1}^{t-1} Y^*(k) \times \mathbb{1}\,[t > 1] + U_{D(t)} \end{pmatrix}\right\}$$

Table 2: Simulation parameters.

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
p_S	0.10	0.10	**0.20**	0.10	0.10
p_W	0.20	0.20	0.20	0.20	0.20
p_H	0.30	0.30	0.30	0.30	0.30
β_0^N	3.00	3.00	3.00	3.00	3.00
β_W^N	-0.10	-0.10	-0.10	-0.10	-0.10
β_H^N	-0.50	-0.50	-0.50	**-1.50**	-0.50
β_A^N	-1.50	-1.50	-1.50	-1.50	-1.50
β_0^A	-1.50	-1.50	-1.50	-1.50	-1.50
β_W^A	-0.50	-0.50	-0.50	-0.50	-0.50
β_A^A	2.50	2.50	2.50	2.50	2.50
β_0^D	-5.50	-5.50	-5.50	-5.50	-5.50
β_W^D	1.00	1.00	1.00	1.00	1.00
$\beta_{\bar{A}}^D$	0.50	0.50	0.50	0.50	0.50
β_Y^D	0.50	**2.00**	0.50	0.50	0.50
β_0^Y	-7.00	-7.00	-7.00	-7.00	**-6.00**
β_W^Y	2.00	2.00	2.00	2.00	2.00
β_A^Y	0.25	0.25	0.25	0.25	0.25
$\beta_{\bar{A}}^Y$	0.20	0.20	0.20	0.20	0.20
β_H^Y	0.70	0.70	0.70	**1.70**	0.70
β_S^Y	0.30	0.30	0.30	0.30	0.30

- If $Y^*(t-1) = 1$ then $Y^*(t) = 1$ otherwise

$$Y^*(t) \sim \text{Bernoulli}\left\{ \text{logit}\left(\begin{array}{r} \beta_0^Y + \beta_W^Y W + \beta_A^Y A(t) + \beta_{\bar{A}}^Y \bar{A}(t-1) \times \mathbb{1}\left[t > 1\right] \\ + \beta_S^Y S \times \bar{A}(t) + \beta_H^Y H(t) + U_{Y^*(t)} \end{array} \right) \right\}$$

- If $t < R$ then $Y(t) = 0$
- If $t \geq R$ then $Y(t) = Y^*(t) \times \mathbb{1}\left[Y^*(\lfloor R \rfloor) = 0\right] \times \mathbb{1}\left[D(t) = 0\right]$.

Five sets of data were generated using these equations to represent five scenarios. Scenario 1 represents the base case where 10% of workers are susceptible to exposure-related effects, the odds ratio of mortality each additional year following cancer diagnosis is about 1.6, and there is moderate time-varying confounding by health status. In scenario 2, we have greater cancer-related mortality by increasing β_Y^D. In scenario 3, we increase p_S, the proportion of the study population susceptible to the carcinogenic effects of MWF exposure. In scenario 4, we consider greater time-varying confounding by health status by increasing β_H^N and β_H^Y. In the last scenario, we have greater background cancer incidence by increasing β_0^Y. The sets of parameters used in the five scenarios are presented in Table 2.

Interventions, potential outcomes, target parameters, and estimation

The substantive question of interest is the causal effect of occupational exposure to MWF on cancer incidence risk. Since occupational MWF exposure occurs only when individuals are at work, we define dynamic exposure regimes that depend on employment status. Under rule a_0, set $D(t) = 0$, and set $A(t) = 0$ while $N(t) = 1$. Under rule a_1, set $D(t) = 0$, and set $A(t) = 1$ while $N(t) = 1$. Under both rules, we prevent censoring by death as if it were intervenable. The causal effect is defined by contrasting the survival function $S_{a_1}(t) = 1 - \mathbb{E}\left[Y_{a_1}(t)\right]$ under rule a_1 to $S_{a_0}(t) = 1 - \mathbb{E}\left[Y_{a_0}(t)\right]$ that under rule a_0. Note that this causal estimand is defined over *a priori counterfactuals* not observable in the real world (Frangakis and Rubin 2002). This approach is standard in epidemiologic studies, however.

The survival function expresses the probability that a person following rule a is cancer-free at the end of time point t. The expected time until cancer under rule a is $\mu_a = \sum_0^K S_a(t)\, dt$. Our causal estimand is summary measure $\psi = \mu_{a_1} - \mu_{a_0}$, the difference in expected time until event under two different interventions over 20 years of follow-up under five different data generating scenarios. Bias will be evaluated by comparing estimates of ψ to its true value in 250 simulations per scenario (the original analysis performed 500). The true value will be calculated by simulating the full data for five hundred thousand individuals (the original analysis used one million) with rules a_0 and a_1 applied deterministically. Estimates of ψ will be obtained by first estimating the survival curves $S_a(t)$ using two estimators: the inverse probability weighted Kaplan-Meier estimator (WKM) and the Aalen-filtered WKM (AWKM). These survival estimators are detailed in the following section.

Kaplan-Meier estimator and extensions

To estimate survival, we applied extensions of the widely-known Kaplan-Meier (KM) estimator for survival (Kaplan and Meier 1958). First, we review the estimator of Xie and Liu (2005), an extension of the KM estimator where units are weighted by the inverse probability of treatment. The standard KM estimator requires counting up the number of cases $c_a^0(t)$ that occurred in interval $(t-1, t]$ and the number of units at risk $R_a^0(t)$ in that interval at all event times t. Assuming cancer status was assessed at the end of regular intervals $t = 1, \ldots, K$, we have:

$$c_a^0(t) = \sum_i^n \mathbb{1}\left[Y_i(t) = 1\right] \times \mathbb{1}\left[Y_i(t-1) = 0\right] \times \mathbb{1}\left[\bar{A}_i(t) = \bar{a}(t)\right]$$

$$R_a^0(t) = \sum_i^n \mathbb{1}\left[Y_i(t-1) = 0\right] \times \mathbb{1}\left[\bar{A}_i(t) = \bar{a}(t)\right].$$

The standard survival estimator is

$$\hat{S}_a^0(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{j \le t}\left(1 - \frac{c_a^0(j)}{R_a^0(j)}\right) & \text{if } t \ge t_1 \end{cases}$$

where t_1 is the first event time.

In observational studies, survival contrasts estimated using the standard KM estimator are biased for the true causal survival contrast. However, if conditional ignorability and positivity are attained, the inverse probability weighted KM (WKM) estimator of Xie and Liu (2005) yields unbiased estimates of the true causal survival curve. The WKM estimator augments the standard KM estimator by weighting units at time t by $w_{i,a}(t)$ the inverse probability of treatment:

$$\begin{aligned} c_a^w(t) &= \sum_i^n w_{i,a}(t) \times \mathbb{1}\left[Y_i(t) = 1\right] \times \mathbb{1}\left[Y_i(t-1) = 0\right] \times \mathbb{1}\left[\bar{A}_i(t) = \bar{a}(t)\right] \\ R_a^w(t) &= \sum_i^n w_{i,a}(t) \times \mathbb{1}\left[Y_i(t-1) = 0\right] \times \mathbb{1}\left[\bar{A}_i(t) = \bar{a}(t)\right] \end{aligned}$$

The WKM survival estimator for rule a is

$$\hat{S}_a^w(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{j \le t} \left(1 - \frac{c_a^w(j)}{R_a^w(j)}\right) & \text{if } t \ge t_1 \end{cases}$$

where t_1 is the first event time.

Finally, to account for (uninformative) left filtering, we applied the Aalen filter, which considers only the units at time t for which the outcome is observed:

$$\begin{aligned} c_a(t) &= \sum_i^n w_{i,a}(t) \times \mathbb{1}\left[Y_i(t) = 1\right] \times \mathbb{1}\left[Y_i(t-1) = 0\right] \times \mathbb{1}\left[\bar{A}_i(t) = \bar{a}(t)\right] \times \mathbb{1}\left[t \ge R_i\right] \\ R_a(t) &= \sum_i^n w_{i,a}(t) \times \mathbb{1}\left[Y_i(t-1) = 0\right] \times \mathbb{1}\left[\bar{A}_i(t) = \bar{a}(t)\right] \times \mathbb{1}\left[t \ge R_i\right] \end{aligned}$$

The Aalen-filtered WKM (AWKM) estimator for rule a is

$$\hat{S}_a(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{j \le t} \left(1 - \frac{c_a(j)}{R_a(j)}\right) & \text{if } t \ge t_1 \end{cases}$$

where t_1 is the first event time.

In the full data, the WKM and AWKM estimators are equivalent, and identification is achieved under overlap (positivity) and sequential ignorability (randomization) assumptions:

$$\begin{aligned} &Y^*_{a,\bar{d}=0}(t') \perp\!\!\!\perp A(t) \mid W,\ \bar{A}(t-1) = \bar{a}(t-1),\ D(t-1) = 0,\ N(t) = 1 \\ &Y^*_{a,\bar{d}=0}(t') \perp\!\!\!\perp D(t) \mid W,\ D(t-1) = 0,\ Y^*(t-1) = 0,\ \bar{A}(t-1) = \bar{a}(t-1) \end{aligned}$$

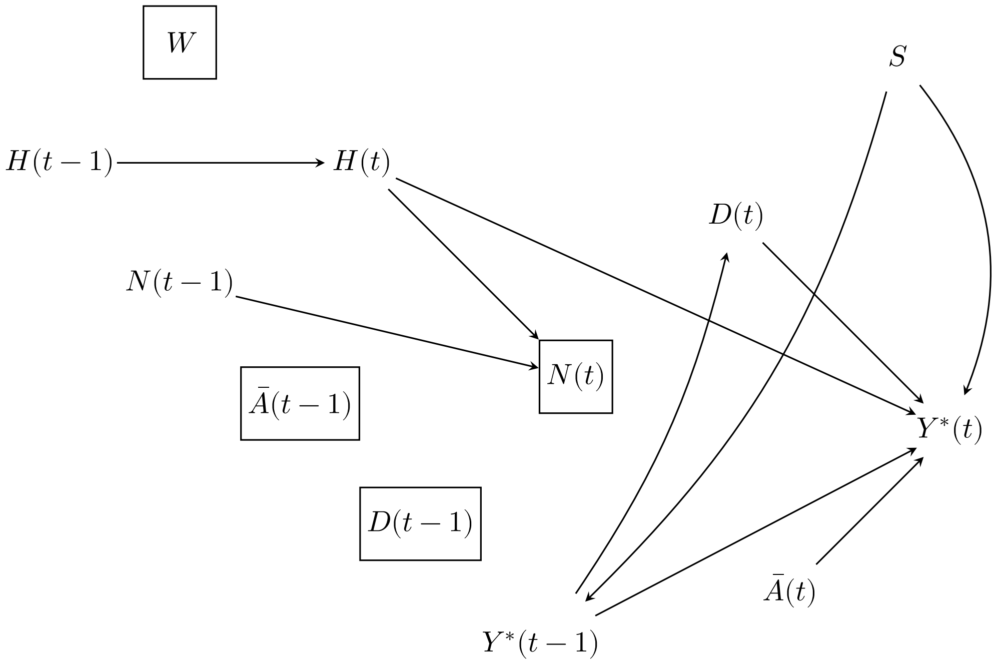
for all times $t' \ge t$, and

$$0 < \mathbb{P}\left(A(t) = 1 \mid W,\ \bar{A}(t-1) = \bar{a}(t-1),\ D(t-1) = 0,\ N(t) = 1\right) < 1$$

$$0 < \mathbb{P}\left(D(t) = 0 \mid W,\ D(t-1) = 0,\ Y^*(t-1) = 0,\ \bar{A}(t-1) = \bar{a}(t-1)\right) < 1.$$

Graphical representations of the first and second components of the ignorability assumption are presented in Figures 2 and 3 where conditioning on boxed variables are represented by the removal of edges pointing away from those variables. The resulting graphs show the fulfillment of Pearl's backdoor criterion for the estimation of the causal effects of $\bar{A}(t)$ on $Y^*(t)$ and $D(t)$ on $Y^*(t)$, respectively. Thus, the causal effect of the joint intervention on $(\bar{A}(t), D(t))$ at each time t is identified. Causal identification is not attained when true cancer status $Y^*(t)$ is not known.

Figure 2: Directed acyclic graph representing the causal relationships encoded in the non-parametric structural equation model at time t after conditioning on $\{W,\ \bar{A}(t-1),\ D(t-1),\ N(t)\}$.



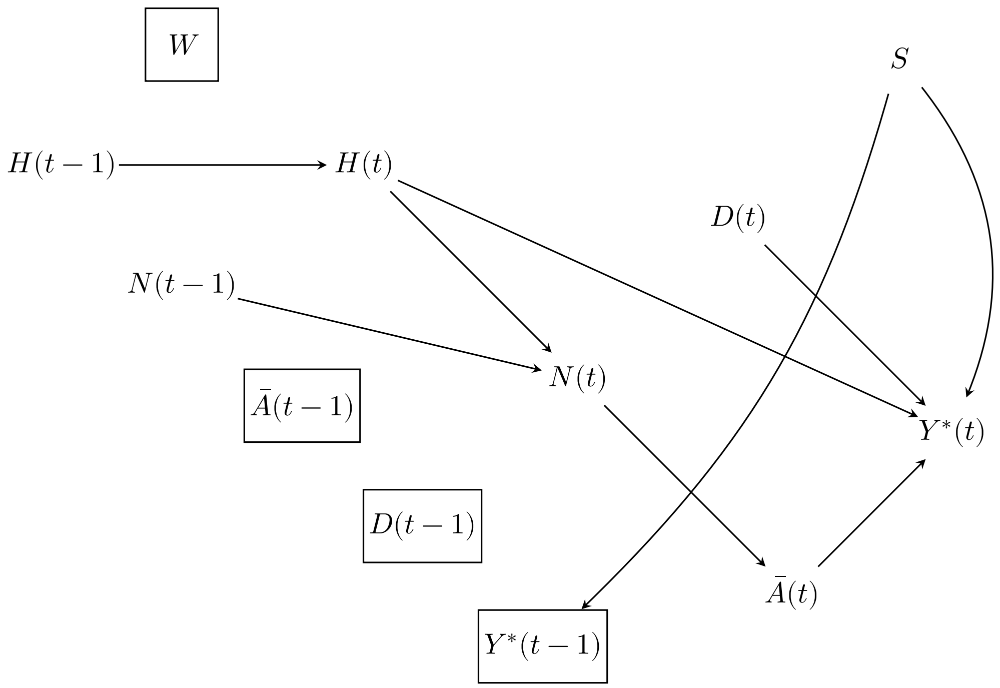
Estimation of weights

To estimate the weights for the WKM and AWKM estimators, we fit two logistic regressions at each time point $t = 1, \ldots 20$:

$$\text{logit}\left(\mathbb{P}\left(A(t) = 1 \mid W,\ \bar{A}(t-1),\ D(t-1) = 0,\ N(t) = 1\right)\right) = \alpha_0 + W\alpha_1 + \bar{A}(t-1)\alpha_2$$
$$\text{logit}\left(\mathbb{P}\left(D(t) = 1 \mid W,\ D(t-1) = 0,\ Y(t-1) = 0,\ \bar{A}(t-1)\right)\right) = \beta_0 + W\beta_1 + \bar{A}(t-1)\beta_2$$

The first will be fitted on data for those alive and at work at time t. The second will be among those alive and (observed to be) cancer-free. For each unit at time t, the weight will be calculated by taking the inverse of the cumulative probability of following the exposure rule and remaining

Figure 3: Directed acyclic graph representing the causal relationships encoded in the non-parametric structural equation model at time t after conditioning on $\{W,\ \bar{A}(t-1),\ D(t-1),\ Y^*(t-1)\}$.



uncensored:

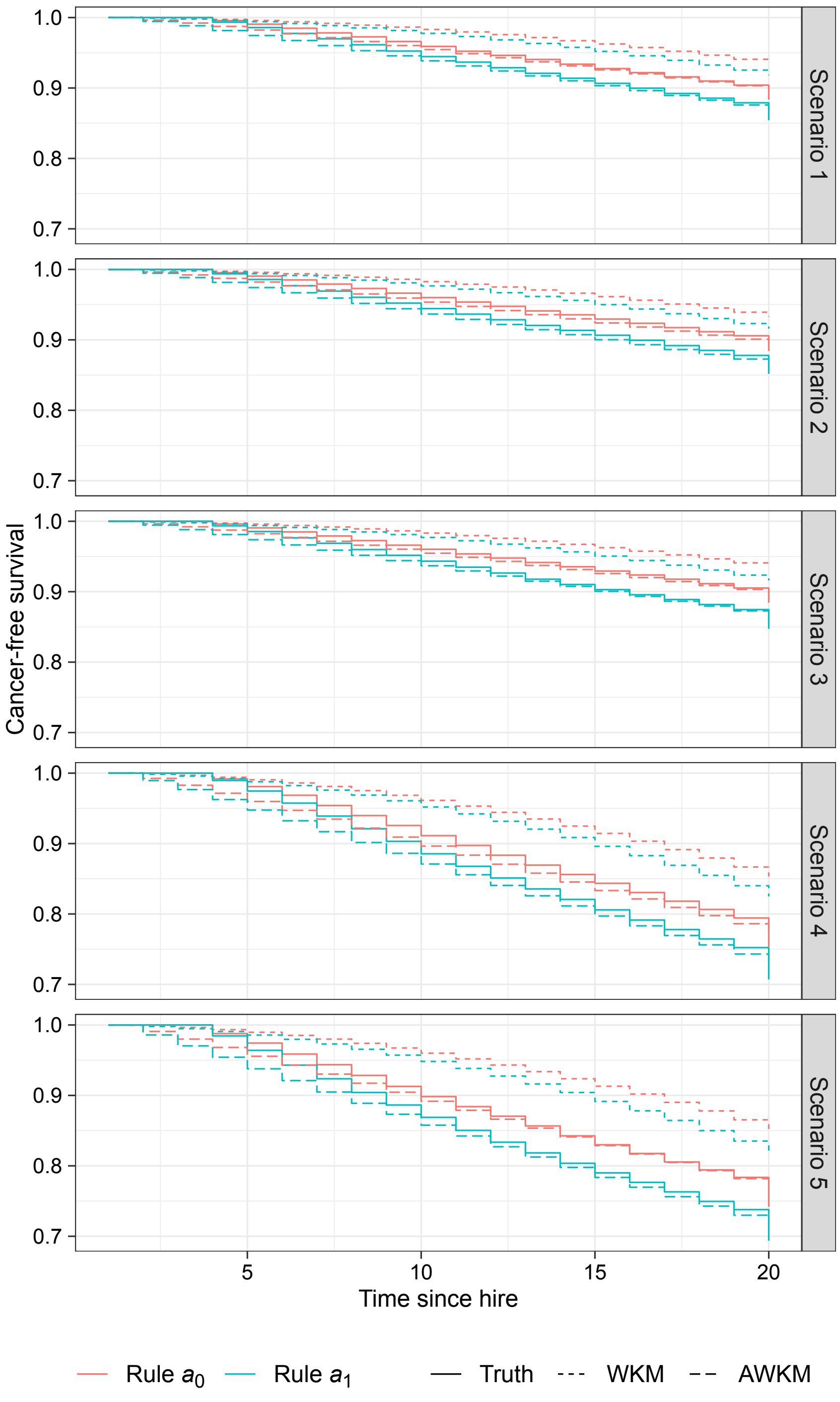
$$\hat{w}_a(t) = \left[\prod_{j=1}^{t} \begin{array}{r} \hat{\mathbb{P}}\left\{A(j)=a(j) \mid W,\ \bar{A}(j-1)=\bar{a}(j-1),\ D(j-1)=0,\ N(j)=1\right\} \times \\ \hat{\mathbb{P}}\left\{D(j)=0 \mid W,\ D(j-1)=0,\ Y(j-1)=0,\ \bar{A}(j)=\bar{a}(j)\right\} \end{array}\right]^{-1}.$$

Results

Figure 4 presents the true survival curves as well as the WKM and AWKM survival curves averaged over 250 replications for each intervention rule and scenario. Qualitatively, the WKM estimator consistently over-estimates survival whereas the the AWKM survival curve is much closer to the truth. The bias of the AWKM survival estimate appears to be larger earlier in follow-up and smaller as follow-up extends forward. The bias of both the the WKM and the AWKM survival curves appears largest in Scenario 5.

Table 3 presents true and estimated average cancer-free survival times under each intervention rule and scenario. Table 4 presents differences in survival time contrasting rule a_1 to rule a_0. Table 5 presents estimates of the bias of the WKM and AWKM estimators for ψ, the difference in average cancer-free survival time over 20 years of follow-up. These numeric results are consistent with the qualitative interpretations of Figure 4. The WKM estimator over-estimates the difference in cancer-free survival time, resulting in bias toward the null, whereas the AWKM estimator under-estimates the cancer-free survival, resulting in bias away from the null. In every scenario, the bias

Figure 4: Cancer-free survival over time since hire in five simulation scenarios. The true (discrete) survival curve is represented by the solid lines. The average inverse probability weighted Kaplan-Meier (WKM) survival curve is represented by the dashed-line with short dashes. The average Aalen-filtered inverse probability weighted Kaplan-Meier (AWKM) survival curve is represented by the dashed-line with long dashes. Estimated survival curves were averaged over 250 replicates. Salmon color indicates survival and survival estimates under rule a_0 when workers are always unexposed. Cyan color indicates those under rule a_1 when workers are always exposed while employed.



of the WKM-derived contrast is several times larger in magnitude than that of the AWKM-derived contrast. The bias of both estimators is greatest for Scenario 5.

The qualitative results here are consistent with those of Izano (2017). However, true and estimated survival in the present analysis is larger than those found previously. Furthermore, the true and estimator average mean differences in survival are smaller in magnitude in the present case. The magnitudes of the bias estimates are also smaller.

Table 3: True cancer-free survival time μ_a over 20-year follow-up and estimator averages over 250 replicates.

Rule	Estimator	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
a_0	Truth	19.08	19.10	19.10	18.02	17.83
	WKM	19.52	19.51	19.53	18.92	18.90
	AWKM	19.02	19.00	19.02	17.79	17.72
a_1	Truth	18.80	18.79	18.76	17.54	17.29
	WKM	19.39	19.37	19.37	18.68	18.62
	AWKM	18.70	18.67	18.67	17.30	17.07

Table 4: Difference in average cancer-free survival time over 20-year follow-up comparing rule a_1 always exposed to rule a_0 never exposed at work: true value ψ and estimator averages over 250 replicates.

Estimator	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Truth	-0.28	-0.31	-0.34	-0.48	-0.54
WKM	-0.14	-0.15	-0.15	-0.23	-0.28
AWKM	-0.31	-0.33	-0.35	-0.50	-0.64

Table 5: Bias estimates of estimators for ψ, the difference in average cancer-free survival time over 20 years of follow-up.

Estimator	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
WKM	0.14	0.16	0.19	0.25	0.26
AWKM	-0.03	-0.02	-0.01	-0.02	-0.10

References

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding. 1993. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer, New York, NY. https://books.google.com/books?id=kBnvAAAAMAAJ.

Arrighi, H. Michael, and Irva Hertz-Picciotto. 1994. "The Evolving Concept of the Healthy Worker Survivor Effect." *Epidemiology* 5 (2): 189–96. http://www.jstor.org/stable/3702361.

Byers, Jerry P. 2006. *Metalworking Fluids*. CRC Press.

Eisen, Ellen A, Judith Bardin, Rebecca Gore, Susan R Woskie, Marilyn F Hallock, and Richard R Monson. 2001. "Exposure-Response Models Based on Extended Follow-up of a Cohort Mortality Study in the Automobile Industry." *Scandinavian Journal of Work, Environment & Health* 27 (4): 240–49.

Eisen, Ellen A, Paige E Tolbert, Richard R Monson, and Thomas J Smith. 1992. "Mortality Studies of Machining Fluid Exposure in the Automobile Industry I: A Standardized Mortality Ratio Analysis." *American Journal of Industrial Medicine* 22 (6): 809–24.

Frangakis, Constantine E, and Donald B Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1): 21–29.

Garcia, Erika, Sally Picciotto, Sadie Costello, Patrick T Bradshaw, and Ellen A Eisen. 2017. "Assessment of the Healthy Worker Survivor Effect in Cancer Studies of the United Autoworkers-General Motors Cohort." *Occupational and Environmental Medicine* 74 (4): 294–300.

Garcia, Erika, Sally Picciotto, Andreas M Neophytou, Patrick T Bradshaw, John R Balmes, and Ellen A Eisen. 2018. "Lung Cancer Mortality and Exposure to Synthetic Metalworking Fluid and Biocides: Controlling for the Healthy Worker Survivor Effect." *Occupational and Environmental Medicine* 75 (10): 730–35.

Izano, Monika A. 2017. "Estimating Causal Effects of Occupational Exposures." PhD thesis, University of California, Berkeley; University of California, Berkeley.

Izano, Monika A, Oleg A Sofrygin, Sally Picciotto, Patrick T Bradshaw, and Ellen A Eisen. 2019. "Metalworking Fluids and Colon Cancer Risk: Longitudinal Targeted Minimum Loss-Based Estimation." *Environmental Epidemiology* 3 (1): e035.

Kaplan, Edward L, and Paul Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53 (282): 457–81.

Naimi, Ashley I, Stephen R Cole, Michael G Hudgens, M Alan Brookhart, and David B Richardson. 2013. "Assessing the Component Associations of the Healthy Worker Survivor Bias: Occupational Asbestos Exposure and Lung Cancer Mortality." *Annals of Epidemiology* 23 (6): 334–41.

Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82 (4): 669–88.

Xie, Jun, and Chaofeng Liu. 2005. “Adjusted Kaplan–Meier Estimator and Log-Rank Test with Inverse Probability of Treatment Weighting for Survival Data.” *Statistics in Medicine* 24 (20): 3089–3110.